Article

Identifying Key Issues in Climate Change Litigation: A Machine Learning Text Analytic Approach

Wullianallur Raghupathi [1,*] , **Dominik Molitor** [1] , **Viju Raghupathi** [2] and **Aditya Saharia** [1]

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: As climate change, environmental, social, and governance (ESG), along with sustainability, become increasingly crucial for businesses and society, there is a noticeable scarcity of information and transparency regarding corporate practices. Often, government agency enforcement actions lead to litigation and are ultimately resolved by court decisions. Moreover, in instances when there is perceived inadequacy in government enforcement, citizens frequently turn to the courts for preventive judgments against businesses or agencies. In an effort to shed light on the multifaceted aspects of climate change, we adopted a novel, exploratory approach to analyze climate change-related litigation cases. Utilizing a blend of machine learning-based text analytics, we have extracted key insights from individual case narratives. Our analysis encompassed over four hundred cases from the Westlaw database through various keyword searches. The emergent topics from our case dataset revolved around four critical environmental themes: forest, land, water, and air emissions. Our findings provide insight into the nature and dimensions of climate change and also carry significant policy implications, laying the groundwork for future research in this domain.

Keywords: climate change; machine learning; text analytics; litigation; legal case

1. Introduction

When considering climate change, thoughts often revolve around its significant environmental consequences, such as increasing sea levels, glacial melting, and rising temperatures. These changes contribute to the degradation of our planet, increase pollution levels, and pose significant health risks [1–5]. Regrettably, the climate crisis continues to worsen rather than abate. Each passing year witnesses a heightened intensity in the impacts of climate change. Hundreds of millions of people bear the brunt of increasingly frequent and severe extreme weather events, resulting in the loss of livelihoods and, tragically, lives. Annually, our economies, and in certain instances, entire nations, grapple with the tangible consequences of unforeseeable events [4–11]. As the Secretary-General of the United Nations stated in the twenty-seventh Conference of the Parties to the United Nations Framework Convention on Climate Change in November 2022, we are facing the most critical battle of our existence and, unfortunately, we find ourselves on the losing side, with potentially devastating consequences for our planet and future generations [5]. But, this is only the beginning of the impact of climate change, which is closely related to every aspect of human society [7–12].

Climate change not only intensifies existing health threats but also gives rise to new and daunting public health challenges. Indeed, it is widely regarded as the foremost threat to public health in the 21st century [2]. Climate change leads to rising temperatures, which in turn elevates the risk of heat-related illnesses and fatalities, worsens air quality which contributes to cardiopulmonary and respiratory diseases, facilitates the transmission of diseases through contaminated food, water, and vectors, and imposes significant stress

on mental health [4–7]. Without substantial worldwide reductions in greenhouse gases (GHGs), these effects will only intensify [13,14]. Climate change encompasses long-term changes in temperatures and weather patterns. Although certain changes may arise naturally, such as those linked to fluctuations in the solar cycle, it is crucial to acknowledge that since the 1800s, human activities have been releasing greenhouse gases such as carbon dioxide (CO_2) and methane (CH4) into the atmosphere. This has led to global warming and the greenhouse effect [7,8,10,11].

These human activities include burning gasoline in cars, using fossil fuels like coal and oil for generating electricity, clearing land and forests that result in the release of stored CO_2, creating landfills, which are significant sources of methane emissions, agricultural practices including livestock production, and industrial processes [10,11,15]. The concentrations of greenhouse gases have surged to levels unseen in more than 2 million years, and emissions are still on the rise. Consequently, the Earth's average temperature has increased by approximately 1.1 °C since the late 1800s. The decade from 2011 to 2020 has the distinction of being the warmest on record. Interestingly, many people mistakenly associate climate change primarily with warmer temperatures [5]. The rise in temperature is just the beginning of the climate change story. Our Earth operates as a complex system where all components are interconnected, meaning that changes in one area can trigger cascading effects across the entire system.

Despite the host of critical challenges that climate change poses, there is limited knowledge regarding the issues, impacts, and mitigation strategies at the micro level for various entities and organizations, including corporations and government agencies [16–21].

Acknowledging the significance of comprehending and addressing the effects of climate change, this exploratory study takes an innovative approach by utilizing machine learning and textual analysis methods [22–28] to analyze climate change litigation cases. The purpose is to derive insights into the phenomenon and key components of climate change utilizing the novel source of litigation cases, given the acknowledged scarcity of climate change data, as well as climate change disclosures within organizations in general [29–33].

Belal et al. found that while many companies in Bangladesh primarily reported information pertaining to the energy usage category, which is a mandatory disclosure, they provided minimal disclosure regarding other aspects of climate change, such as GHG emissions. [34]. Another study found that, while companies disclosed information regarding corporate governance, there was limited disclosure and insight concerning climate change risks and the potential for mitigation [35]. Yet another study has revealed that institutional investors have turned to private reporting processes as a means of mitigating the recognized deficiencies in public climate change reporting [36]. Nurunnabi discovered that, on average, Bangladeshi companies provided climate change-related information amounting to 2.23% [37]. Surprisingly, even multinational corporations did not meet the expected standards for satisfactory disclosure [37]. Rouas explored access to justice and corporate accountability in Europe, specifically within the context of how effective litigation had been against multinational enterprises and found that there is a need for increased commitment from multinationals in general [38]. In the U.K., despite the Climate-related Financial Disclosures (TCFD) framework requiring companies to report on their strategies for addressing climate change risks and opportunities, there is potential to stimulate increased disclosure [39]. In Bangladesh, while the mean disclosure index for all companies in general was low, the indices of those in industries with substantial pollution levels were lower than those in industries with minimal pollution levels [40]. Another study reported that there was a general lack of information about the contributions and management of climate change aspects [41], and an urgent need for more climate change reporting and due diligence [42].

Companies with more proficient managers are inclined to provide greater disclosures regarding climate change. This implies that across the board, mandatory disclosure is required to elicit more information regarding corporate climate change practices [43,44]. In

general, companies are reluctant to report on many aspects of their climate change [34,42] and, even if they did report, the disclosure is scant and sporadic [36,37,40]. Therefore, in summary, there is a dearth of information and insight regarding climate change issues and practices due to poor data availability. Climate change litigation cases have the potential to provide this information and fill the gap.

Climate change litigation, an emerging trend in recent years, stands as a pioneering solution to reshape the dynamics of this battle [45–51]. An escalating trend shows individuals resorting to legal action to address the climate crisis. Both private and public sector entities are facing escalating challenges and greater accountability [52–57]. Young people, women's organizations, local communities, Indigenous Peoples, and various other groups are assuming increasingly influential roles in initiating cases and propelling climate change governance reform in numerous countries worldwide [58–62]. The legal foundations for such cases are expanding. The Human Rights Council, as well as the General Assembly of the United Nations, have officially acknowledged the right to an environment that is unpolluted, conducive to well-being, and sustainable [4–6,63,64]. We are witnessing the emergence of fresh claims that revolve around the breach of laws pertaining to NetZero targets, assessments of environmental impacts, rising standards in advertising, and commitments outlined in the Paris Agreement [4,5]. Climate change litigation has established important precedents for climate action on a global scale, transcending the borders of their original jurisdictions and inspiring and propelling similar efforts in other nations [4,5,52].

The impetus for this research is drawn from several angles. First, climate change is deteriorating rapidly, and the effects are being felt in all walks of life all over the world, while simultaneously, the various stakeholders (individual citizens, companies, government agencies, non-profits, and other policy-making bodies, etc.) are desperate to obtain more information and insight. Second, it naturally implies that the numerous constituents and participants such as concerned citizens and activists, companies, regulatory agencies, and non-profit entities are actively seeking to shape policy and mitigate the threats of climate change. Third, while businesses are attempting to proactively disclose their climate change initiatives using mandatory and voluntary disclosure, studies have shown that these disclosures are lackadaisical at best. Fourth, it is universally recognized that studies on climate change and its impact at the granular level are still ad hoc and sporadic at best.

This current analytical study applies sophisticated machine learning and textual analytical methods to extract and thoroughly examine the most noteworthy climate change details in litigation cases in the courts. Though past studies have primarily focused on information sources that are easily accessible such as press publications, news dissemination agencies, corporate reports, and voluntary disclosure of climate change, ESG, and sustainability reports, our study delves into the specific climate change information embedded in the openly accessible litigation information in the different levels of courts and jurisdictions in the United States. It is noteworthy that legal documents, while perceived to be a legitimate source of information, are still subjective to an extent since they are, by and large, corpuses of textual data. The analysis of large amounts of text data requires the application of advanced machine learning and natural language processing. In analyzing and modeling legal cases related to climate change, our primary focus was on categorizing and detailing the different aspects of climate change, including their characteristics, as well as the relevant statutes and more.

We supplemented existing research on climate change as it relates to organizations and litigants in several ways. First, our study sought to analyze court filings (e.g., affidavits, verdicts, etc.). Therefore, our insights were derived from more compelling sources. Second, we leveraged the most recent data present in the legal documents, with the latest data being from 2021, encompassing a substantial dataset spanning an entire decade. Through an extensive longitudinal study encompassing a bigger sample size of climate change cases, we expanded the scope of our research to facilitate analysis from multiple perspectives and methods. Additionally, this approach allowed us to highlight a time-modeled comprehension of the contemporary status of significant climate change issues. Moreover, the

dataset enabled us to explore the various facets that constitute climate change, including key climate change issues and categories, stakeholder demographics, the utilization of laws, financial aspects, and more. Third, our research modestly advanced our comprehension of the applications of machine learning, Natural Language Processing (NLP), as well as text analytics, within the expansive realm of climate change and the legal domain, particularly when dealing with extensive text datasets. Fourth, our research shed light on climate change from a legal lens and surfaced the issues the key parties to the litigation were fighting about. The multiple stakeholders can leverage the insights obtained from this research, which is expected to promote more effective mitigation and prevention strategies to address the impacts of climate change. This, in turn, is likely to strengthen the overall response to climate change.

The rest of this paper is organized as follows: Section 2 discusses the concepts of climate change, climate change litigation, and the methods of machine learning and natural language processing text analysis. ML-based text analytics in terms of the methodology. Section 3 outlines the methods employed. Section 4 offers the associated results and analysis. Section 5 provides a comprehensive discussion of the scope and limitations of the study. Finally, Section 6 concludes with suggestions for future research.

2. Research Background

2.1. Climate Change

Climate change pertains to long-term alterations in weather patterns and temperatures. These shifts may occur naturally, stemming from variations in the sun's activity or significant volcanic eruptions [7,8,15,65]. However, since the 1800s, human activities have become the predominant drivers of climate change, chiefly because of the combustion of fossil fuels such as oil, gas, and coal. The burning of these fossil fuels produces greenhouse gas emissions, which function like a blanket enveloping the Earth, trapping heat from the sun and leading to a rise in temperatures [7,8,15,25]. The primary greenhouse gases responsible for driving climate change are carbon dioxide and methane. The sectors that are key to greenhouse gas emissions include energy, buildings, transportation, agriculture, land use, and industry [10,11]. People are encountering the effects of climate change in a multitude of ways, impacting various life aspects, including health, food production, housing, safety, and employment. Certain populations, such as those residing in small island developing States, are already more susceptible to the repercussions of climate change [4,5]. Conditions such as rising sea levels and saltwater intrusion have progressed to the extent that entire communities have been compelled to relocate. In the years to come, it is anticipated that the number of climate refugees will be on the increase [4,5,66]. Therefore, every increment of global warming holds significant importance.

In a 2018 report, a consensus among thousands of scientists and government reviewers concluded that constraining global temperature increases to a maximum of 1.5° centigrade would serve as a crucial measure in averting the most severe climate-related consequences, and in preserving a habitable climate [7,8]. However, if carbon dioxide emissions continue their current trajectory, global temperatures could potentially rise by up to 4.4 °C by the close of this century. Emissions responsible for climate change originate from all regions across the globe and impact people worldwide. However, certain nations contribute significantly more than others. In fact, the 100 countries with the lowest emissions collectively account for just three percent of the total emissions [5], while the 10 largest emitters are responsible for 68 percent of emissions. While climate action is a shared responsibility, those individuals and nations contributing more significantly to the problem bear a greater responsibility to take the lead in addressing it.

Climate change presents a formidable challenge, but, on the bright side, numerous solutions that can yield economic advantages, enhance quality of life, and safeguard the environment have already been identified [25,67–72]. There are also international agreements in place to steer our collective efforts, including the Paris Agreement and the UNFCCC (United Nations Framework Convention on Climate Change) [4,5]. Three

overarching categories of action include adapting to climate impacts, reducing emissions, and financing required adjustments. Transitioning from fossil fuel-based energy systems to renewables, such as solar power, will mitigate the emissions contributing to climate change. The urgency of beginning these actions cannot be overstated [4,5]. While an increasing number of countries are pledging to achieve net-zero emissions by 2050, it is essential to recognize that approximately half of the necessary emissions reductions must be achieved by 2030 to limit global warming to under 1.5 °C [7,8]. Achieving this goal entails an approximate six percent annual reduction in fossil fuel production from 2020 to 2030 [7,8]. Adaptation efforts will be necessary worldwide, but there is an immediate need to prioritize those who are most vulnerable and have the fewest resources to address climate-related risks [5]. The potential return on investment can be substantial. For example, the implementation of early warning systems for disasters not only saves lives and property but can also yield benefits up to 10 times the initial cost [4,5]. The choice is to invest in proactive measures now or face significantly higher costs in the future. Addressing climate change necessitates substantial financial commitments from both governments and businesses. However, the costs of climate inaction far outweigh these investments. A crucial measure is for industrialized nations to honor their commitment to provide $100 billion annually to developing countries, enabling them to adapt and transition towards more sustainable economies [7–9,11,12].

However, three significant limitations exist in the literature. Firstly, despite the ample coverage of climate change in both academic literature and the media, there is a surprisingly limited body of research dedicated to thoroughly examining the phenomenon at the granular level of companies and entities to accurately assess associated risks and trends. Unless more detailed information regarding climate change is available at the organizational and grassroots level, these entities are unable to take decision steps to mitigate climate change [5,7,8]. Secondly, there is a scarcity of studies that have applied data analytic techniques like machine learning and text analytics for conducting descriptive analyses of granular text data. Thirdly, the limited studies available tend to be more conceptual and less empirical. This study aims to address these gaps.

2.2. Climate Change Litigation

Over the past few years, there has been a substantial increase in climate litigation on a global scale, encompassing a broader array of legal theories and spanning diverse geographical regions [6,64]. This surging wave of climate-related lawsuits is instigating essential transformations. Climate litigation is pressuring corporate entities and governments to adopt high-reaching goals for both mitigating and adapting to climate change. An emerging and noteworthy trend involves cases that prioritize fundamental human rights related to a stable climate. Additionally, there is an increasing number of cases on the right to a healthy environment, a right enshrined in the constitutions of more than 100 countries. These cases are compelling enhanced climate-related disclosures and putting an end to deceptive corporate greenwashing on climate change. Citizens are demanding accountability from their governments, striving to prevent further extraction of fossil fuels and contesting the lack of enforcement of climate-related laws and policies [4].

As part of this wave, more citizens and organizations around the world are going to court to seek a fair judgment in climate change law cases, and the number of cases brought against climate change inaction has increased dramatically [2]. For instance, while the Clean Air Act (CAA) empowers the U.S. Environmental Protection Agency (EPA) to regulate emissions of both carbon dioxide and other air pollutants, non-governmental organizations have resorted to legal action to compel the EPA to fulfill its obligations in safeguarding public health from air pollution. Additionally, they have initiated legal actions against entities believed to be breaching relevant emission standards or permit regulations. The British Institute of International and Comparative Law research project examines climate litigation globally and produces a toolbox for implementing climate law.

This shows that there are at least 2000 climate change litigation cases filed globally since November 2022 [73].

Climate change litigation offers civil society, individuals, and various stakeholders a potential avenue to confront insufficient responses from the private sector and governments in dealing with the climate crisis. In climate-related cases, individuals or parties referred to as plaintiffs employ diverse legal tactics across various national and international jurisdictions. Their primary aim is typically to compel the public and private sectors to adopt more ambitious goals for both mitigation and adaptation. Nonetheless, there are instances where plaintiffs may also aim to contest climate laws and lower climate objectives. In its Sixth Assessment Report, the Intergovernmental Panel on Climate Change (IPCC) acknowledged that climate litigation has, for the first time, impacted the results and level of ambition within climate governance [7,8,66]. The IPCC has also recognized climate litigation as a significant channel through which stakeholders can influence climate policy beyond the formal UNFCCC processes [7,8,66]. Furthermore, winning cases pursued by plaintiffs have inspired the initiation of analogous claims in different legal jurisdictions. As an example, the ruling in the Urgenda Foundation v. State of the Netherlands case, where a court held a government accountable for greenhouse gas emissions mitigation, has catalyzed a series of ambition-driven cases in other countries. Many of these cases explicitly reference the Urgenda decision even though it lacks legal authority beyond the Netherlands [74]. In a separate instance, a cohort of young individuals in Montana achieved a groundbreaking legal victory when a judge ruled that it was unconstitutional for the state to approve fossil fuel projects without considering climate change [59].

The scope of climate litigation will continue to broaden as research on climate science expands, and new legal theories get explored nationally and internationally. [5]. Each passing year sees climate change litigation assume a progressively vital role, either driving forward or hindering substantial action on climate change. The Intergovernmental Panel on Climate Change (IPCC), in 2022, acknowledged that litigation has an impact on shaping the outcome and the level of ambition in climate governance [7,8,75]. The Global Climate Litigation Report: 2023 Status Review reveals that by December 2022, a total of 2180 climate-related cases had been submitted across 65 jurisdictions. These encompassed tribunals, international/regional courts, quasi-judicial bodies, and other adjudicatory entities, including Special Procedures at the United Nations and arbitration tribunals [5]. This marks a consistent rise in case numbers from 884 to 1550 between the years 2017 and 2020. Notably, local communities, women's groups, children and youth, and Indigenous Peoples are assuming a significant role in initiating these cases and spearheading reforms in the governance of climate change worldwide [5]. In 2019, the body of climate litigation literature saw considerable expansion, with notable attention directed toward new landmark judgments, emerging legal pathways, diverse actors involved, shifting litigation objectives, and an extended range of jurisdictions [63]. A recent comprehensive review systematically examines significant literature on climate litigation released from 2000 to 2018 [76].

This study adopts the definition of climate change litigation used by the Sabin Center in the creation and upkeep of its databases. According to this definition, climate change litigation encompasses cases that present significant legal or factual matters pertaining to adaptation, mitigation, or the scientific aspects of climate change [64]. These cases are presented to various administrative, judicial, and other decision-making bodies. The Sabin Center generally identifies cases through keywords such as 'climate change', 'greenhouse gas', 'global change', 'global warming', 'sea level rise', and 'GHGs'. Additionally, cases that address legal or factual aspects of climate change but do not explicitly employ these specific terms are also encompassed [6,64]. The domain of climate litigation is expanding, with a noticeable increase in both the volume of case filings and the range of jurisdictions where they have been presented over the last few years [4,5].

Considering the paucity of information on corporate ESG and climate change/ sustainability practices, the application of machine learning and text analytics to ana-

lyze textual narratives from legal cases holds promise for yielding valuable insights into the climate change phenomenon.

2.3. Text Analytics and Machine Learning

The utilization of text analytics as a machine learning technique has become widespread, thanks to the growing accessibility of electronic documents from various sources [23]. In the realm of electronic data, there are two primary categories: structured and unstructured. Structured data is characterized by a clear organization and ease of searchability, whereas unstructured data is more intricate, lacking a defined structure, and encompasses various types such as audio, video, and graphics [77]. Given the ubiquity of unstructured data, the retrieval of knowledge from these sources holds significant importance within both research and practical applications. By harnessing NLP, text analytics has the capacity to transform unstructured data into a structured format, enabling its analysis and utilization in conjunction with machine learning algorithms. Furthermore, the utilization of text analytics enables researchers to assess various aspects of fundamental concepts. Many text analytics investigations rely on dictionaries resembling thesauruses, which consist of words or phrases with shared meanings [78]. To analyze a collection of texts, this approach entails examining the frequencies of entries and categories while assessing the significance of key concepts within the text. The key benefit of text analytics lies in its ability to process extensive volumes of data [79]. In the present research context, investigating court cases related to climate change litigation presents a suitable domain. Firstly, it constitutes a substantial collection of unstructured text. Additionally, it encompasses significant factual information, legal elements, and precedents related to climate change subjects, serving as a relevant repository for exploration. Hence, climate change litigation emerges as a significant research domain worthy of exploration. The text analytics methodology has been applied in research using unstructured data, including the analysis of vaccination-related tweets [80], legal pharma patent validity [81], health blogs [82], shareholder resolutions on sustainability [83], and understanding corporate sustainability disclosures [84] among others. A few studies, for example, have applied machine learning and text analytics in climate change research [22,24,26–28,85].

This exploratory research utilized machine learning-based text analytics for climate change litigation legal cases. The frequency distribution of word counts and the identification of common words in each case were visualized. Following that, word clouds based on Text Rank [86] and Term Frequency were created. This was followed by topic modeling, which included the Latent Dirichlet Allocation (LDA) Gensim and LDA Mallet Model [87–89] to explore the topic modeling distribution and sub-topic visualization. Next, Word2vec was applied to average word embeddings [90], K-Means models were constructed for clustering, and LDA was employed to extract topic words within each cluster. In this part, Spacy and Regex expressions were utilized to extract frequent law acts and inference cases. Additionally, Bi-grams and Tri-grams were generated, and keywords were examined. The TF-IDF Document Similarity method was also employed as a valuable tool to assess document similarity [91].

3. Methodology

This exploratory study scrutinized climate change court cases to extract insights regarding the diverse categories of climate change, applicable laws, practices, and evolving trends by employing machine learning-based text analysis methods [77,92,93]. The primary source of the climate change court cases was Westlaw, the legal database (https://legal.thomsonreuters.com/en/westlaw (accessed on 20 October 2023)). Westlaw is an online legal research service and a proprietary database that encompasses over 40,000 case laws, and state and federal regulations. The text analytics approach based on machine learning facilitates the efficient processing and examination of extensive textual data [94–96]. The findings from this study served as an alert to inform a wide range of parties, including company management, stakeholders, climate change experts, policymakers, activists,

consumers, government and regulatory bodies, and NGOs, regarding the recognition, mitigation, prevention, and the implications and future trajectory of climate change.

Using the online query search, approximately 2656 cases were identified and obtained from the Westlaw database for the period 1 June 2019 to 31 May 2021. The Python selenium package (pypi.org/project/selenium/ (accessed on 6 June 2022)) was further utilized for scraping and transforming the pdf files, a file for each case. Several cases were redundant in the results of the query search. Moreover, numerous PDFs lacked complete case descriptions with decisions; instead, they consisted of supplemental filings with the court. Additionally, many others were unrelated, such as instances where 'climate change' was mentioned in different contexts, like divorce-related cases. Also, footnotes are present on every page, such as Chief Justice John Roberts, etc. Therefore, the stop words also included the words appearing frequently in the footnotes. Legal terms such as 'court', 'defendant', and 'plaintiff' were also included in the dictionary. The unconnected hyperlinks were also removed to clarify the documents. A total of 2252 cases' narratives were eliminated due to the above-mentioned reasons, resulting in a final data set of 404 cases. These 404 cases were then transformed into text strings in Python. The Pickle utility (https://docs.python.org/3/library/pickle.html (accessed on 11 July 2022)) was employed to locally store the text data for the purpose of later reloading it for analysis. Figure 1 outlines the overall methodology and the various ML and textual analysis algorithms and methods utilized in this research.
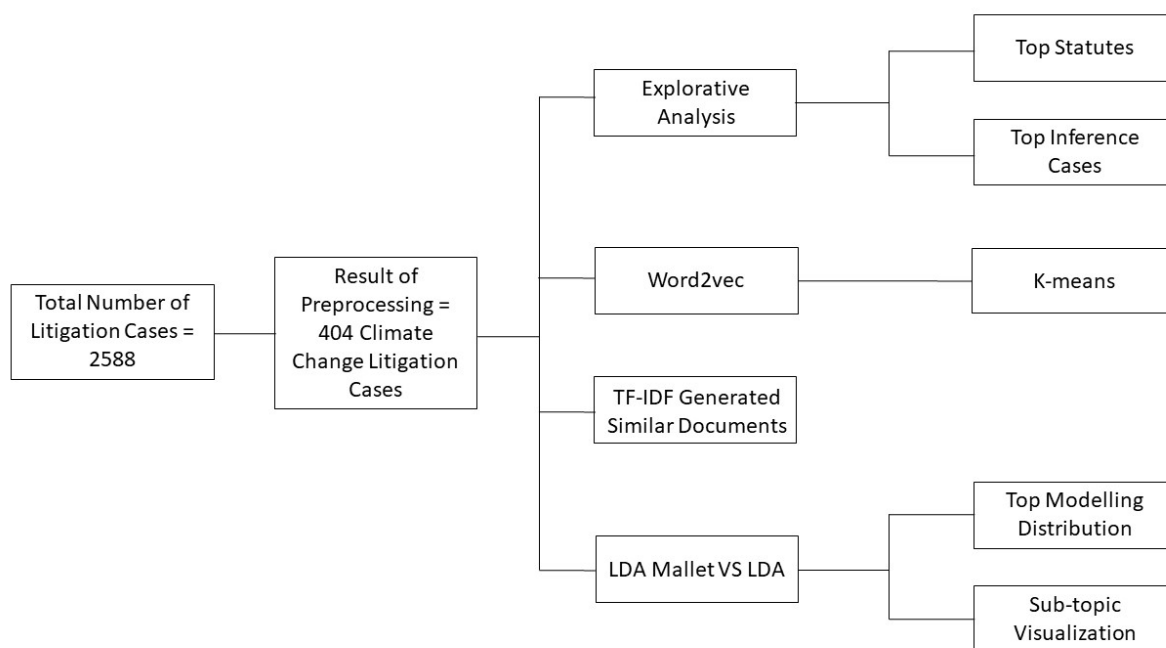

Figure 1. Methodology.

The data contained approximately 4,609,307 words, averaging 11,409 words per case narrative. The median number of words among the 404 cases was 9254. Through descriptive analysis, Figure 2 presents a bar chart illustrating the frequency of keywords found in these climate change cases. The term "Environment" was the most frequent, occurring 8336 times, followed by "Water" at 6202 occurrences. "Area" appeared 6168 times, and "Forest" was mentioned 5681 times. Following these, the words "Habitat" and "Land" appeared quite frequently. Together, these words indicate that many of the climate change litigation cases dealt mostly with environmental issues in the context of land and water use habitats and forests.
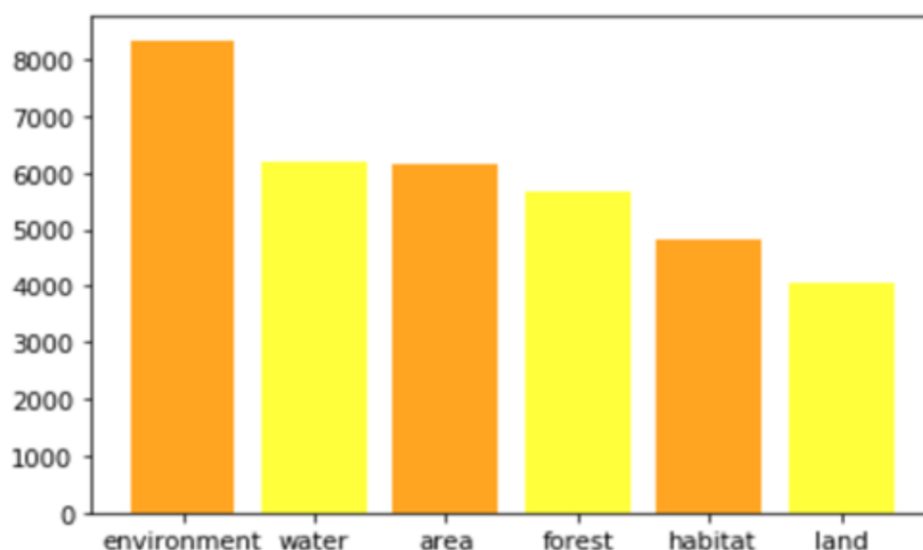
Figure 2. Keyword frequency in climate change.

3.1. Text Analytics

Subsequently, machine learning-based text analytics were employed to examine climate change litigation cases. Pre-processing was done to the data before applying the text analytics. The retrieved data for each case was saved as a text file before vectorization. A few cases were removed due to relevance, link functionality, or clarity issues. The creation of predictive models using text data introduced some challenges to the process of modeling. First, it is important to note that textual data is not suitable for input in many mathematical models. Consequently, an implementation of an NLP system was employed to convert the text into essential components for subsequent analysis. Second, text-based datasets tended to be more extensive in size compared to numerical datasets. Consequently, developing an effective model required the extraction of pertinent information by identifying key data points. In the data pre-processing stage, the summary section that was scraped was transformed into plain text documents. Redundant items like numbers, spaces, punctuations, and standard stop words were eliminated using the Natural Language Toolkit (NLTK). Subsequently, the text was transformed to lowercase, utilizing NLTK and TextBlob (https://textblob.readthedocs.io/en/dev/ (accessed on 20 October 2023)). Afterward, lemmatization was employed to transform words into their root forms, such as substituting "bought" and "buying" with "buy". Lemmatization groups inflected various forms of a word together, facilitating their analysis as a single term and providing a contextual understanding of the words. Using the pandas package (https://pandas.pydata.org/ (accessed on 20 October 2023)), the scraped data was filtered and processed into appropriate data frames that were suitable for analysis. The sklearn package (https://scikit-learn.org/stable/index.html (accessed on 20 October 2023)) was employed for result refinement. The number of features was set to a maximum of 4000. Tokenized words comprised those having more than four characters. These steps helped decrease the impact on the generated model downstream. LDA was applied for the purpose of identifying themes and their distribution in large corpora [96–98]. This approach helped analyze the documents without relying on a pre-defined list of terms, providing a more total perspective regarding the content of climate change court cases compared to prior studies [99–101]. The 'term frequency-inverse document frequency' (TF-IDF) method [99–101] was employed to compute a term's weight and its significance within a document. TF-IDF assigns the weight of a term based on its frequency of occurrence (TF) and inverse document frequency (IDF). Each term was assigned both of these scores, and the term's weight was subsequently computed by multiplying these scores together. Further details about this technique can be found in the Results section. The K-Means clustering model was used to uncover the key climate change concepts. K-Means

clustering is a well-known machine learning algorithm that uses similarity measures to classify cases, specifically the distance between cases. It finds applications in fields like pattern recognition and classification. There were four clusters created with the word cloud package. Subsequently, a KNN classifier was used to perform the classification, in line with the supervised approach to machine learning. The data was split for training and testing to evaluate the classification.

4. Results and Analysis

This section showcases the outcomes of the machine learning-based text analytics applied to the climate change litigation dataset. First, a word cloud analysis was conducted, generating the most frequently occurring keywords and word clouds to obtain a general overview of the data. This was done using Text Rank and Term Frequency approaches [102]. The word clouds visually represented the most frequently occurring words within the corpus of cases. These word clouds were generated using wordclouds.com, where the size of each word corresponded to its frequency in the corpus. The results confirmed our existing understanding, which was based on anecdotal evidence and preliminary manual examination and analysis of the underlying inferred issues in the court cases. Following this, cluster analysis was carried out on the central themes and related subjects to pinpoint fundamental concepts and critical factors within each cluster. The Word2vec models that have been developed were used as inputs for the K-Means model. The two primary machine learning models that were utilized encompass clustering and LDA [103,104]. To ensure the reliability of the results, both LDA and LDA Mallet were employed to delve into the topics discussed in the documents. Additionally, TF-IDF was employed to retrieve documents with similar content [91,99,105].

4.1. Word Cloud

The word cloud visualizations in Figures 3–5 depict the words that appeared most frequently within the body of litigation cases. Text Rank and Term Frequency models were employed to generate word cloud maps. Text Rank, a versatile graph-based ranking algorithm commonly used in NLP, assesses the significance of text segments within a document by recursively considering information from the entire document [102,106]. Term Frequency indicates how often a specific piece of text appeared in the entire document. Three word-count methods were employed to identify the significantly relevant pertinent information in all case files: text rank splitting by page, text rank splitting by case, and term frequency. Given that a single case consists of various sections, including background information (such as holdings, attorneys, and law firms), pertinent concepts (like petitioners' allegations and standing), and judicial opinions, dividing cases by page can assist in excluding background information, which predominantly comprises less relevant content in terms of word count. Consequently, separating cases by page was expected to enhance the retrieval of crucial information related to the relevant concepts within the cases. Figure 3 displays a word cloud based on the text rank approach split by page. All cases were individually split by page. Subsequently, the Text Rank algorithm was applied, and common stop words like 'itself' and 'he' were removed before creating the word cloud. As demonstrated, one of the significant keywords identified across all cases was NEPA—the National Environmental Policy Act ('NEPA'). NEPA is a pivotal U.S. environmental law that was enacted on the first of January 1970, aimed at fostering environmental improvement. It also established the President's Council on Environmental Quality (CEQ) [107]. The cases also highlighted keywords such as 'biological diversity' and 'water'. In addition, terms like greenhouse gas emissions, CEQA, and EPA held relative significance compared to other words. CEQA represents the California Environmental Quality Act that was passed in 1970, following the passage of the NEPA by the U.S. federal government. CEQA's primary purpose is to establish a comprehensive statewide environmental protection policy [108]. EPA stands for the United States Environmental Protection Agency. In summary, the prominent topics in the cases included 'biological diversity', 'greenhouse gas emissions',

'habitat', and 'water', as well as 'wildlife'. Additionally, the two prominent statutes were 'NEPA' and 'CEQA', with the state of 'California' being the primary location for climate change litigation.

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

Figure 4 displays the word cloud generated using the Text Rank approach when cases are separated. Compared to the word cloud by page in Figure 3, this 'split by case' approach provided less insight. It seemed that isolating the actual case content (separating cases by page rather than by case) results in Text Rank by page, revealing more information. In Figure 4, the most prominent words were 'climate change', followed by 'impacts', 'agency action', etc. Additionally, 'greenhouse gas emission', 'judicial review', 'claims', 'district court', and others were significant keywords. This word cloud further reinforced California's status as the foremost state in climate change litigation, suggesting its leadership in endeavors related to climate change mitigation.

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

Regarding the Term Frequency word cloud, Figure 5 highlights common litigation and legal terminology within the realm of climate change litigation. Terms like 'action', 'agency', 'case', 'claim', 'defendant', 'government', 'issue', 'state', and others emerged. In summary, these three methods collectively revealed the keywords associated with climate change cases and illuminated the diverse concepts and vocabulary related to climate change. Nevertheless, keywords alone do not significantly enhance our comprehension of climate change. Therefore, next is the topic modeling using LDA and LDA Mallet [96,109,110].

Figure 5. Word cloud of term frequency approach.

4.2. Topic Modeling

This section applied the topic modeling analysis to find the most frequent content that occurred in the 404 climate-change-related cases. Topic modeling involves the unsupervised learning process of automatically identifying topics within a set of documents. It entails analyzing the documents themselves to uncover hidden structures such as topics, per-document topic distributions, and per-document per-word topic assignments [96,98,100]. Document representations in these semi-automated methods are typically in vector form. In the simplest form, each vector contains the frequency of each term in the document.

However, this type of vector results in a large number of dimensions, with each dimension corresponding to a unique term. Therefore, it is necessary to reduce the dimensionality of these vectors [111] to handle the extensive dataset effectively. LDA was used to accomplish this. In the generated LDA vector, every dimension represented a specific concept or topic [111]. A topic represented the probability distribution of all the terms that occur together in the underlying documents [111]. The document itself represented the probability distribution of all topics in the corpus [112,113]. In essence, an author could describe a topic by selecting words with a certain probability from a word pool relevant to the topic [113]. For example, when discussing climate change, terms such as carbon dioxide, GHG, climate, warming, emissions, and temperature were highly likely to appear, whereas terms like gains, social responsibility, or employee benefits had a lower likelihood. Topics were determined by examining the frequent co-occurrence of terms. Therefore, if terms appeared together frequently in a document, it is highly possible that they were associated with that same topic. Each litigation case encompassed multiple topics, and the probability distribution of a particular document indicated the significance of the identified topics in that specific case. The next step was to tokenize the documents by breaking them into tokens comprising words and special symbols such as punctuation marks. The text was standardized by converting all characters to lowercase and eliminating special characters and numbers. Following that, the text underwent lemmatization using the WordNetLemmatizer V3.1.1. Standard stop words (like articles, pronouns, and conjunctions) were removed. The NLTK package in Python provided the "English" stop word list used for this purpose. Additionally, terms appearing in fewer than two documents were excluded. The remaining vocabulary was manually reviewed to eliminate additional irrelevant terms, such as individual names. The LDA process aimed to identify a blend of topics in each document, with each topic characterized by a combination of terms [113]. Consequently, the probability distribution for the combination of topics was distinct from the distribution for the combination of terms 113]. The hyperparameter α defines the shape of the distribution of topics within each document, while the hyperparameter β defines the shape of the distribution of words within each topic [114]. The algorithm estimated these distributions using Dirichlet priors [113]. There are efficient and effective implementations of LDA, such as Gensim for Python and Mallet for Java. [111]. After preprocessing, the next step involved running the model. Initially, the widely utilized model LDA was employed. LDA is a probabilistic generative model used for analyzing collections of discrete data, like text corpora [112]. It treats case document files as collections of words, organizing them into word bags where sets of keywords are grouped together to form topics. Each case word bag may contain multiple topics, and the number of topics within each case can be customized. Multiple parameters can be adjusted and the default maximum number of iterations was set to 50, along with the hyperparameters α and β, which defined the structure of the per-document topic distribution and per-topic word distribution. Each training chunk contained 1000 data files selected from a total of 404 cases utilized. Moreover, there were ten complete passes through the corpus during training. Additionally, the model generated a list of topics ranked in descending order based on their likelihood for each word, including their phi values multiplied by the feature length (i.e., word count) when the parameter per_word_topics was set to True. An initial attempt at 20 potential topics was made. To assess model efficiency, the Cross Validated (C_V) coherence score, which quantifies the frequency of co-occurrence of words belonging to the same topic in the corpus, was used as the standard. The C_V coherence score relied on four criteria: data segmentation into word pairs, computation of probabilities for individual words or word pairs, evaluation of a confirmation measure assessing the strength of one word set's support for another, and the aggregation of individual confirmation measures into an overall coherence score [115]. A coherence score of 0.39 was achieved. The coherence score lacked a standardized value as it varied with the corpus size. In this case, the LDA model's visualization did not meet expectations. In Figure 6, depicted below, the gap between the bubbles reflects the similarity between topics based on word distribution. The size of the bubbles indicates the topic's

prevalence within the corpus. As topic modeling aimed to reduce overlap among each node (represented by the bubbles in the chart), it was desirable for these bubbles (sub-topics) to be distributed as widely apart as possible, ideally spanning the four quadrants (PC 1–PC 4) on average. However, when certain nodes cluster closely together, effective differentiation between groups of cases becomes challenging.
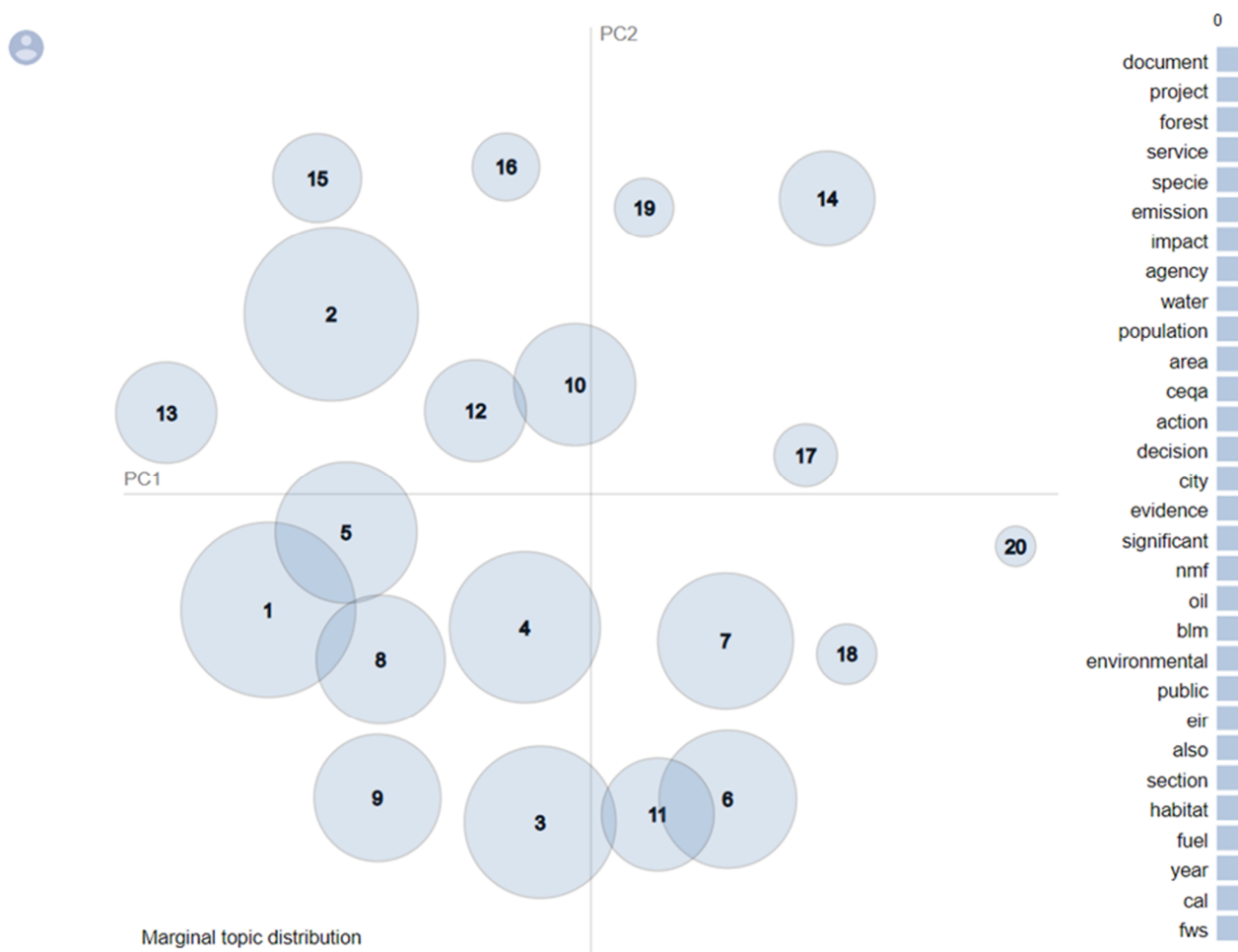

Figure 6. Distribution of LDA model with 20 topics.

Consequently, the Mallet implementation, with its automatic estimation of hyperparameters α and β, was utilized to perform the LDA analysis. The LDAMallet model was brought in with the goal of achieving the highest correlation with all the available human topic ranking data when developing the coherence score [115]. Mallet is a toolkit that is based on Java, designed for statistical natural language processing, encompassing tasks such as information extraction, clustering and classification, topic modeling, and various other text-related machine learning applications [88]. This toolkit can be applied to unlabeled text analysis, particularly in the context of topic modeling. Most of its parameters can be left at their default values, with the exception of the number of topics. The predetermined number of topics depends on the degree of topic specialization that is desired. [113]. The intention of the current study was to ensure that each resulting topic was assigned an appropriate label. However, cases with fewer dimensions tended to have more generalized topics since the wide assortment of terms restricted the specificity of labels that could be assigned. Conversely, cases with many dimensions get highly specific labels. The algorithm was applied to various sets of dimensions, and the results were compared. The decision was made to concentrate on 70 dimensions, as this count provided a diverse range of topics without delving excessively into specifics. The algorithm generated two sets of results

for each topic. The initial result set included all terms within the corpus, along with their likelihood of contributing to the topic [113]. The next result set encompassed all documents in the corpus, along with the probability of occurrence of the topic within each document. Even though the number of topics remained at 20, the coherence score saw an improvement from 0.39 to 0.45. The LDAMallet model performed better than the regular LDA model. When interpreting the results, it is common practice to scrutinize from about 5 to about 20 of the most probable terms for a topic. This examination helped determine the level of shared characteristics and, consequently, assisted in defining the topic's label. [111]. As the number of topics significantly impacts the coherence score, an experiment was conducted with various numbers of topics to compare their respective coherence scores. The model was executed with a range of topic counts spanning from 5 to 50. The coherence line chart is shown in Figure 7. The number of topics indicates the frequency of nodes (sub-topics) employed within each topic in the LDAMallet model.
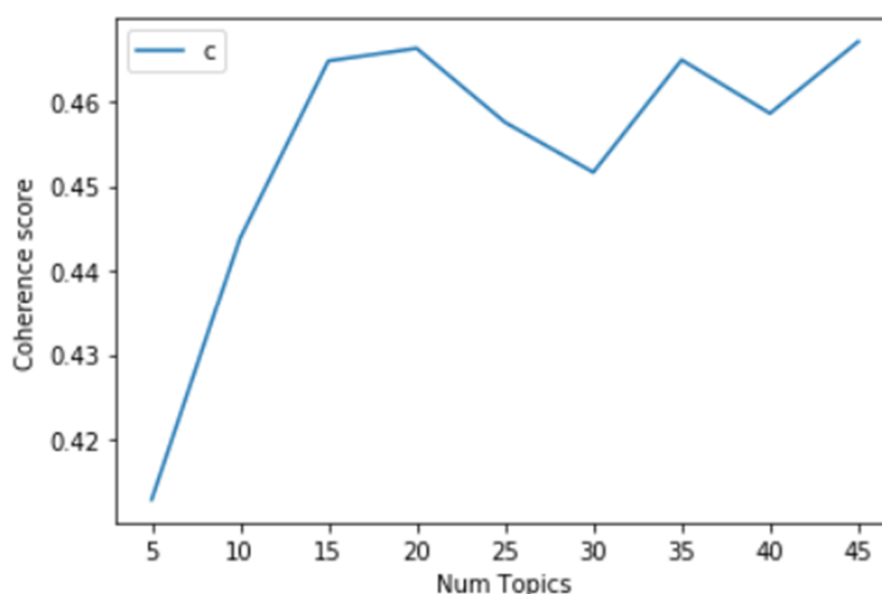
Figure 7. Coherence score for a number of topics.

Figure 8 illustrates a gradual rise in coherence scores as the number of topics increases, reaching its peak at 20 topics before tapering off. Notably, at 20 topics, there might be an overlap in the bubbles.

```
Num Topics = 5   has Coherence Value of 0.4129
Num Topics = 10  has Coherence Value of 0.4438
Num Topics = 15  has Coherence Value of 0.4648
Num Topics = 20  has Coherence Value of 0.4663
Num Topics = 25  has Coherence Value of 0.4576
Num Topics = 30  has Coherence Value of 0.4516
Num Topics = 35  has Coherence Value of 0.4649
Num Topics = 40  has Coherence Value of 0.4586
Num Topics = 45  has Coherence Value of 0.4671
```

Figure 8. Coherence score for a number of topic iterations.

Therefore, a conservative choice was made in selecting 15 as the number of sub-topics. Figure 9 illustrates the most commonly occurring terms for these 15 topics. It has keywords like 'water', 'emission', 'forest', 'specie', 'city', 'ceqa', (refers to California Environmental Quality Act), 'greenho_gas', 'oil', etc.
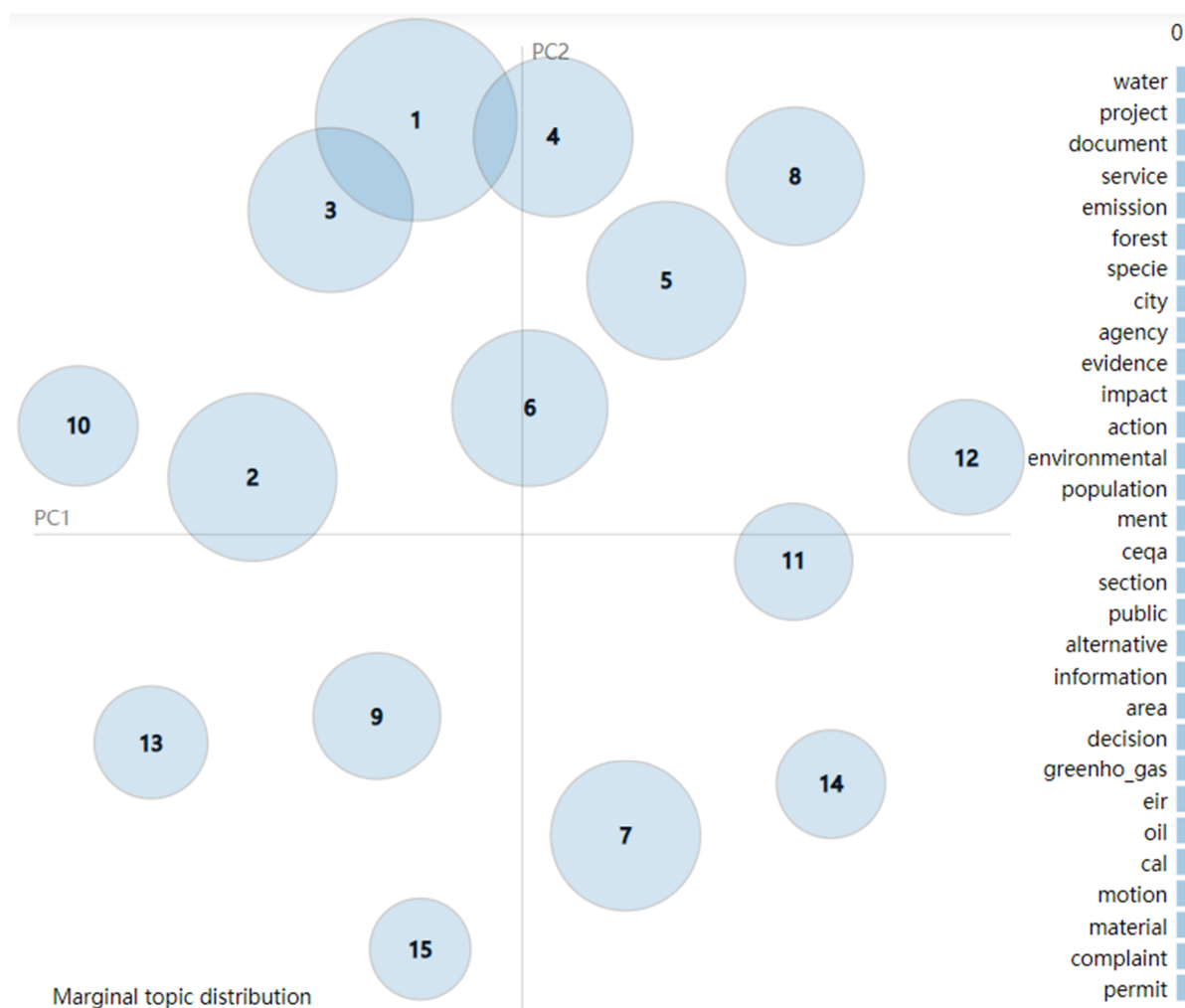
Figure 9. General distribution of 15 topics in LDAMallet model.

The distinct nodes represent cases that consist of various keywords, with each node containing specific meaningful keywords. Figure 10a below shows the multidimensional scaling focused on node 3, while Figure 10b displays the red bar for its keywords, illustrating the frequency of these keywords within a particular node compared to the overall term frequency. Keywords with red bars that span a significant portion of the blue bars were regarded as genuine keywords that exclusively appear and cluster within their respective nodes. It is worth noting that node three primarily focused on "oil, gas, coal, mining, and resource".

Examining Figure 9, node 1 stands out for "ceqa (California Environmental Quality Act)", node 4 emphasizes "fire, forest, tree, habitat", and node 5 focuses on "fish, survival, recovery, salmon". Node 6 is linked to "species, conservation", while node eight is associated with "water, reclamation, groundwater, flow". Node 9 encompasses "emission, greenhouse gas, fuel, LCFS (low carbon fuel standard), ethanol", and node 13 is noteworthy for "damage, pollution", among the more prominent nodes. These nodes suggested that the set of 404 cases could be categorized into eight distinct groups related to climate change: CEQA, oil/gas/coal resource emissions, forest fires, salmon survival and recovery, species conservation, reclamation, greenhouse gas emissions, and pollution damage. Further to the visual representation of topic modeling, it is possible to display the keywords directly. The fundamental concept behind topic modeling is to identify the topic frequency that holds the highest percentage contribution within a document. Initially, keywords were extracted from each sentence, and their frequencies were added to the entire case. In reverse, after confirming the topic for each sentence, the Pandas "group by" function was

employed to identify the representative case for each sub-topic. As an illustration, the most common keyword set could be relevant to 150 out of the 404 cases, consisting of terms like "action, agency, impact, environmental, decision, alternative, effect, land, analysis, and area". To identify the exact topic distribution across all 404 cases, the topic distribution was calculated as the topic_counts divided by the sum of topic_counts. This allowed us not only to gain a comprehensive understanding of the overall topic modeling results but also to extract the key concepts within each individual case. For this purpose, the topic number with the most significant percentage contribution to each litigation case was determined. As an example, consider the keywords "water, action, species, agency, year, effect, fish, population, project, operation". These keywords made up 90% of the keyword composition in the case titled "United States District Court, E.D. California, The Consolidated Salmonids, San Luis & Delta–Mendota Water Authority; Westlands Water District, v., Gary F. Locke, as Secretary of the United States Department of Commerce; et al., Stockton East Water District, et al., v., National Oceanic and Atmospheric Administration, et al., Water Contractors, v., Gary F. Locke". In another case titled "United States District Court, N.D. California, Center for Biological Diversity, v., Office of Management and Budget, No. C07–4997", the keywords "service, forest, document, agency, species, area, project, habitat, decision, information" comprised 92.85% of the keyword content. A cross-section of these are displayed in Figure 11. By adopting this method, researchers can identify the sub-topic in any given case.

(a)
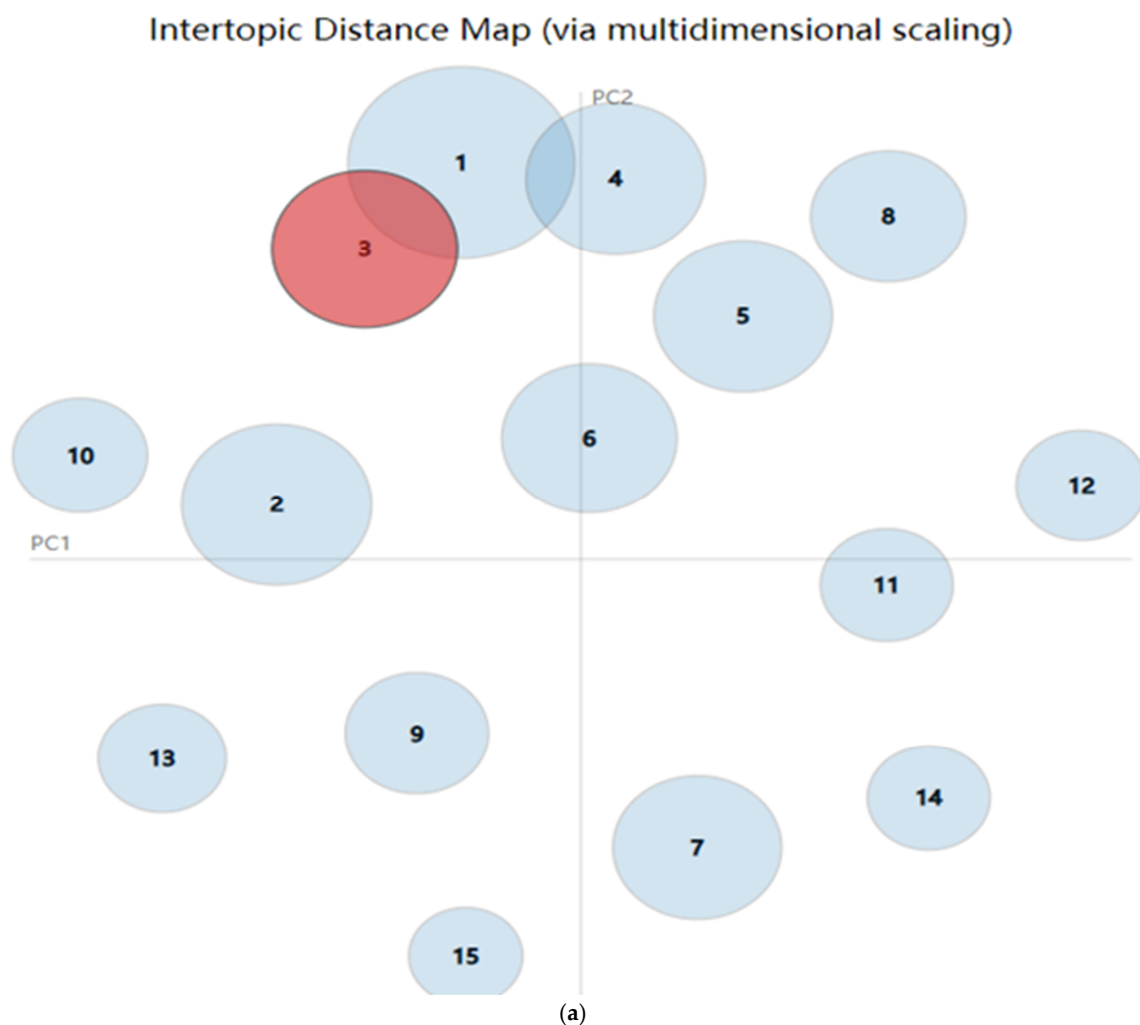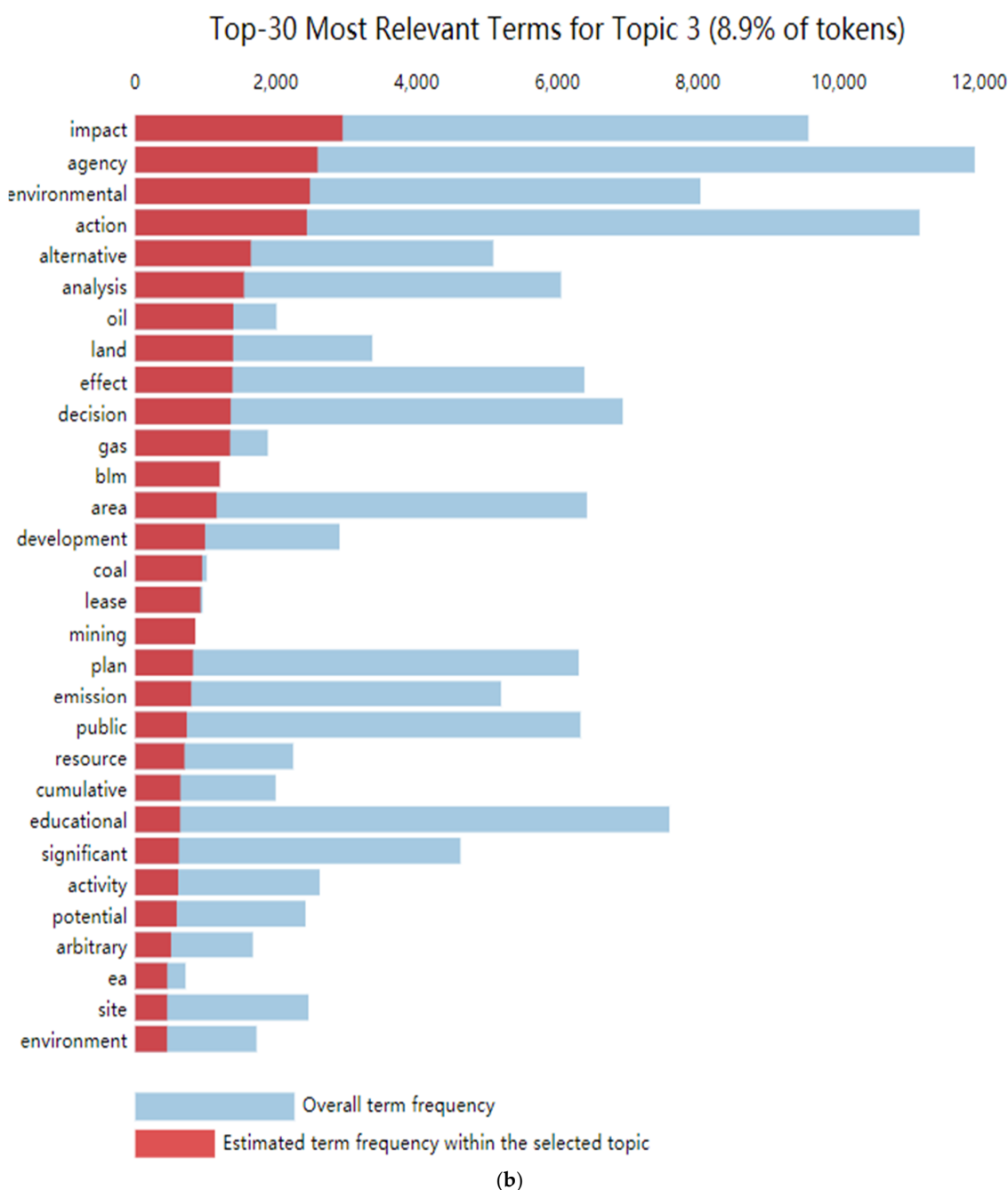
Figure 10. *Cont.*

Figure 10. (**a**) Multidimensional scaling focusing on node 3. (**b**) Distribution of keywords in LDAMallet node 3.

Figure 11. Examples of sub-topics in the first two litigation case files.

4.3. K-Means and Document Similarity

This section contains a description of the application of the word2vec model, K-Means, and TF-IDF for text analyzing litigation cases to gain further insight. As described in Section 3.1, TF-IDF is an approach that assigns a comprehensive weight to a term based on the combination of its frequency (TF) and the inverse document frequency (IDF). The K-Means clustering model was utilized to extract the primary concepts of sustainability. K-Means clustering stands out as a well-known machine learning algorithm utilized for categorizing cases by assessing their similarities, often measured as the distance between cases. It finds frequent applications in fields related to pattern recognition and classification. Subsequently, the data underwent classification using the KNN classifier, an algorithmic method within supervised machine learning. To evaluate the classifier's performance, the data was divided into two segments: one for training and the other for testing. Prior to applying these machine learning methods to textual data, bigrams were generated, and document similarities were assessed. This was essential since keywords alone may not significantly enhance our understanding of the nature and dimensions of climate change litigation cases. Hence, the co-occurrence of words was subsequently examined. In the field of linguistics, co-occurrence refers to the probability of two terms appearing in a specific order within a substantial corpus of data. It serves as an indicator of the semantic proximity of these terms [116]. This model provided insights into the associations between different issues.

Figure 12 presents a bar chart depicting the 30 most commonly occurring bigrams within the dataset. Examples of these frequent bigrams included phrases like 'air quality', 'gas emission', 'greenhouse gas', and 'forest service'. Additionally, it is noteworthy that 'government' and 'New York' had the highest frequencies, indicating a significant government involvement in climate change-related matters.
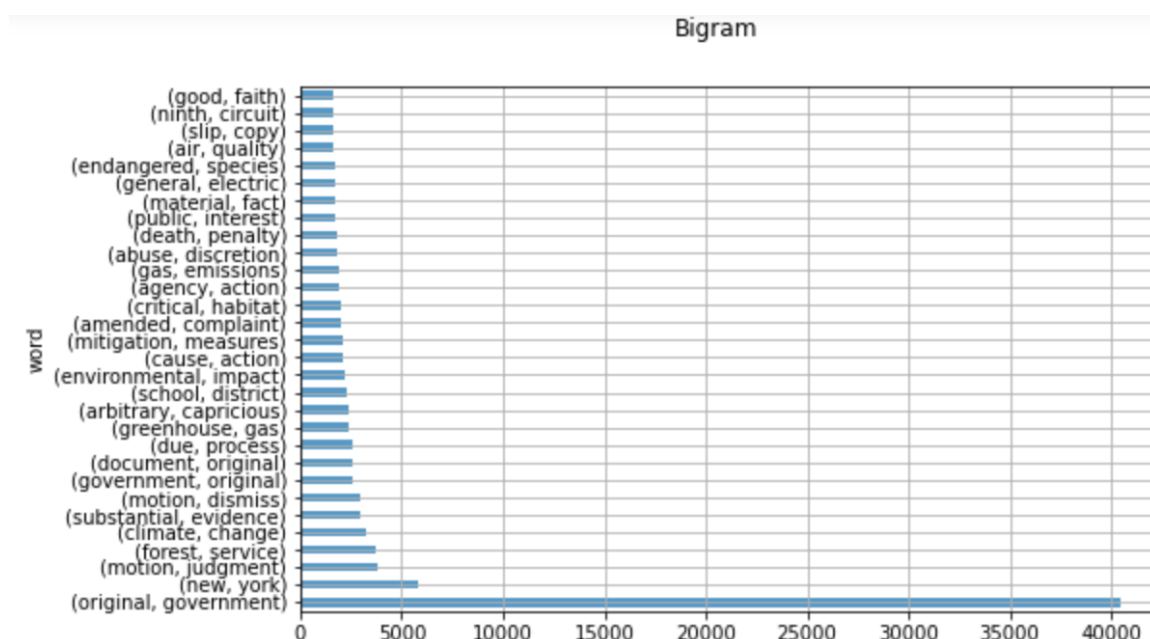

Figure 12. Top 30 Words in Bigrams.

Figure 13 displays the top 30 trigrams, representing three words frequently appearing together in the corpus. Notably, 'greenhouse gas emission' emerged as the most frequent phrase in the corpus. Additionally, 'fish wildlife services' and 'natural resources defense' occurred frequently, indicating the presence of climate change discussions as well as the active involvement of government agencies and non-governmental organizations (NGOs) in addressing climate change-related cases. 'Hostile work environment' refers to climate-change-related hazards that can put vulnerable workers (e.g., first responders, industrial

workers, and others) at risk, such as heat stress, extreme weather conditions, and exposure to chemicals and emissions. 'Employee benefits' refers to how employers can redesign the benefits package to encourage workers to minimize and offset their carbon footprints, such as clean commutes and subsidies for energy-efficient homes or locally sourced food.
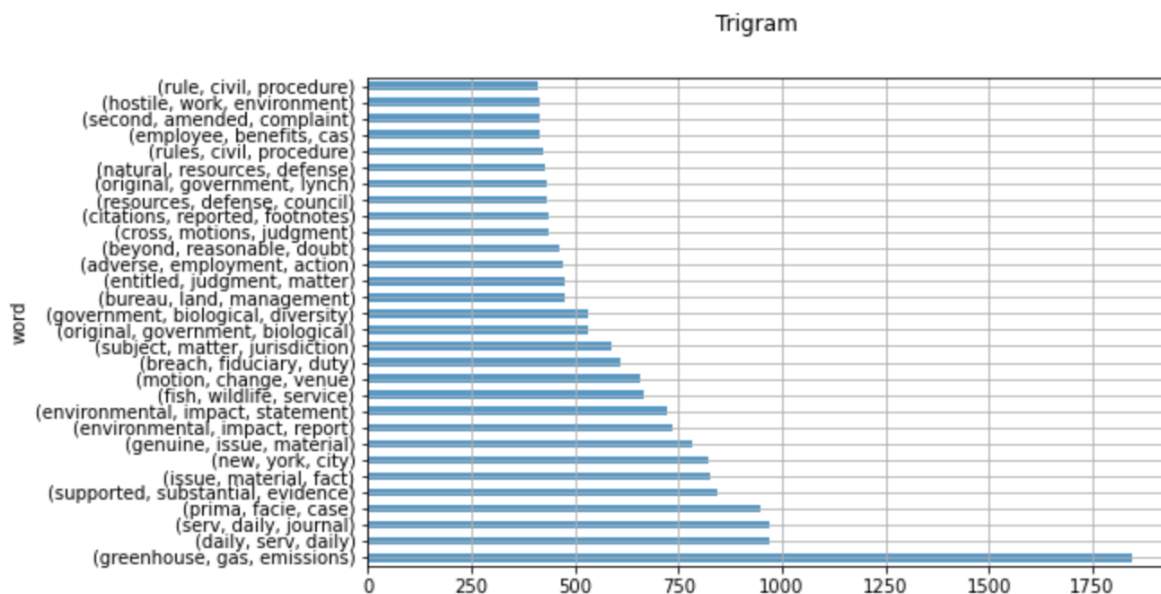

Figure 13. Top 30 words in Trigrams.

Figure 14 illustrates an example of document similarity. The TF-IDF technique was deployed to identify the top five most similar documents for a selected document. This method assigned a value to each word in the corpus and calculated the cosine similarity between the chosen document and the remaining documents in the corpus to determine their similarities. The function took a selected document as input and computed the top five most similar documents based on their similarity scores. The outcome included the title of the initially selected document, followed by the titles of the other five similar documents, along with their respective similarity scores.

```
Original document :
047 - Citizens Against Airport Pollution v City of San Jose.pdf

Most similar document:
--------------------
074 - Juliana v United States.pdf
TF-IDF Score:
0.7955347412388492
--------------------
026 - Association of Irritated Residents v Kern County Bd of Supervisors.pdf
TF-IDF Score:
0.7704299036383888
--------------------
034 - Warner v Marshall.pdf
TF-IDF Score:
0.7558638953546932
--------------------
058 - City of Holland v Federal Insurance Company.pdf
TF-IDF Score:
0.7463729630721814
--------------------
053 - North Carolina Alliance for Transp Reform Inc v US Dept of Transp.pdf
TF-IDF Score:
0.7428205906035956
```

Figure 14. Document Similarity.

Efforts were made to identify the primary statutes that exhibited the strongest associations with climate change litigation. The top statutes (Figure 15) involved wilderness, natural resources, and maritime commerce. Generally, a majority of the climate change cases in this dataset were related to wildlife preservation, managing natural resources, and marine protection.
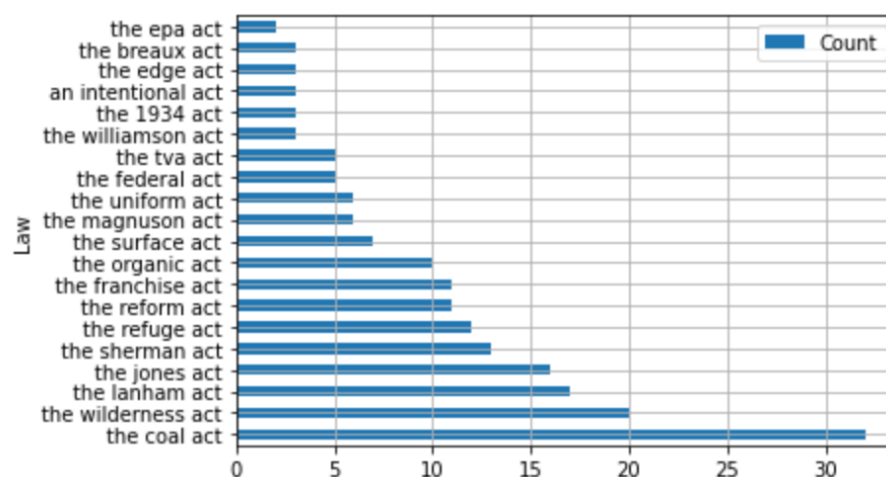

Figure 15. Top 20 Legal Statutes.

As shown in Figure 16, the top 25 litigation cases cited in the dataset are also presented. This analysis was essential for identifying the frequently cited landmark cases that contribute to legal precedents, given their importance in court proceedings. It was observed that these cases encompassed a multitude of climate protection topics. For example: 'Missouri Office of Public Counsel V. Public' discussed utility usage, 'Robertson v. Methow' was related to forest services, and 'Kanuk v. State' was a climate change litigation case.
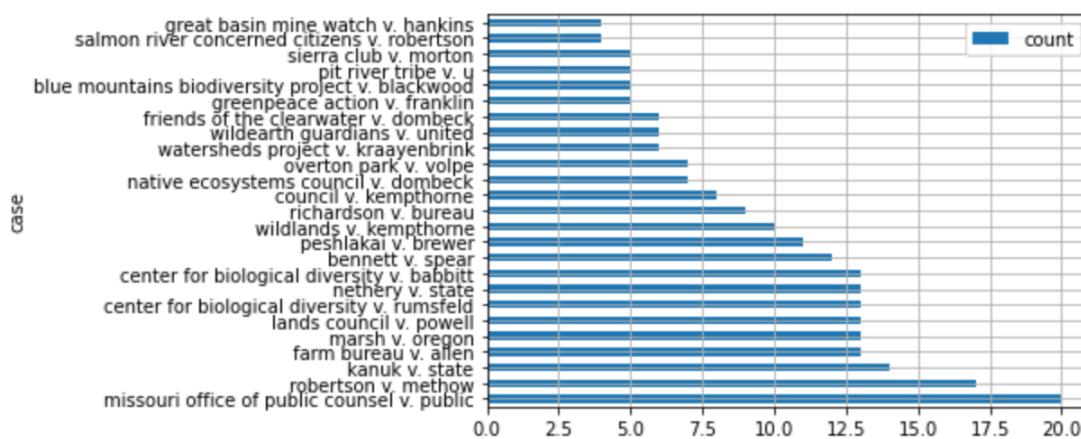

Figure 16. Top 25 litigation cases cited.

4.4. Clustering

Finally, the textual data was vectorized and clustering algorithms were applied. To input words into machine learning models, it was necessary to vectorize words based on their linguistic context so that the model could interpret the words accordingly. Numerous tasks rely on the widely recognized yet simplistic method of a bag of words (BOW) approach, such as TF-IDF. However, the results tended to be largely inconsequential, as the BOW approach lacked the representation of word order and semantic contexts. Word2vec was implemented to vectorize words. Word2vec is a technique that involves training a shallow neural network using individual words from a text and using nearby words as

labels to make predictions [110,111]. Prior to document vectorization, the unprocessed summary section obtained from web scraping underwent a transformation into plain tokens. This transformation involved the removal of punctuation, standard stop words, and lemmatization using Regex and NLTK.

Moreover, a 150-dimensional word vector in Gensim's Word2vec model was loaded to vectorize our filtered tokens. Obtaining document-level embeddings from the words present in each document was preferred, despite having obtained a dense vector for each word within our corpus. Hence, the approach involved averaging the word embeddings for each word within a document, resulting in a single embedding for each document consisting of 150 dimensions. Document features were generated for the corpus, and the documents were prepared for clustering. The objective of clustering is to investigate the possibility of discerning the focal points or emphases within the documents. The clusters were generated using the K-Means algorithm. Following the training of the K-Means model, the next step involved selecting the ten most central documents within each cluster. Subsequently, to extract prominent keywords for each cluster, the LDA model was utilized to identify the top 15 topic words in each cluster. Table 1 displays the document count in each cluster, while Figure 17 consists of word clouds for each cluster, which will then be subject to detailed analysis.

Table 1. Number of documents in each cluster.

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of Documents	84	121	86	113

(**a**)

(**b**)

Figure 17. *Cont.*

(c)

(d)

Figure 17. Cluster 1 (Forest Protection) (**a**); Cluster 2 (Emission Pollution Control) (**b**); Cluster 3 (Land Management) (**c**); and Cluster 4 (Water Habitat Protection) (**d**).

Cluster 1, as depicted, included terms like 'forest', 'fire', 'greenhouse', 'fuel', and so on. This suggests that the cases in Cluster 1 predominantly pertained to the protection of forest resources. A suitable label for this cluster could be 'Forest Protection'. Cluster 2 comprised keywords such as 'air', 'vehicle', 'emission', and 'MPCA', which stands for the Minnesota Pollution Control Agency. These cases were primarily concerned with matters related to emissions and pollution control. Thus, an appropriate label for this cluster might be 'Emission Pollution Control'. Cluster 3 pertains to terms like 'soil', 'waste', and 'bor', the latter being a shortened notation for the chemical element "boron" that is often associated with mining activities. This cluster appears to involve cases related to mineral mining. A fitting overarching term for this group could be 'Land Exploitation'. Cluster 4 consisted of terms like 'fish', 'habitat', 'water', 'coast', and others, suggesting a focus on water-related issues. Consequently, this cluster could be appropriately labeled as 'Water Habitat Protection'. In the clustering analysis, Word2vec was employed to acquire average word embeddings at the document level, effectively representing documents in a vectorized space. As the results underscore, the primary topics of discussion in the dataset included 'forest', 'land', 'water', and 'air emissions'. Additionally, our analysis revealed the active involvement of government offices and agencies in these cases, all dedicated to environmental protection efforts.

5. Discussion

In the quest to comprehend the nature and scope of climate change, various stakeholders involved in its identification and mitigation are increasingly exploring multiple perspectives. These include entities like The Intergovernmental Panel on Climate Change (IPCC) (https://www.ipcc.ch/ (accessed on 11 October 2023)), the World Economic Forum

(https://www.weforum.org/ (accessed on 11 October 2023)), national agencies and task forces, corporate-level ESG and sustainability initiatives, among others. However, there remains a shortage of detailed data regarding the nuanced aspects of climate change understanding. Recently, researchers have begun to explore novel data sources, with climate change litigation emerging as one such valuable resource. Along this line, the current research analysis of climate change litigation drew attention to matters related to the prime concepts of climate change, ESG (Environmental, Social, and Governance), and sustainability. The findings showed how climate change litigation cases represented a crucial source of information concerning the root causes of climate change. Consequently, various stakeholders, including activists, management, government bodies, non-governmental organizations (NGOs), and global agencies, can derive valuable insights into the character of climate change through this avenue.

In addition, this exploratory investigation employed a range of machine-learning techniques in text analytics to scrutinize and extract vital climate change information from a dataset of 404 cases spanning a 3-year timeframe. In this manner, the current research leveraged advancements in information processing technology to uncover insights from extensive text corpora, a task that was formerly reliant on manual study and subjective evaluation. Aligning with the research question, the methodology identified four clusters and 15 sub-topics associated with climate change, reflecting the primary areas of concern. The analysis revealed that stakeholders, such as litigants, expressed significant concerns about major topics related to CEQA, oil/gas/coal resource emissions, forest fires, salmon survival and recovery, species conservation, reclamation, greenhouse gas emissions, and pollution damage. More importantly, these topics could be categorized into four primary clusters: forest protection, emission pollution control, land management, and water habitat protection.

This research offers significant practical implications. The analysis of climate change litigation cases provides clarity and insight into the spectrum of issues that concern litigants (stakeholders). Examining climate change litigation cases also assists organizations in anticipating how these issues may develop in the public's perception. Furthermore, it enables organizations to gauge their susceptibility to public scrutiny in connection with these matters. For litigators and government officials, the empirical identification of the clusters reflecting prime areas of concern offers avenues to channel legal and policy initiatives that can mitigate climate change. This can set up a strong global precedent for utilizing legal avenues to support efforts driving action on climate change.

6. Conclusions and Future Research

In this exploratory-descriptive study, the goal was to identify key macro-level climate trends from a novel dataset comprising a corpus of legal cases. The methodology overcomes the limitations of survey-based and other types of findings that are somewhat subjective. This approach involved the application of LDA, TF-IDF, and K-Means clustering to reveal critical climate change themes, which can be categorized into four main groups: forest, emissions, land, and water. Additionally, Spacy and Regex techniques were used to extract frequently cited statutes and cases. These findings provide valuable information for key stakeholders involved in climate change mitigation efforts, moving forward.

Nonetheless, there are some matters deserving of examination. For instance, to what degree can text documents like climate change litigation cases be employed to predict climate change trends and managerial responses? Conducting trend analysis allows for an assessment of the novelty of cases and the influence of rulings as they evolve over time. The influence of climate change litigation on public perception, as well as the reciprocal relationship, can be explored through a social media lens, particularly in terms of public sentiment analysis. Achieving reproducibility as well as validation poses significant challenges in implementing machine learning and text analytics. Moreover, the intuitive categorization of clusters and the interpretation of machine learning outcomes can introduce subjectivity. Nevertheless, there is confidence in the results of the analysis, including the

identification of key topics and cluster labels, as they generally align with the descriptions found in the cases, a validation supported by prior literature. As such, this research offers a robust methodology for investigating and comprehending climate change. Further limitations pertain to the reliability of the documents and the soundness with which the data can be prepared for analysis. The applicability of the identified topics (clusters) may be somewhat uncertain due to the examination of a limited sample of cases. These findings may not fully represent comprehensive national-level endeavors and strategies concerning climate change on a macro scale. Additionally, machine learning models have limitations in their ability to extract comprehensive insights. Furthermore, it is worth noting that this study exclusively relied on case documents. Subsequent research endeavors could enhance the understanding gained from litigation cases by incorporating additional empirical data sources.

Notwithstanding these constraints, our study makes significant contributions to policy and research. First, the findings offer a valuable resource for practitioners and researchers to prioritize climate change initiatives. They can also explore various clusters from alternative perspectives, such as those of NGOs or other activists. Second, this study serves as a demonstration of the effectiveness of machine learning and text analytics in comprehending climate change litigation cases, thereby enabling informed decision-making through descriptive, predictive, and prescriptive analytics. Third, given the scarcity of data related to climate change, particularly within the corporate sector, extracting insights from litigation cases represents an innovative contribution to the research. Fourth, the macro-level analysis provides valuable insights into the overarching critical issues pertaining to climate change.

Subsequent research endeavors can further investigate and implement advanced techniques, such as deep learning, to delve deeper into case analysis for enhanced content analysis. For instance, there is room for exploring prescriptive analytics, which not only predicts outcomes but can also explore potential impacts and strategies. Future research opportunities encompass conducting comparisons across industries and states, exploring global disparities, and investigating the cost-benefit dynamics of litigation versus settlement, which influence corporate responses to climate change and their correlation with company performance. Furthermore, the application of discovery analytics to the resolutions could provide insights into innovation and the generation of new product ideas.

For future studies looking to conduct topic analysis of legal cases, the maturation of deep learning will facilitate the extraction of insights from textual data. In addition, predictive studies can be done with the objective of supporting litigation as well as minimizing the costs of litigation. Artificial intelligence can contribute to this difficult task of predicting the rulings of judges or the likely outcomes of cases. If litigants had access to predictions about the likely case outcomes, they may choose to settle, rather than engage in prolonged and uncertain litigation. In the broader context, the fusion of law and data science holds the promise to provide deeper insights into the dynamics of climate change.

References

1. McCormick, S.; Simmens, S.J.; Glicksman, R.L.; Paddock, L.; Kim, D.; Whited, B.; Davies, W. Science in litigation, the third branch of US climate policy. *Science* **2017**, *357*, 979–980. [CrossRef]
2. McCormick, S.; Simmens, S.J.; Glicksman, R.; Paddock, L.; Kim, D.; Whited, B. The role of health in climate litigation. *Am. J. Public Health* **2018**, *108*, S104–S108. [CrossRef] [PubMed]
3. McCormick, S.; Glicksman, R.L.; Simmens, S.J.; Paddock, L.; Kim, D.; Whited, B. Strategies in and outcomes of climate change litigation in the United States. *Nat. Clim. Chang.* **2018**, *8*, 829–833. [CrossRef]
4. UNEP. Global Climate Litigation Report: 2020 Status Review. Nairobi. 2020. Available online: https://wedocs.unep.org/handle/20.500.11822/34818 (accessed on 8 October 2023).
5. UNEP. Global Climate Litigation Report: 2023 Status Review. 2023. Available online: https://www.unep.org/resources/report/global-climate-litigation-report-2023-status-review (accessed on 8 October 2023).
6. Burger, M.; Tigre, M.A. *Global Climate Litigation Report: 2023 Status Review*; UNEP—UN Environment Programme: Nairobi, Kenya, 2023.
7. IPCC. Press_Release_WGI_AR6_Website-Final (ipcc.ch). 2021. Available online: https://www.ipcc.ch/site/assets/uploads/2021/08/IPCC_WGI-AR6-Press-Release_en.pdf (accessed on 8 November 2023).
8. IPCC. *Climate Change Widespread, Rapid, and Intensifying—IPCC*; IPCC: Geneva, Switzerland, 2021.
9. The World Bank. *Climate Change Overview: Development News, Research, Data*; World Bank: Washington, DC, USA, 2023.
10. UN. *What Is Climate Change?* United Nations: New York, NY, USA, 2023.
11. UN. Fastfacts-What-Is-Climate-Change.pdf (un.org). 2023. Available online: https://www.un.org/en/climatechange/what-is-climate-change (accessed on 9 November 2023).
12. EPA. Climate Change Science Facts (epa.gov). 2023. Available online: https://www.epa.gov/climatechange-science (accessed on 9 November 2023).
13. Blattner, C.E.; Vicedo-Cabrera, A.M.; Frölicher, T.L.; Ingold, K.; Raible, C.C.; Wyttenbach, J. How science bolstered a key European climate-change case. *Nature* **2023**, *621*, 255–257. [CrossRef] [PubMed]
14. Liu, Z.; Deng, Z.; Davis, S.J.; Giron, C.; Ciais, P. Monitoring global carbon emissions in 2021. *Nat. Rev. Earth Environ.* **2022**, *3*, 217–219. [CrossRef] [PubMed]
15. Lee, H.; Calvin, K.; Dasgupta, D.; Krinner, G.; Mukherji, A.; Thorne, P.; Ruane, A.C. *Climate Change 2023 Synthesis Report: Summary for Policymakers*; IPCC: Geneva, Switzerland, 2023.
16. Carattini, S.; Hertwich, E.; Melkadze, G.; Shrader, J.G. Mandatory disclosure is key to address climate risks. *Science* **2022**, *378*, 352–354. [CrossRef]
17. Dawkins, C.; Fraas, J.W. Coming clean: The impact of environmental performance and visibility on corporate climate change disclosure. *J. Bus. Ethics* **2011**, *100*, 303–322. [CrossRef]
18. Giannarakis, G.; Zafeiriou, E.; Arabatzis, G.; Partalidou, X. Determinants of corporate climate change disclosure for European firms. *Corp. Soc. Responsib. Environ. Manag.* **2018**, *25*, 281–294. [CrossRef]
19. Ihlen, Ø. Business and climate change: The climate response of the world's 30 largest corporations. *Environ. Commun.* **2009**, *3*, 244–262. [CrossRef]
20. Stanny, E.; Ely, K. Corporate environmental disclosures about the effects of climate change. *Corp. Soc. Responsib. Environ. Manag.* **2008**, *15*, 338–348. [CrossRef]
21. Wright, C.; Nyberg, D. An inconvenient truth: How organizations translate climate change into business as usual. *Acad. Manag. J.* **2017**, *60*, 1633–1661. [CrossRef]
22. Aversa, D. Scenario analysis and climate change: A literature review via text analytics. *Br. Food J.* 2023, *ahead of print*.
23. Alencar, A.B.; de Oliveira, M.C.F.; Paulovich, F.V. Seeing beyond reading: A survey on visual text analytics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 476–492. [CrossRef]
24. Casey, J. Text analytics techniques in the digital world: A sentiment analysis case study of the coverage of climate change on US news networks. *Ir. Commun. Rev.* **2018**, *16*, 7.
25. Dahlmann, F.; Branicki, L.; Brammer, S. Managing carbon aspirations: The influence of corporate climate change targets on environmental performance. *J. Bus. Ethics* **2019**, *158*, 1–24. [CrossRef]
26. Gao, L.; Calderon, T.G. Climate Change Risk Disclosures and Audit Fees: A Text Analytics Assessment. *J. Emerg. Technol. Account.* **2023**, *20*, 71–93. [CrossRef]
27. Huntingford, C.; Jeffers, E.S.; Bonsall, M.B.; Christensen, H.M.; Lees, T.; Yang, H. Machine learning and artificial intelligence to aid climate change research and preparedness. *Environ. Res. Lett.* **2019**, *14*, 124007. [CrossRef]
28. Rolnick, D.; Donti, P.L.; Kaack, L.H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. Tackling climate change with machine learning. *ACM Comput. Surv.* **2022**, *55*, 1–96. [CrossRef]
29. Ahmad, F.M. *Beyond the Horizon: Corporate Reporting on Climate Change*; Center for Climate and Energy Solutions: Arlington, VA, USA, 2017. Available online: www.c2es.org/site/assets/uploads/2017/09/beyond-horizon-corporate-reporting-climate-change.pdf (accessed on 17 August 2018).

30.　Ajax, C.M.; Strauss, D. Corporate Sustainability Disclosures in American Case Law: Purposeful or Mere "Puffery"? *Ecol. Law Q.* **2018**, *45*, 703–734. [CrossRef]

31.　Aragón-Correa, J.A.; Marcus, A.; Hurtado-Torres, N. The natural environmental strategies of international firms: Old controversies and new evidence on performance and disclosure. *Acad. Manag. Perspect.* **2016**, *30*, 24–39. [CrossRef]

32.　Eleftheriadis, I.M.; Anagnostopoulou, E.G. Relationship between corporate climate change disclosures and firm factors. *Bus. Strategy Environ.* **2015**, *24*, 780–789. [CrossRef]

33.　Le Ravalec, M.; Rambaud, A.; Blum, V. Taking climate change seriously: Time to credibly communicate on corporate climate performance. *Ecol. Econ.* **2022**, *200*, 107542. [CrossRef]

34.　Belal, A.R.; Kabir, M.R.; Cooper, S.; Dey, P.; Khan, N.A.; Rahman, T.; Ali, M. Corporate environmental and climate change disclosures: Empirical evidence from Bangladesh. In *Research in Accounting in Emerging Economies*; Emerald Group Publishing Limited: Bingley, UK, 2010; Volume 10, pp. 145–167.

35.　Haque, S.; Deegan, C. Corporate climate change-related governance practices and related disclosures: Evidence from Australia. *Aust. Account. Rev.* **2010**, *20*, 317–333. [CrossRef]

36.　Solomon, J.F.; Solomon, A.; Norton, S.D.; Joseph, N.L. Private climate change reporting: An emerging discourse of risk and opportunity? *Account. Audit. Account. J.* **2011**, *24*, 1119–1148. [CrossRef]

37.　Nurunnabi, M. Who cares about climate change reporting in developing countries? The market response to, and corporate accountability for, climate change in Bangladesh. *Environ. Dev. Sustain.* **2016**, *18*, 157–186. [CrossRef]

38.　Rouas, V. *Achieving Access to Justice in a Business and Human Rights Context an Assessment of Litigation and Regulatory Responses in European Civil-Law Countries*; University of London Press: London, UK, 2022.

39.　Eccles, R.G.; Krzus, M.P. Why companies should report financial risks from climate change. *MIT Sloan Manag. Rev.* **2018**, *59*, 1–6.

40.　Masuma, M.H.; Hassanb, N.; Jahana, T. Corporate climate change reporting: Evidence from Bangladesh. *Account. Manag. Inf. Syst.* **2019**, *18*, 399–416. [CrossRef]

41.　Gulluscio, C.; Puntillo, P.; Luciani, V.; Huisingh, D. Climate change accounting and reporting: A systematic literature review. *Sustainability* **2020**, *12*, 5455. [CrossRef]

42.　Hösli, A.; Weber, R.H. Climate change reporting and due diligence: Frontiers of corporate climate responsibility. *Eur. Co. Financ. Law Rev.* **2022**, *18*, 948–979. [CrossRef]

43.　Daradkeh, H.; Shams, S.; Bose, S.; Gunasekarage, A. Does managerial ability matter for corporate climate change disclosures? *Corp. Gov. Int. Rev.* **2023**, *31*, 83–104. [CrossRef]

44.　Park, J.D.; Nishitani, K.; Kokubu, K.; Freedman, M.; Weng, Y. Revisiting sustainability disclosure theories: Evidence from corporate climate change disclosure in the United States and Japan. *J. Clean. Prod.* **2023**, *382*, 135203. [CrossRef]

45.　Cadez, S.; Czerny, A.; Letmathe, P. Stakeholder pressures and corporate climate change mitigation strategies. *Bus. Strategy Environ.* **2019**, *28*, 1–14. [CrossRef]

46.　Hamman, E. Save the reef! Civic crowdfunding and public interest environmental litigation. *QUT Law Rev.* **2015**, *15*, 159–173. [CrossRef]

47.　Hsu, S.L. A realistic evaluation of climate change litigation through the lens of a hypothetical lawsuit. *Univ. Colo. Law Rev.* **2008**, *79*, 701.

48.　Markell, D.; Ruhl, J.B. An empirical assessment of climate change in the courts: A new jurisprudence or business as usual. *Fla. Law Rev.* **2012**, *64*, 15. [CrossRef]

49.　Peel, J.; Lin, J. Transnational climate litigation: The contribution of the Global South. *Am. J. Int. Law* **2019**, *113*, 679–726. [CrossRef]

50.　Peel, J.; Markey-Towler, R. Recipe for success?: Lessons for strategic climate litigation from the Sharma, Neubauer, and Shell cases. *Ger. Law J.* **2022**, *22*, 1484–1498. [CrossRef]

51.　Vanhala, L.; Hilson, C. Climate change litigation: Symposium introduction. *Law Policy* **2013**, *35*, 141–149. [CrossRef]

52.　Butti, L. The Tortuous Road to Liability: A Critical Survey on Climate Change Litigation in Europe and North America. *Sustain. Dev. Law Policy* **2010**, *11*, 32.

53.　Hunter, D.; Salzman, J. Negligence in the air: The duty of care in climate change litigation. *Univ. Pa. Law Rev.* **2006**, *155*, 1741.

54.　Peel, J. Issues in climate change litigation. *Carbon Clim. Law Rev.* **2011**, *5*, 15–24. [CrossRef]

55.　Peel, J.; Osofsky, H.M. A rights turn in climate change litigation? *Transnatl. Environ. Law* **2018**, *7*, 37–67. [CrossRef]

56.　Savaresi, A.; Setzer, J. Rights-based litigation in the climate emergency: Mapping the landscape and new knowledge frontiers. *J. Hum. Rights Environ.* **2022**, *13*, 7–34. [CrossRef]

57.　Tigre, M.A. A Look Back at Significant Decisions in Climate Litigation in 22 December 2022. Available online: https://blogs.law.columbia.edu/climatechange/2022/12/22/a-lookback-at-significant-decisions-in-climate-litigation-in-2022/ (accessed on 10 April 2023).

58.　Houniuhi, C. Why I'm leading Pacific Islands students in the fight on climate change. *Nature* **2023**, *618*, 9. [CrossRef] [PubMed]

59.　Gelles, D.; Baker, M. Judge Rules in Favor of Montana Youths in a Landmark Climate Case. *The New York Times*. 14 August 2023. Available online: https://www.nytimes.com/2023/08/14/us/montana-youth-climate-ruling.html?smid=url-share (accessed on 11 October 2023).

60. Posner, E.A. Climate change and international human rights litigation: A critical appraisal. *Univ. Pa. Law Rev.* **2006**, *155*, 1925. [CrossRef]
61. Thorpe, A. Tort-based climate change litigation and the political question doctrine. *J. Land Use Environ. Law* **2008**, *24*, 79.
62. Toussaint, P. Loss and damage and climate litigation: The case for greater interlinkage. *Rev. Eur. Comp. Int. Environ. Law* **2021**, *30*, 16–33. [CrossRef]
63. Peel, J.; Osofsky, H.M. Climate change litigation. *Annu. Rev. Law Soc. Sci.* **2020**, *16*, 21–38. [CrossRef]
64. Sabin Center for Climate Change Law. Climate Change Litigation | Sabin Center for Climate Change Law. 2023. Available online: https://climate.law.columbia.edu/ (accessed on 10 October 2023).
65. Pielke, R.A., Jr. Misdefining "climate change": Consequences for science and action. *Environ. Sci. Policy* **2005**, *8*, 548–561. [CrossRef]
66. Chapter 13: National and Sub-National Policies and Institutions. In *Climate Change 2022: Mitigation of Climate Change. Working Group III Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Intergovernmental Panel on Climate Change: Geneva, Switzerland, 2022. Available online: https://www.ipcc.ch/report/ar6/wg3/downloads/report/IPCC_AR6_WGIII_Chapter_10.pdf (accessed on 10 October 2023).
67. Kemp, L.; Xu, C.; Depledge, J.; Ebi, K.L.; Gibbins, G.; Kohler, T.A.; Rockström, J.; Scheffer, M.; Schellnhuber, H.J.; Steffen, W.; et al. Climate Endgame: Exploring catastrophic climate change scenarios. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2108146119. [CrossRef]
68. Demertzidis, N.; Tsalis, T.A.; Loupa, G.; Nikolaou, I.E. A benchmarking framework to evaluate business climate change risks: A practical tool suitable for investors decision-making process. *Clim. Risk Manag.* **2015**, *10*, 95–105. [CrossRef]
69. Gasbarro, F.; Pinkse, J. Corporate adaptation behaviour to deal with climate change: The influence of firm-specific interpretations of physical climate impacts. *Corp. Soc. Responsib. Environ. Manag.* **2016**, *23*, 179–192. [CrossRef]
70. Gouldson, A.; Sullivan, R. Long-term corporate climate change targets: What could they deliver? *Environ. Sci. Policy* **2013**, *27*, 1–10. [CrossRef]
71. Nikolaou, I.; Evangelinos, K.; Leal Filho, W. A system dynamic approach for exploring the effects of climate change risks on firms' economic performance. *J. Clean. Prod.* **2015**, *103*, 499–506. [CrossRef]
72. Pinkse, J.; Kolk, A. Challenges and trade-offs in corporate innovation for climate change. *Bus. Strategy Environ.* **2010**, *19*, 261–272. [CrossRef]
73. British Institute of International 1 and Comparative Law. Global Perspectives on Corporate Climate Legal Tactics. 2023. Available online: https://www.biicl.org/projects/global-perspectives-on-corporate-climate-legal-tactics?cookiesset=1&ts=1699734136 (accessed on 11 October 2023).
74. Wewerinke-Singh, M.; McCoach, A. The State of the Netherlands v Urgenda Foundation: Distilling best practice and lessons learnt for future rights-based climate litigation. *Rev. Eur. Comp. Int. Environ. Law* **2021**, *30*, 275–283. [CrossRef]
75. Setzer, J.; Higham, C. *Global Trends in Climate Change Litigation: 2022 Snapshot*; Grantham Research Institute on Climate Change and the Environment: London, UK, 2022.
76. Setzer, J.; Vanhala, L.C. Climate change litigation: A review of research on courts and litigants in climate governance. *Wiley Interdiscip. Rev. Clim. Chang.* **2019**, *10*, e580. [CrossRef]
77. 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. [CrossRef]
78. Landmann, J.; Zuell, C. Identifying events using computer-assisted text analysis. *Soc. Sci. Comput. Rev.* **2008**, *26*, 483–497. [CrossRef]
79. Shelley, M.; Krippendorff, K. Content Analysis: An Introduction to its Methodology. *J. Am. Stat. Assoc.* **1984**, *79*, 240. [CrossRef]
80. 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]
81. 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]
82. Raghupathi, V.; Zhou, Y.; Raghupathi, W. Exploring big data analytic approaches to cancer blog text analysis. *Int. J. Healthc. Inf. Syst. Inform.* **2019**, *14*, 1–20. [CrossRef]
83. Raghupathi, V.; Ren, J.; Raghupathi, W. Identifying corporate sustainability issues by analyzing shareholder resolutions: A machine-learning text analytics approach. *Sustainability* **2020**, *12*, 4753. [CrossRef]
84. Raghupathi, W.; Wu, S.J.; Raghupathi, V. Understanding Corporate Sustainability Disclosures from the Securities Exchange Commission Filings. *Sustainability* **2023**, *15*, 4134. [CrossRef]
85. Dahal, B.; Kumar, S.A.; Li, Z. Topic modeling and sentiment analysis of global climate change tweets. *Soc. Netw. Anal. Min.* **2019**, *9*, 1–20. [CrossRef]
86. Chandrapaul; Soni, R.; Sharma, S.; Fagna, H.; Mittal, S. News analysis using word cloud. In *Advances in Signal Processing and Communication: Select Proceedings of ICSC 2018*; Springer: Singapore, 2019; pp. 55–64.
87. Haider, M.M.; Hossin, M.A.; Mahi, H.R.; Arif, H. Automatic text summarization using gensim word2vec and k-means clustering algorithm. In Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 5–7 June 2020; IEEE: New York, NY, USA, 2020; pp. 283–286.

88. McCallum, A.K. Mallet: A Machine Learning for Language Toolkit. 2002. Available online: http://mallet.cs.umass.edu (accessed on 17 September 2023).

89. Řehůřek, R.; Sojka, P. Gensim—Statistical Semantics in Python. 2011. Available online: https://pypi.org/project/gensim/ (accessed on 17 September 2023).

90. Sarkar, R.; McCrae, J.P.; Buitelaar, P. A supervised approach to taxonomy extraction using word embeddings. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.

91. Aninditya, A.; Hasibuan, M.A.; Sutoyo, E. Text mining approach using TF-IDF and naive Bayes for classification of exam questions based on cognitive level of bloom's taxonomy. In Proceedings of the 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), Bali, Indonesia, 5–7 November 2019; IEEE: New York, NY, USA, 2019; pp. 112–117.

92. Carnot, M.L.; Bernardino, J.; Laranjeiro, N.; Gonçalo Oliveira, H. Applying text analytics for studying research trends in dependability. *Entropy* **2020**, *22*, 1303. [CrossRef]

93. Han, H.J.; Mankad, S.; Gavirneni, N.; Verma, R. What Guests Really Think of Your Hotel: Text Analytics of Online Customer Reviews. 2016. Available online: https://ecommons.cornell.edu/items/658a3400-e42f-4be9-b5ac-c25e1bc36efd (accessed on 9 November 2023).

94. Dietterich, T.G. Machine learning in ecosystem informatics and sustainability. In Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence, Pasadena, CA, USA, 14–17 July 2009.

95. Shahi, A.M.; Issac, B.; Modapothala, J.R. Analysis of supervised text classification algorithms on corporate sustainability reports. In Proceedings of the 2011 International Conference on Computer Science and Network Technology, Harbin, China, 24–26 December 2011.

96. Székely, N.; Vom Brocke, J. What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PLoS ONE* **2017**, *12*, e0174807. [CrossRef]

97. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]

98. Zhou, Y.; Wang, X.; Yuen, K.F. Sustainability disclosure for container shipping: A text-mining approach. *Transp. Policy* **2021**, *110*, 465–477. [CrossRef]

99. Lavin, M. Analyzing Documents with TF-IDF. 2019. Available online: https://digitalcommons.denison.edu/cgi/viewcontent.cgi?article=2064&context=facultypubs (accessed on 9 November 2023).

100. 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]

101. Liu, C.Z.; Sheng, Y.X.; Wei, Z.Q.; Yang, Y.Q. Research of text classification based on improved TF-IDF algorithm. In Proceedings of the 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), Lanzhou, China, 24–27 August 2018; IEEE: New York, NY, USA, 2018; pp. 218–222.

102. Pan, S.; Li, Z.; Dai, J. An improved TextRank keywords extraction algorithm. In Proceedings of the ACM Turing Celebration Conference, Chengdu, China, 17–19 May 2019; pp. 1–7.

103. Wang, Y.; Xu, W. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decis. Support Syst.* **2018**, *105*, 87–95. [CrossRef]

104. Wei, X.; Croft, W.B. LDA-based document models for ad-hoc retrieval. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 6–11 August 2006; pp. 178–185.

105. Zhang, W.; Yoshida, T.; Tang, X. A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Syst. Appl.* **2011**, *38*, 2758–2765. [CrossRef]

106. Xiong, C.; Li, X.; Li, Y.; Liu, G. Multi-documents summarization based on TextRank and its application in online argumentation platform. *Int. J. Data Warehous. Min.* **2018**, *14*, 69–89. [CrossRef]

107. Dreyfus, D.A.; Ingram, H.M. The National Environmental Policy Act: A view of intent and practice. *Nat. Resour. J.* **1976**, *16*, 243.

108. Varner, S.S. The California Environmental Quality Act (CEQA) after two decades: Relevant problems and ideas for necessary reform. *Pepperdine Law Rev.* **1991**, *19*, 1447.

109. Graham, S.; Weingart, S.; Milligan, I. Getting Started with Topic Modeling and MALLET; The Editorial Board of the Programming Historian, 2012. Available online: https://www.uwspace.uwaterloo.ca/handle/10012/11751?show=full (accessed on 9 November 2023).

110. 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]

111. 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. In *Mining Text Data*; Springer: Boston, MA, USA, 2012; pp. 129–161.

112. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.

113. Krestel, R.; Fankhauser, P.; Nejdl, W. Latent dirichlet allocation for tag recommendation. In Proceedings of the Third ACM Conference on Recommender Systems, New York, NY, USA, 23–25 October 2009; pp. 61–68.

114. Debortoli, S.; Müller, O.; Junglas, I.; Vom Brocke, J. Text mining for information systems researchers: An annotated topic modeling tutorial. *Commun. Assoc. Inf. Syst.* **2016**, *39*, 7. [CrossRef]

115. 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, 19–21 October 2017; IEEE: New York, NY, USA, 2017; pp. 165–174.

116. 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; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2012; pp. 958–962.