Original Research Article

Multiclass Text Classification of Climate Change Reports: Insights into Natural Disasters, Impacts, Locations, and Time Periods

ABSTRACT

In recent years the quantity of unstructured information sources, such as documents that include complex information about geospatial concepts, places, as well as events and phenomena occurring in space and time has increased exponentially. Machine learning techniques may be used to extract semantic information from these unstructured sources of geospatial information that remain unexploited to enable semantic organization and classification, as well as a deeper comprehension of the inherent knowledge.

The paper describes the application of a text classification technique on an extensive report related to climate change to predict the category of a document based on four geospatial aspects. More specifically, a multinomial logistic regression learning approach is applied to train multiclass models and classify climate change reports based on classes related to natural disasters, their impacts, as well as their spatial and temporal dimensions.

Keywords: text classification, multinomial logistic regression, unstructured geospatial data, climate change, natural disasters, spatiotemporal information.

1. INTRODUCTION

In recent years, a growing body of research on geospatial information has shifted the focus from structured information sources, such as spatial databases to unstructured information sources, such as social media posts, travel blogs, or scientific reports. These sources include an abundance of information about geographic concepts, places, events and phenomena occurring in space and time; however, they lack organization and structure.

In order to overcome the significant challenge in extracting and organizing the semantic content that is immanent in these sources, semantic information extraction and text classification processes are used to identify and organize the inherent spatial knowledge. These processes are critical in various domains, such as citation analysis, crisis management, geospatial analysis, and natural disaster detection, especially when dealing with large amounts of unstructured text. They implement various techniques to identify relevant domain concepts, topics, and sentiments within textual data, classify documents based on their content, and uncover immanent patterns.

The present paper focuses on the implementation of a text classification technique specifically tailored to climate change reports that aims to predict the category of a document based on a wide range of classes related to natural disasters, their impacts, as well as their spatial and temporal dimensions. These extensive and elaborate reports offer important new insights into the many facets of climate change, such as risks, impacts, and vulnerability. Risks from climate change, such as extreme weather events, loss of ecosystems, reduced food security, damaged livelihoods, health and security of people, migration and displacement, and increased inequality "differ through space and time and cascade across and within regions and systems" [1]. These facets of climate change may be extracted and analyzed by utilizing machine learning techniques, enabling a greater understanding of the correlations and patterns that emerge. More specifically, a multinomial logistic regression learning approach is employed using the Amazon Comprehend tool (https://aws.amazon.com/comprehend/) provided by AWS cloud services to train and evaluate machine learning models for multiclass text classification.

The paper is structured as follows. Section 2 presents related work. Section 3 describes the methodology and more specifically data collection, preprocessing, and multiclass text classification using multinomial logistic regression. Section 4 covers training and validation steps, performance evaluation of text classification models, and analysis of predictions for different classes. Finally, Section 5 presents the conclusions of the present research and suggests areas for future work.

2. RELATED WORK

Text classification is a machine learning technique that classifies text documents into a set of predefined classes. It is considered as a supervised learning approach that uses a training set of documents labelled with a fixed number of classes to create a classification model that predicts the class label of a new and unknown document based on the training model [2]. Training algorithms may be classified into supervised, semi-supervised, ensemble, active, transfer, or multi-view learning methods [3].

Various machine learning approaches have been employed to solve several text classification challenges. Kamath et al. [4] compared convolutional neural networks and traditional machine learning methods such as Logistic Regression, Support Vector Machine, Multinomial Naive Bayes, and Random Forest Classifier. Mirończuk and Protasiewicz[3] surveyed several learning approaches to deconstruct the essential phases of text classification, including feature selection and dimensionality reduction techniques that aim to improve classification accuracy and computational efficiency.

Altınel and Ganiz[5]explored semantic text classification approaches that focus on the meaning of words and hidden semantic connections between words and accordingly between documents, and organized them into five fundamental categories: domain knowledge-based approaches, corpus-based approaches, deep learning-based approaches, character sequence enhanced approaches and linguistic enriched approaches. Kunnath et al. [6] also analyzed semantic classification techniques, including manual, rule-based, machine learning, and deep learning approaches applied to citation information analysis.

An important point that emerged from these analyses of text classification approaches is that the domain and context in which text classification is applied, as well as the specific requirements and available data are important factors for the selection of the appropriate text classification technique [3], [5], [6].

In the geospatial domain several research topics, such as semantic information extraction, geographic information retrieval, geospatial Semantic Web, place semantics, etc. focus on identifying spatiotemporal information from various semi-structured and unstructured sources, such as texts, social media, and images [7].

Karimi et al. [8] conducted a comparison of machine learning techniques including Logistic Regression (LogReg), Support Vector Machine, Stochastic Gradient Descent (SGD), k-Nearest Neighbors, Random Forest, Gradient Boosting, Decision Tree, Multinomial Naïve Bayes (MNB) and Multi-Layer Perception, for the extraction of place functionality from textual data. When using lemmatized words and action verbs, MNB was found to be the quickest algorithm among text classification algorithms using the Bag-of-Words (BoW) feature selection method. Logistic Regression and Stochastic Gradient Descent were the next fastest algorithms using all three feature selection methods (BoW, Word2Vec, and Doc2Vec). Yang et al. [9] reviewed visually enabled deep learning techniques for text and image classification. They discussed various methods for applying labels to train data and highlighted the importance of using visual information to guide active learning. Häberle et al. [10] analyzed geo-tagged Twitter data in Los Angeles urban region using pre-trained word vectors obtained from word2vec, fastText and GloVe and classification with Naïve Bayes, support vector machines, and a convolutional neural network in order to classify building types into commercial and residential.

Information extraction and classification have also been employed for natural disaster detection and management, especially from social media. Burel et al. [11] implemented a dual convolutional neural network to identify relevant events in emergency situations from social media data. Wang and Ye [12] focused on the analysis of social media based on the dimensions of space, time, content, and network and reviewed various techniques that can be used for mining these dimensions. Said et al. [13] surveyed natural disaster detection methods from social media and satellite imagery. They focused on text-based data from Twitter for disaster detection, employing NLP methods such as Named Entity Recognition (NER) along with machine learning algorithms like Support Vector Machines. Additionally, they investigated the detection of disasters from images through the implementation of various forms of neural networks. Ye et al. [14] developed an information model of typhoon events and proposed an information aggregation method for extracting information related to typhoon events and detecting the evolution process of such events from microblog text.

The present research focuses on the use of multiclass text classification to analyze climate change reports across four different dimensions: natural disasters, impacts, locations, and time periods. More specifically, multinomial logistic regression is applied using four custom classifiers trained to identify and categorize unstructured data according to these four dimensions that are highly relevant to climate change. The research aims to enable organizing and extracting insights from large volumes of text based on specific geospatial information tailored to the general topic of the text corpus.

3. METHODOLOGY

Sections 3 and 4 describe the methodology used to analyze climate change reports and identify patterns and relationships among locations, time periods, natural disasters, and impacts. The workflow includes the following steps: data collection, preprocessing, classifier training and testing, subset analysis, and visualization of results (Figure 1).

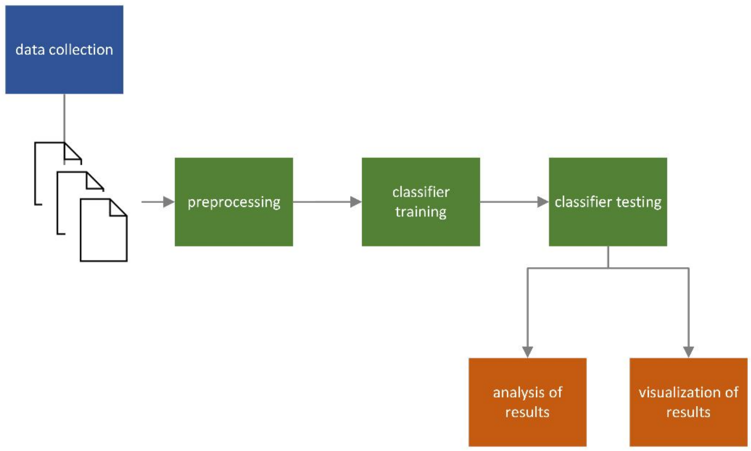


Fig.1. Workflow of the proposed approach.

3.1 DataCollection and Preprocessing

The first step consisted in gathering unstructured data in text format from various sources, such as Wikipedia, encyclopedias, historical sites, and news articles with a significant focus on those provided by the Centre for Research on Epidemiology of Disasters [15]. The objective was to acquire a comprehensive representation of natural disasters and their impacts, as well as different countries and time periods from multiple sources to ensure that the corresponding labeled dataset provided a solid foundation for training the custom classifiers in order to facilitate the subsequent analysis of the Intergovernmental Panel on Climate Change (IPCC) report [16].

A broad dataset was collected that included examples of locations, time periods, natural disasters, and climate change impacts. To ensure the effectiveness of the custom classifiers, each classifier was trained with 50 texts, each ranging in size from approximately 300 to 350 words. Table 1 lists the labels used to represent the classifiers.

Twelve countries threatened by (and vulnerable to) climate change were selected: USA, China, India, Indonesia, Colombia, Brazil, France, Peru, Philippines, Vietnam, Mexico, and Malaysia. Regarding time periods, the focus was on capturing the 20th - 21st century span, covering decades from the 1920s to the 2010s in order to analyze climate change trends and impacts during the most recent era.

A range of natural disasters that cause great damage and loss were selected: floods, droughts, heatwaves, wildfires, storms, cyclones, hurricanes, tsunamis, earthquakes, avalanches, and typhoons. Similarly, a wide range of impacts were selected including loss of life and health impacts, loss of livelihoods, environmental damage, social and economic disruption, and property damage.

In order to optimize the text data for the subsequent text classification and enhance the accuracy of the models, a preprocessing stage was implemented that includes the following techniques:

- the removal of common stop words, such as "the" "and" and "in", punctuation marks and special characters, which can introduce noise into the classification process,
- case normalization which consists in converting all characters to lowercase, to ensure that words are recognized as the same item by the models independent of the use of capital letters, and
- lemmatization which consists in identifying the base or dictionary form of a word (lemma).

Tables should be explanatory enough to be understandable without any text reference. Double spacing should be maintained throughout the table, including table headings and footnotes. Table headings should be placed above the table. Footnotes should be placed below the table with superscript lowercase letters. Sample table format is given below.

Table 1. Labels of each Classifier

Locations	**Time Periods**	**Natural Disasters**	**Impacts**
USA	1920s	Floods	Health Impacts
China	1930s	Droughts	Loss of Livelihoods
Brazil	1940s	Heatwave	Environmental Damage
Colombia	1950s	Wildfire	Social and Economic Disruption
Peru	1960s	Storms	Loss of Life
Mexico	1970s	Cyclones	Property Damage
India	1980s	Hurricanes	
Indonesia	1990s	Tsunami	
Malaysia	2000s	Earthquakes	
Philippines	2010s	Avalanches	
Vietnam		Typhoons	
France		Blizzards	

3.2. Multiclass Text Classification using Multinomial Logistic Regression

The classifiers were trained using AWS Comprehend's(https://docs.aws.amazon.com/comprehend/index.html) custom classification capabilities, employing the multi-class mode with multinomial logistic regression as the learning algorithm.
Logistic regression is an important data analysis technique in the field of Machine Learning (ML) that attempts to determine the correlations between two data variables. The relationship is then used to forecast the value of one of the variables based on the other. Usually, predictions have a limited number of outcomes, such as yes or no.

Logistic regression function
Logistic regression uses a logistic function or logit function as an equation between the x and y variables. The logit function maps y as a sigmoid function of x. The variable x is often used to represent the independent or predictor variables while y represents the dependent or outcome variable. The independent variables, denoted as x, are the inputs used to predict the probability of a binary outcome represented by the dependent variable y.

$$(1) f(x) = \frac{1}{1 + e^{-x}}$$

The logit function returns only numbers between 0 and 1, regardless of the values of the independent variable. This is how logistic regression calculates the dependent variable's value. Logistic regression methods can also be used to model equations involving numerous independent variables and a single dependent variable.

Logistic regression analysis with multiple independent variables

Multiple explanatory variables often influence the value of the dependent variable. Logistic regression formulas presuppose a linear relationship between the various independent variables in order to model such input datasets. The sigmoid function may be changed, and the final output variable is calculated as follows:

(2) $y = f(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n)$

The symbol β represents the regression coefficient. The logit model can reverse calculate these coefficient values when a sufficiently large experimental dataset is given with known values of both dependent and independent variables.

Multinomial Logistic Regression

For the purposes of this study, multinomial logistic regression is applied. As long as the number of alternative outcomes is finite, multinomial regression can be used to assess issues with many outcomes. Multinomial logistic regression is a machine learning technique for handling multi-class classification problems. It is an extension of binary logistic regression that handles problems with more than two classes and estimates the probability of each class to represent the link between predictors and multiple categories. The algorithm generates individual logistic regression models for each class and compares them to a reference class. It converts the outputs of these models into probabilities using a softmax function **σ**:

$$(3)\,\sigma(\overrightarrow{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

where:

- **z**: Input vector of the given data.
- e^{z_i}: Exponential function for the input vector.
- e^{z_j}: Exponential function for the output vector.
- **K**: Number of classes.

Multinomial logistic regression was selected as the learning algorithm during the training phase of the customized classifiers because of its ability to handle multi-class situations. The training consists in matching input sample documents with corresponding classes. The algorithm identified relations and patterns between textual features and the target class and optimized the weights to reduce the loss function.

When it comes to creating unique classifiers, multinomial logistic regression has several benefits. The technique permits the accurate classification of information into predetermined classes (locations, time periods, natural disasters, and impacts of natural disasters). Furthermore, by predicting the probability for each class, it enables not only the classification of documents, but also the identification of the level of confidence associated with each prediction.

4. TRAINING PROCESS AND RESULTS

In this section the custom classification using Amazon Comprehend, a NLP service provided by Amazon Web Services (AWS) is presented.

4.1. Training and validation processes

A two-step process was implemented to train the custom models. For each classifier, the classes were first created by carefully selecting the labels that best reflected the intended notions. Subsequently, sample documents were provided to demonstrate the various patterns and variations that were anticipated to be encountered in the text data from the input sources.

The two categories of classifier models supported by Amazon Comprehend are plain-text models and native document models. Plain-text models were chosen for the present research as they categorize documents solely based on their textual content. By making this decision, we were able to take advantage of Amazon Comprehend's built-in features while focusing on comprehending the textual material.

In order to train the custom classification models, a dataset of UTF-8-encoded text documents was provided. The training materials for each classifier were deliberately confined to the English language to ensure accuracy and uniformity. The performance of the model depends on the learning algorithm used during training. Its main job is to figure out the model's weights, which indicate how likely it is that the patterns in the training data accurately reflect the underlying relations. Multinomial logistic regression was implemented in our case as discussed in the previous chapter.

Two essential parts of the learning algorithm are a loss function and an optimization method. The penalty calculated when the model's predictions differ from the actual target labels is measured by the loss function. Three loss functions are offered in Amazon Machine Learning, each of which is intended to handle a certain prediction issue. The multinomial logistic loss function was chosen for the present multi-class classification challenge to assess the difference between the predicted and the actual class labels. Stochastic Gradient Descent (SGD) was used as the optimization strategy to reduce the loss and enhance model performance. During successive passes over the training data, SGD incrementally modifies the feature weights. The objective is to repeatedly approach the ideal weights that reduce loss. By updating the weights one example at a time, SGD allows the model to adapt to the training data and generalize well to unseen documents.

4.2. Performance evaluation of the text classification model

Instead of being limited to overfitting by simply memorizing the data presentedduring training, the ML model seeks to discover patterns that generalize well to new data. Once the model is complete, it is crucial to observe how it performs on cases that were not used in training the model. To achieve this, the projected target is first compared to the actual answer by using the model to predict the answer on the evaluation dataset. The prediction accuracy of a model is measured using a variety of measures in machine learning. The ML task determines the accuracy metric to use. Reviewing these indicators is crucial for determining the performance of the model.

The effectiveness of the customized classification models was evaluated using the following metrics:

- Precision estimates the proportion of correctly predicted instances for a given class and is a measure of the model's accuracy in identifying true positives while minimizing false positives.
- Accuracy estimates the proportion of examples that were correctly classified across all classes and thus is a measure of the total correctness of predictions.
- Micro-precision estimates the total true positives, false positives, and false negatives across all classes and is considered a global measure of precision.
- Recall estimates the proportion of properly predicted instances for a particular class among all the actual instances of that class and is considered a measure of the model's ability to identify true positives while minimizing false negatives.

- Micro-recall estimates the total true positives, false positives, and false negatives across all classes and is considered an overall metric of recall.
- The harmonic mean of precision and recall, or F1-score, balances the two metrics into a single measure that may be used to evaluate model performance for each class.
- Micro F1-score estimates the total true positives, false positives, and false negatives across all classes and provides a global measure of F1-score.
- Hamming loss estimates the proportion of incorrectly categorized labels across all classes and is considered as a measure of label prediction accuracy.

These metrics allow for a thorough assessment of the efficiency and reliability of the specific classification models (Table 2).

Overall, the evaluation of each classifier offers some insight into how effectively it performs (Figure 2).

Table 2. Performance of Classifiers

1st Classifier - Locations								
Accuracy	**Precision**	**Recall**	**F1 Score**	**Hamming loss**	**Micro Precision**	**Micro Recall**	**Micro F1 Score**	**Test Data Source**
0.98	0.98	0.98	0.98	0.0167	0.98	0.98	0.98	Autosplit - 20%
2nd Classifier - Natural Disasters								
Accuracy	**Precision**	**Recall**	**F1 Score**	**Hamming loss**	**Micro Precision**	**Micro Recall**	**Micro F1 Score**	**Test Data Source**
0.90	0.93	0.90	0.90	0.0909	0.90	0.90	0.90	Autosplit - 20%
3rd Classifier - Impacts of Natural Disasters								
Accuracy	**Precision**	**Recall**	**F1 Score**	**Hamming loss**	**Micro Precision**	**Micro Recall**	**Micro F1 Score**	**Test Data Source**
0.98	0.98	0.98	0.98	0.0167	0.98	0.98	0.98	Autosplit - 20%
4th Classifier - Time Periods								
Accuracy	**Precision**	**Recall**	**F1 Score**	**Hamming loss**	**Micro Precision**	**Micro Recall**	**Micro F1 Score**	**Test Data Source**

0.80	0.85	0.80	0.80	0.20	0.80	0.80	0.80	Autosplit - 20%

The Locations and Impacts of Natural Disasters classifiers led to identical results exhibiting high performance across all metrics, which indicates their reliability in detecting instances related to their respective classes. More specifically, the two classifiers resulted in:

- High accuracy (0.98) which indicates accurate classification of instances related to locations and impacts of natural disasters.
- High precision (0.98) which indicates a significant number of correctly predicted instances and reduced false positives.
- High recall (0.98) which demonstrates their capacity to identify a significant number of true positives while limiting false negatives.
- High F1 score (0.98) which indicates a balanced performance in terms of recall and precision.
- Low hamming loss value (0.0167) which indicates few misclassifications across all classes.
- High micro precision (0.98), micro recall (0.98), and micro F1 (0.98) scores, which were all 0.98 which confirm the consistent level of performance across all classes.

The performance of the Natural Disasters classifier was satisfactory for all evaluated measures. More specifically, the classifier resulted in:

- A 0.90 accuracy score which indicates that the model was able to correctly classify instances of natural disasters.
- A 0.93 precision score which indicates a high proportion of correctly predicted events, while minimizing false positives.
- A 0.90 recall score which also indicates the ability of the model to detect a sizeable number of true positives while limiting false negatives.
- A 0.90 F1 score which indicates a balanced performance in terms of precision and recall.
- A 0.0909 hamming loss value highlights that the classifier resulted in some misclassifications across classes.
- The 0.90 scores for the micro precision, micro recall, and micro F1 measures indicate a consistent performance across all classes.

On the contrary, the Time Periods classifier resulted in modest performance for all tested measures signifying the need for improvement in the way time-related instances are classified:

- The 0.80 accuracy score indicates relatively fewer accurate time period classifications.
- The 0.85 precision score indicates that a sufficient proportion of instances for the Time Periods class were correctly classified, despite the presence of some false positives.
- The 0.80 recall score indicates that the classifier identified a sufficient number of true positives while limiting false negatives.
- The 0.80 F1 score indicates a balanced performance in terms of recall and precision.
- The 0.20 hamming loss score indicates that the classifier may have incorrectly classified several instances.
- The 0.80 scores for the micro precision, micro recall, and micro F1 scores indicate a uniform performance of the classifier across all classes.

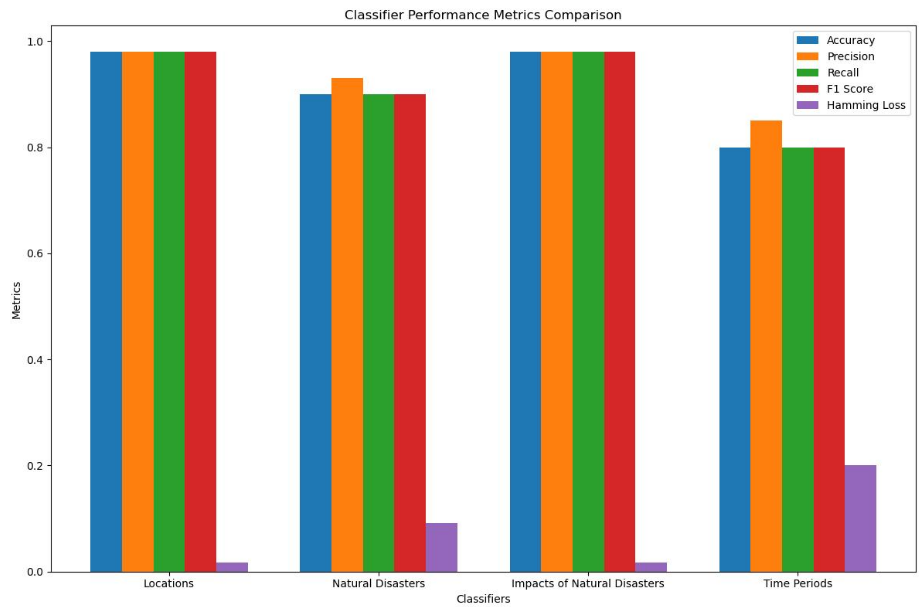


Fig.2. Classifier Performance Comparison Chart

4.3. Analysis of predictions for different classes

The ability of the custom classification models to effectively classify climate change texts was evaluated using an extended report with diversified content and complex vocabulary from the Intergovernmental Panel on Climate Change [16]. The report includes 18 chapters, is 3,068 pages long and 396 megabytes in size. Due to computational constraints the dataset was divided randomly into five equally sized subsets. This approach allowed for testing of each classifier on different subsets, facilitating the comparison of the results among them.
The classifiers' performance in classifying specific locations, natural disasters, impacts, and time-periods is shown in Table 3.

Based on the findings obtained from testing the Locations classifier on the five subsets of the report, USA was identified as the most likely location in subset 1, followed by Colombia, China, and India. In subset 2, USA, along with China, Colombia, and Mexico, emerged as the top anticipated locations. The top projected locations in subset 3 were India, Indonesia, Philippines, Colombia, and China. The most likely destinations predicted for subset 4 were China, India, Philippines, Brazil, and Colombia. USA was the leading location in subset 5, followed by France, Colombia, and China.

Based on the results of testing the Natural Disaster classifier on the five subsets of the reports, our study revealed consistent identification of natural disasters (Table 3). The classifier assigned scores to probable natural disasters mentioned in the report and showed consistent performance across all subsets. In subset 1, the classifier identified droughts as the most likely event followed by floods, earthquakes, tsunamis, storms, wildfires, and avalanches. The top anticipated disaster in subset 2 was droughts as well, followed by floods, earthquakes, storms, tsunamis, wildfires, and avalanches. The most often predicted natural disasters in subset 3 were droughts, wildfires, storms, tsunamis, earthquakes, floods, and avalanches. The most likely natural disasters identified in subset 4 were droughts, storms, tsunamis, wildfires, floods, and avalanches. In subset 5, the classifier identified droughts as the most likely natural disaster, followed by storms, wildfire, tsunamis, earthquakes, and avalanches.

Regarding the results of the Impacts of Natural Disasters classifier, social and economic disruption was the most likely impact in subset 1, followed by loss of life, property damage, and environmental harm. Similar results occurred for subsets 2 and 3: social and economic disruption was identified as the most likely impact, followed by loss of life, loss of livelihoods, and property damage. The most likely impacts identified in subset 4 were social and economic disruption, loss of life, and health impacts. In subset 5 social and economic disruption was also identified as the most likely impact, followed by loss of life and loss of livelihoods.

Regarding the results of the Time Periods classifier, the 2000s was identified as the most probable decade in subset 1, followed by the 1990s, 2010s, 1980s, and 1940s. In subset 2 the 2000s emerged as the most likely time period, followed by the 2010s, 1990s, 1970s, and 1980s. In subset 3, the 1990s, 2000s, 2010s, 1980s, 1970s, and 1940s were the top predicted time periods. Subset 4 yielded predictions of the 1940s, 2000s, 2010s, 1990s, 1980s, and 1970s as the most probable time periods. Lastly, subset 5 featured the 2000s as the most prominent time period, followed by the 2010s, 1940s, 1990s, and 1970s.

Table 3: Predictions of Subsets

Subset 1

Locations	Location_Probabiblity	Impacts	Impact_Probability
USA	68,00%	Social and Economic Disruption	63,00%
Colombia	14,00%	Loss of Life	23,00%
China	13,00%	Property Damage	11,00%
India	5,00%	Environmental Damage	3,00%

Natural Disasters	Natural_Disaster_Probability	Time Periods	Decade_Probability
Droughts	41,00%	2000s	57,00%
Floods	15,00%	1990s	21,00%
Earthquakes	13,00%	2010s	12,00%
Tsunami	12,00%	1980s	9,00%
Storms	9,00%	1940s	1,00%
Wildfire	8,00%		
Avalanches	2,00%		

Subset 2

Locations	Location_Probabiblity	Impacts	Impact_Probability
USA	88,00%	Social and Economic Disruption	57,00%
China	7,00%	Loss of Life	24,00%
Colombia	3,00%	Loss of Livelihoods	12,00%

Mexico	2,00%	Property Damage	7,00%

Natural Disasters	Natural_Disaster_Probability	Time Periods	Decade_Probability
Droughts	35,00%	2000s	41,00%
Floods	26,00%	2010s	33,00%
Earthquakes	15,00%	1990s	10,00%
Storms	9,00%	1970s	9,00%
Tsunami	8,00%	1980s	7,00%
Wildfire	5,00%		
Avalanches	2,00%		

Subset 3

Locations	Location_Probabiblity	Impacts	Impact_Probability
India	38,00%	Social and Economic Disruption	79,00%
Indonesia	22,00%	Loss of Life	15,00%
Philippines	15,00%	Loss of Livelihoods	6,00%
Colombia	13,00%		
China	12,00%		

Natural Disasters	Natural_Disaster_Probability	Time Periods	Decade_Probability
Droughts	27,00%	1990s	30,00%
Wildfire	22,00%	2000s	25,00%
Storms	21,00%	2010s	22,00%
Tsunami	13,00%	1980s	13,00%
Earthquakes	8,00%	1970s	7,00%
Floods	7,00%	1940s	3,00%
Avalanches	2,00%		

Subset 4

Locations	Location_Probabiblity	Impacts	Impact_Probability
China	42,00%	Social and Economic Disruption	49,00%
India	21,00%	Loss of Life	33,00%
Philippines	16,00%	Health Impacts	18,00%
Brazil	12,00%		
Colombia	9,00%		

Natural Disasters	Natural_Disaster_Probability	Time Periods	Decade_Probability
Droughts	78,00%	1940s	31,00%
Storms	15,00%	2000s	27,00%
Tsunami	4,00%	2010s	18,00%
Wildfire	2,00%	1990s	12,00%
Floods	1,00%	1980s	8,00%
		1970s	4,00%

Subset 5

Locations	Location_Probabiblity	Impacts	Impact_Probability
USA	61,00%	Social and Economic Disruption	69,00%
France	36,00%	Loss of Life	19,00%
Colombia	2,00%	Loss of Livelihoods	12,00%
China	1,00%		
Natural Disasters	**Natural_Disaster_Probability**	**Time Periods**	**Decade_Probability**
Droughts	52,00%	2000s	39,00%
Storms	23,00%	2010s	23,00%
Wildfire	14,00%	1940s	16,00%
Tsunami	8,00%	1990s	13,00%
Earthquakes	3,00%	1970s	9,00%

Figure 3 shows the map-based visualization for the prediction of locations across the five subsets. The most anticipated locations, which include USA, Colombia, China, and India, have been identified by the classifiers almost across all subsets. Figure 4 shows the polar plots that represent the predicted classes for the rest classifiers (time periods, natural disasters, impacts) for the five subsets and provide a visually concise way to discern and compare the distribution of probabilities across subsets of natural disasters, time periods, and impacts. Each subset's polar plot illustrates the expected probabilities of natural disasters, impacts, and time periods.

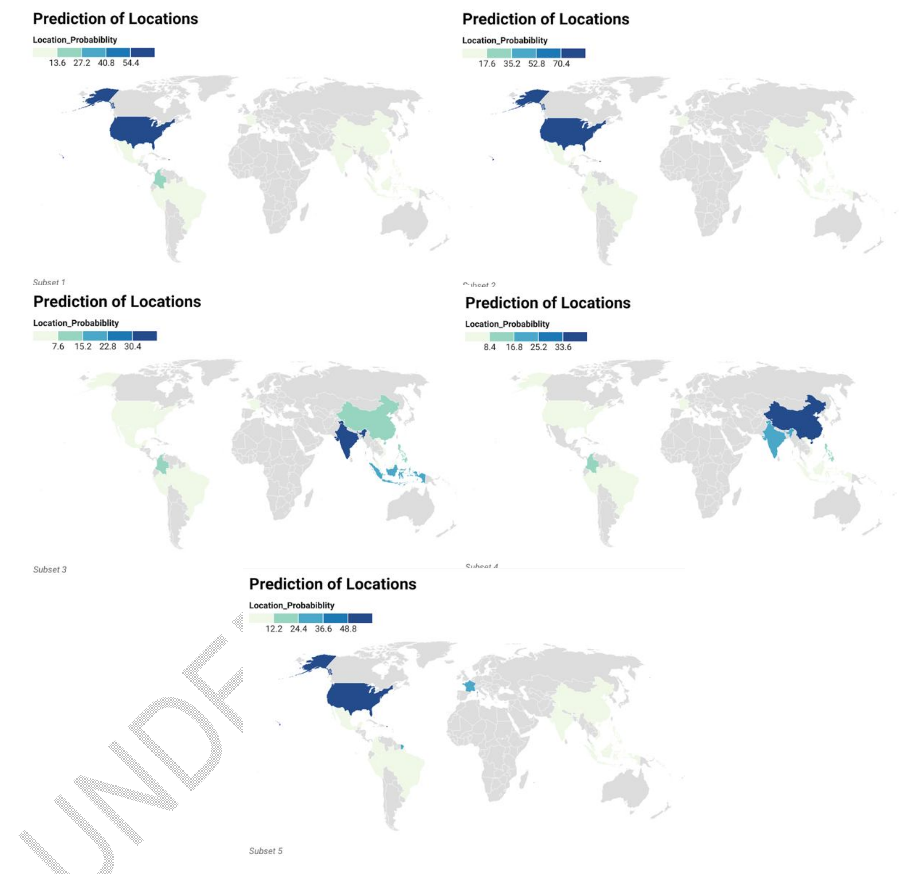
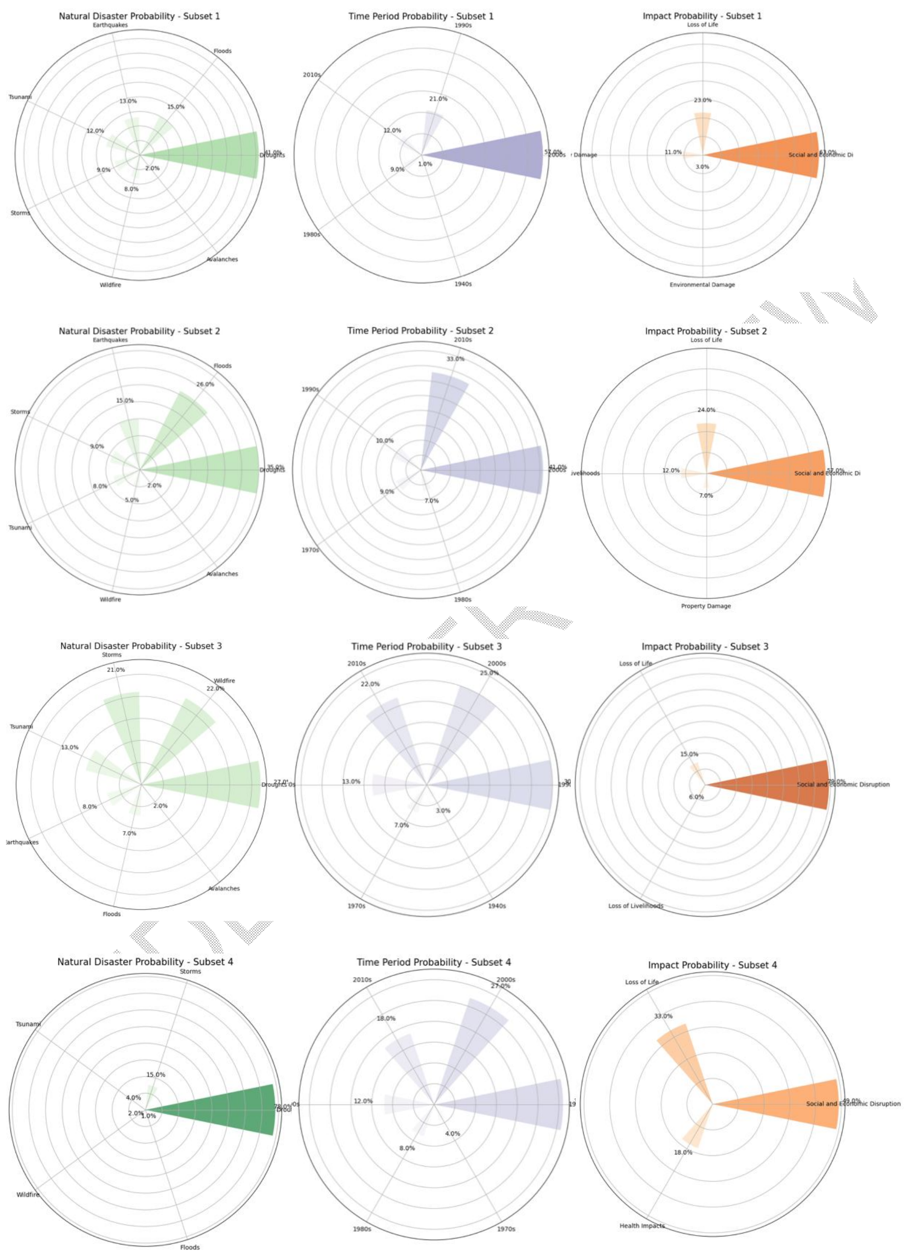


Fig. 3. Maps representing the prediction of locations for the five subsets.

Comment [G1]: Need some field study photos for validation


Natural Disaster Probability - Subset 1
Earthquakes
Floods
Tsunami
Droughts
Storms
Wildfire
Avalanches
13.0%
15.0%
12.0%
9.0%
2.0%
8.0%
Time Period Probability - Subset 1
1990s
2010s
2000s
1980s
1940s
21.0%
12.0%
9.0%
1.0%
Impact Probability - Subset 1
Loss of Life
Environmental Damage
23.0%
11.0%
3.0%
Natural Disaster Probability - Subset 2
Earthquakes
Floods
Storms
Droughts
Tsunami
Wildfire
Avalanches
15.0%
26.0%
9.0%
8.0%
5.0%
2.0%
Time Period Probability - Subset 2
2010s
1990s
2000s
1970s
1980s
33.0%
10.0%
9.0%
7.0%
Impact Probability - Subset 2
Loss of Life
Property Damage
24.0%
12.0%
7.0%
Natural Disaster Probability - Subset 3
Storms
Wildfire
Tsunami
Earthquakes
Floods
Avalanches
21.0%
22.0%
13.0%
8.0%
7.0%
2.0%
Time Period Probability - Subset 3
2010s
2000s
1970s
1940s
22.0%
25.0%
13.0%
7.0%
3.0%
Impact Probability - Subset 3
Loss of Life
Social and Economic Disruption
Loss of Livelihoods
15.0%
6.0%
Natural Disaster Probability - Subset 4
Storms
Tsunami
Wildfire
Floods
15.0%
4.0%
2.0%
1.0%
Time Period Probability - Subset 4
2010s
2000s
1980s
1970s
18.0%
27.0%
12.0%
8.0%
4.0%
Impact Probability - Subset 4
Loss of Life
Social and Economic Disruption
Health Impacts
33.0%
18.0%

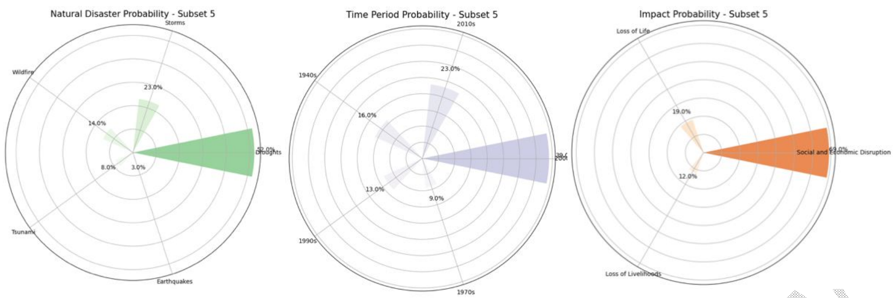


Fig. 4. Classification Results for Natural Disasters, Time-Periods, and Impacts.

5. CONCLUSION

The paper focused on the implementation of multinomial logistic regression for text classification in order to predict the class of a document based on its semantic content. Four custom classifiers were implemented for climate change reports that are tailored to identify and categorize natural disasters and their impacts, locations, and time periods.

The IPCC climate change report [16] was used to evaluate the classifiers' reliability in detecting and classifying specific locations, natural disasters, impacts, and time-periods in the context of climate change literature. The predictions from the classifiers highlighted various aspects of the reports regarding these four dimensions. Countries such as USA, China, India, Colombia, and France that are featured in recent discussions on climate change have been consistently identified in the report. The classifiers have also identified various types of natural disasters, such as droughts, tsunamis, wildfires, and earthquakes, with droughts being the top predicted natural disaster class. The most prominent identified impacts were social and economic disruption, loss of life and loss of livelihoods, while the most frequently mentioned time periods were the 2000s and 2010s.

However, climate change is a multifaceted issue with several complex, interacting dimensions. Future research could delve into understanding the complex relationships and interactions between the different dimensions of climate change by refining and expanding the existing classifier models used in this study. Developing alternative algorithms and methods could enhance text classification performance and accuracy, while adding further features and data sources, such as social media data or climate models could improve the classifiers' predicting powers. Advanced data analytics techniques regarding text classification such as cluster, behavioral and predictive analytics, using complex machine and especially deep learning algorithms or causal modeling may also be applied to uncover patterns and complex relations between locations, natural disasters, and their impacts.

REFERENCES

1.Begum, R.A., R. Lempert, E. Ali, T.A. Benjaminsen, T. Bernauer, W. Cramer, X. Cui, K. Mach, G. Nagy, N.C. Stenseth, R. Sukumar, and P. Wester, 2022: Point of Departure and Key Concepts. In: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama

(eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 121-196, doi:10.1017/9781009325844.003.
2. Vijayan, V. K., Bindu, K. R., & Parameswaran, L. (2017, September). A comprehensive study of text classification algorithms. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1109-1113). IEEE.
3. Mirończuk, M. M., &Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106, 36-54.
4. Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018, August). Comparative study between traditional machine learning and deep learning approaches for text classification. In Proceedings of the ACM Symposium on Document Engineering 2018 (pp. 1-11).
5. Altınel, B., &Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. Information Processing & Management, 54(6), 1129-1153.
6. Kunnath, S. N., Herrmannova, D., Pride, D., & Knoth, P. (2021). A meta-analysis of semantic classification of citations. Quantitative Science Studies, 2(4), 1170-1215.
7. Hu, Y. (2017). Geospatial semantics. arXiv preprint arXiv:1707.03550.
8. Karimi, M., Mesgari, M. S., & Purves, R. S. (2022). Transactions in GIS, 26, 3225–3252. https://doi.org/10.1111/tgis.12999
9. Yang, L., MacEachren, A. M., Mitra, P., & Onorati, T. (2018). Visually-enabled active deep learning for (geo) text and image classification: a review. ISPRS International Journal of Geo-Information, 7(2), 65.
10. Häberle, M., Werner, M., & Zhu, X. X. (2019). Geo-spatial text-mining from Twitter–a feature space analysis with a view toward building classification in urban regions. European journal of remote sensing, 52(sup2), 2-11.
11. Burel, G., Saif, H., Fernandez, M., & Alani, H. (2017). On semantics and deep learning for event detection in crisis situations.
12. Wang, Z., & Ye, X. (2018). Social media analytics for natural disaster management. International Journal of Geographical Information Science, 32(1), 49-72.
13. Said, N., Ahmad, K., Riegler, M., Pogorelov, K., Hassan, L., Ahmad, N., & Conci, N. (2019). Natural disasters detection in social media and satellite imagery: a survey. Multimedia Tools and Applications, 78, 31267-31302.
14. Ye, P., Zhang, X., Huai, A., & Tang, W. (2021). Information detection for the process of typhoon events in microblog text: A spatio-temporal perspective. ISPRS International Journal of Geo-Information, 10(3), 174.
15. CRED. 2022 Disasters in numbers. Brussels: CRED; 2023.
16. IPCC, 2022: Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press. Cambridge University Press, Cambridge, UK and New York, NY, USA, 3056 pp., doi:10.1017/9781009325844.