Seasonal Analysis and Develop Text Classification Model using Earthquakes Online Report Data

Seonho Woo
clairewo@umich.edu

1 Introduction

Earthquakes are natural disasters that have the potential to cause widespread destruction and loss of life. As such, predicting and understanding their occurrence is crucial for mitigating their impact. One approach to studying earthquakes is seasonal analysis, which involves analyzing patterns in earthquake occurrences over time. With the vast amount of information available in online news articles, text extraction techniques can be used to extract data relevant to seasonal analysis. This research explores the effectiveness of text extraction from online news articles for seasonal analysis in earthquake prediction and understanding. By doing so, it may be possible to improve earthquake prediction or explore relationships between multiple attributes (magnitude, location, season, year) and inform disaster management strategies.

2 Data

2.1 Web Crawling

To get earthquake reports online without bias, I selected an online news archive with many links and online articles available, which is *Newslookup.com*. First, I crawled the text data of online reports from the HTML text of each website and connected it to the keyword search result ("earthquake," in this case) from the website. I would like to observe the seasonal pattern during a single year so that one year of data has been extracted. (In future research, more years will be used for larger data size and cross-validation purposes).

2.2 Body Text Extraction

All article bodies will be saved, containing their time and location, and their temporal information is also extracted and saved in the collected/processed dataset, which is used as the data source for classification models. Once all data collection steps were done, I used those data to: First, split them into seasonal data of earthquake reports in Spring, Summer, Fall, and Winter; and second, track the location of occurrence using a location tagger. Data are split into four seasons using these standards: Spring (March to May), Summer (June to August), Fall (September to November), and Winter (December to February). And the results from this data extraction for capturing numerical data and location information are stored in a single big dataframe and also four different data frames corresponding to each season.

2.3 Text Classification Data

To conduct and evaluate the text classification model, I selected one season's (Fall 2020) data that has a size of 200 data points approximately and split them into three sets of data: 60% for training, 20% for evaluation, and 20% for the test. To generate ground truth data for this set, I manually label and annotate their scores into 0 or 1, whether it is a real report of an earthquake or not. This is comparatively small data to train the model, but due to lacking time, only 200 data could be annotated and used to train the model.

2.4 Spatial Analysis in Different Seasons

Since I have extracted the locational data using a location tagger package, I have the latitude and longitude information on the map for all of the cities that are reported earthquakes. Using the US Cities coordinate dataset, which is retrieved from this US Cities Database website[1], the city names from location tagging can be cross-matched with the US Cities coordinate dataset to pick up each city's coordinate data (their latitude and longitude). This information is also stored in the original dataframe that can be used to be marked on the map.

3 Related Work

3.1 Spatial and temporal pattern of wildfires in California from 2000 to 2019 [2]

This paper covers various spatial and temporal pattern analysis methods in wildfire cases in California (the state of the most frequent wildfire cases). Since this analysis addresses a geometrical feature of California, I might refer to their approach to how to use geometrical information.

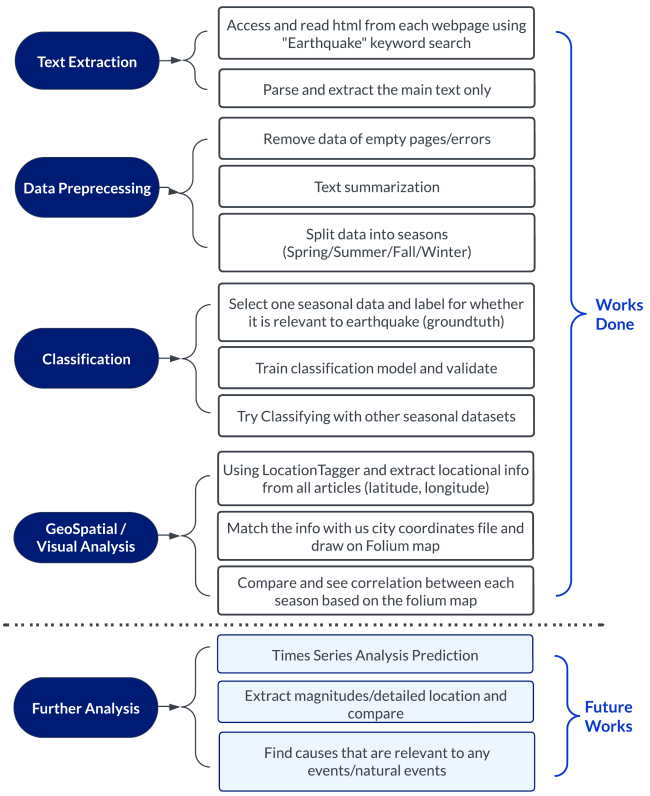
3.2 Analysis of Online News Coverage on Earthquakes Through Text Mining [3]

This paper also used text mining from international news reports to generate the geometrical maps of Earthquakes and do a frequency analysis of occurrence. It handles how to deal with online news text extraction specifically for a natural disaster; it will be a good resource to look up.

3.3 Casualty Information Extraction and Analysis from News [4]

This was referred to in order to see how previous studies extract casualty or damage information from text analysis (tokenized text, specifically) from the news articles.

4 Methods



As shown in the flowchart below, I first extracted the body text from online news articles that are accessed via URLs and links in the news archive. I conducted some tasks to cleanse the dataset to have fewer junk characters or articles with non-accessible links. As a part of data preprocessing, I tried text summarization on the body text since some of the articles do include a lot of redundant or broad information about the locations and the past occurrences, and not only about the current report that is talking about. Also, it is helpful in doing model training since it has a shorter length of the main text that is easy to read and interpret by the machine.

Regarding the development of the Classification model, I first get the datasets and preprocess them by encodings, then feed them into training argument models to classify between labels 0 and 1 based on their relevancy to the actual earthquake occurrences. As stated in the Data section, ground truth data annotated by myself was used to train the model, and F1 scores and other metrics in the confusion matrix are used for evaluating the model performance.

A series of the tasks above was done to develop an effective and accurate information retrieval model while I was diagnosing the seasonal pattern if they have a rough pattern or noticeable correlation between their locations and frequency. As this part was done by using location information that had been extracted from the LocationTagger package and cross-checked with the US cities database, and their longitude and latitude information are available from the steps described above, I could illustrate and mark on the Folium map to visually compare the trend and likelihood of earthquake occurrences in some regions.

5 Evaluation and Results

5.1 Text Classification Model

The following table (Table 1) shows the classification report scores from the confusion matrix.

Scores	
Recall	0.85
Precision	0.92
F1	0.89
Accuracy	0.92

Table 1: Classification Report on test data of Classification Model

As each score shows different aspects of performance metrics, from the most used score F1 (which has a value of 0.89 in this model implementation), I can deduce that this model performs quite well on this testing dataset. However, taking account of the data size and nature, the ground truth data is imbalanced in having most of the label values of 1, as only 5% of the total data in the test dataset are zero-valued. As it is biased on positive cases, it might have a high True Positive (TP) rate, which refers to a sample belonging to the positive class being classified correctly, and a high False Positive (FP) rate, which refers to a sample belonging to the negative class but being classified wrongly as belonging to the positive class. As expected, this rough trend is found in the dataset that there are a very small number of negative cases to train the model to perform well enough to predict its case.

(Continued in the next page)

6 Discussion

6.1 Seasonal Analysis on Geospatial Pattern

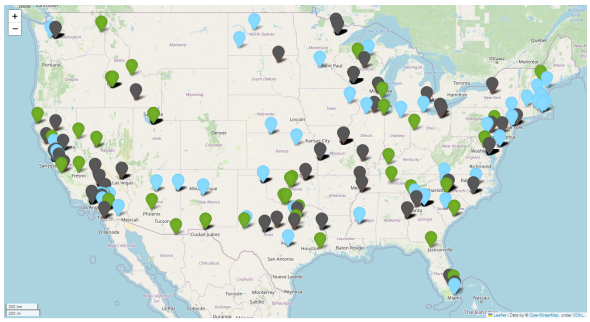


Figure 1: Combined folium map of all seasonal earthquake reports in 2020

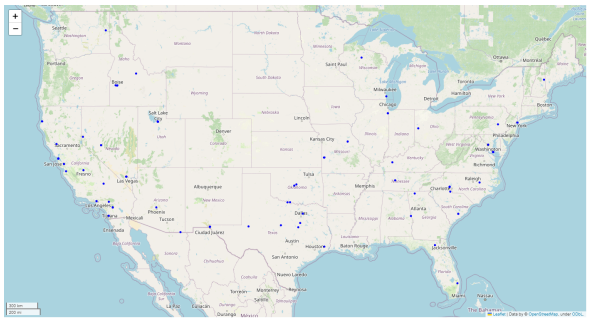


Figure 2: Map of earthquake reports in Spring 2020

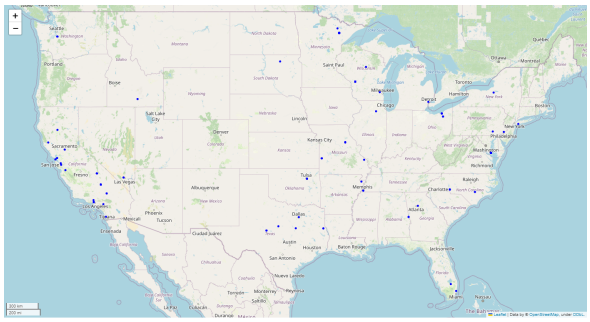


Figure 3: Map of earthquake reports in Summer 2020

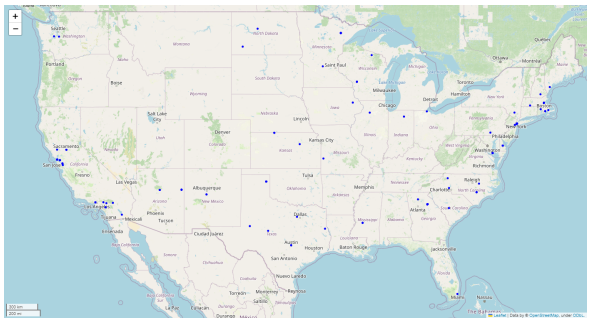


Figure 4: Map of earthquake reports in Fall 2020

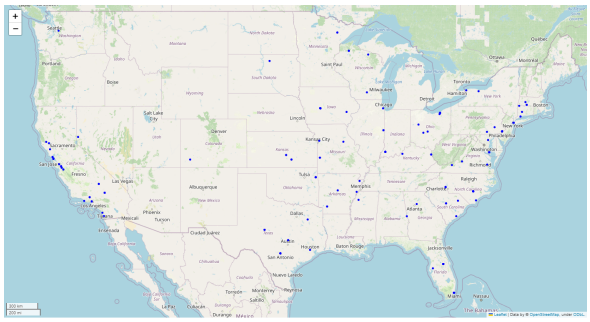


Figure 5: Map of earthquake reports in Winter 2020

As illustrated in Figure 1, the green mark indicates Spring, the red mark represents Summer, the blue mark shows Fall, and the gray mark indicates Winter earthquake occurrences. We cannot find a red mark on the map, but it is hidden behind the gray marks, which could imply that the occurrences pattern between winter and summer are highly correlated.

As geometrical regions on the West Coast and East Coast are more likely to occur earthquakes including some Mideast regions based on some decent reasons (i.e., California tends to experience more earthquake cases since it sits on top of the meeting point, or fault, of two plates that it is more likely to scrape against each other) regardless of their seasons.

Regarding their seasonal difference in occurrence patterns, apart from their geospatial characteristics, it shows some scattered shapes in the occurrence area during spring. It is less likely to occur in the very east coast area(i.e., Washington and New York States). However, during the summer, its red marks are all hidden behind some green and gray marks, and this is interesting or possibly because some of the dates of the event lie in the border of summer and fall or summer and spring seasons, such as May 31st and June 1st, or September 30th and October 1st. Opposed to the marks on the map, summer is generally known (early summer especially) and empirically (from this data) as the season with the most earthquakes occurring. It is theoretically not true to have a tendency of earthquakes depending on the season, but recently, there have been some abnormally high temperatures and some outliers in pressure due to climate change are taking place, and this is probably the reason that the summer of 2020 has the most earthquake cases.) Interestingly, during the fall, it shows a different pattern than the other seasons, that there are more cases in the middle of the continent rather than the area closer to the west or east coast, comparatively to other seasons. And during the winter, it has a similar distribution to the Spring season's occurrence that it is an expected result.

7 Conclusion

In short, my text extraction and classification model performed well in the dataset using online earthquake reports. However, it is likely to be not a robust model since I used data with a comparatively small size for training, and the data was very

imbalanced in containing a remarkably large number of positive cases and very rare negative cases that might cause the model to be biased or poor in predicting negative cases with lacking inputs.

Regarding the spatial and seasonal analysis of earthquakes, I could find that more than 30% of earthquakes happened during the summer, at least in 2020, and there are some noticeable different or shift in the location of events during fall compared to other seasons.

8 Things I have tried and what would have done differently

I tried to generate a good information retrieval model for extracting the reports from the online source, which are the actual reports for earthquake occurrences that belong to none of the general information or cause of earthquakes or trends in occurrences based on their location. One trouble I encountered was that I should manually make ground truth data to make training and evaluation datasets to evaluate my model performance to classify the data (body text of articles) based on whether it is a real report. If I had begun this project earlier and stuck to this topic from the very beginning, I would have been able to label all 2000 data with labels of 0 and 1. However, I changed the topic within the last minute of the project deadline so that I was able to do this task only on 200 data approximately. I think this is the potential and probable reason that my model performance seems to be good but cannot be a robust model since it has only been trained with a small amount of data as opposed to the property of deep learning.

If I could start over from the very beginning, I would crawl more data from other years to observe the pattern between years and seasons. I would figure out this ground truth data at my earliest availability. Also, I would figure out the methods or ways to effectively demonstrate and analyze the spatial correlation(i.e., using Lee's L test, haversine, or geopy distance) between two locations and expand this into seasonal analysis and cross-validation from each year's result. I also plan to do further research on Keyword or search-word sensitivity and how the search results from news archives are sensitive to the search words (i.e., is it better to be detailed or broad?) or some keywords that most of them include. Regarding the contents of the news reports, I would find some causations or else-like relationships between earthquake magnitude, location, and season that can generate visual or numerical data that are easily interpreted and used to find meaningful patterns.

As shown in the flowchart in the $Methods$ section, I have tried several methods for getting a geospatial correlation between seasons. Still, I could not make it work until the end, so I would like to analyze this part of my study and am willing to expand and conduct further analysis in this area since there are more possibilities and interesting sources that I could find from the online news reports about earthquakes. There are two aspects that I can dig into: first, with respect to the nature of natural disaster data, and second, in terms of characteristics of online crawled data. And the combination of both could make better reasons to make further analysis in the future.

Acknowledgement and Additional Notes

Just to add some notes on my personal plan and interests, I would like to do research using simulation and NLP in Industrial Engineering. I am starting my Ph.D. program at Purdue University this Fall, and I am very glad to start one project in NLP that I would be eager to further study in the area and expand during my Ph.D. years. I am thankful to my brother, Seungho, who is a 2nd year IE Ph.D. student at Purdue University that he advised me regarding which tools/resources are available to use and what other methods could work on my data.

9 References

[1] United States Cities Database, *SimpleMaps*. Retrieved from https://simplemaps.com/data/us-cities. [2] Li, S., Banerjee, T. Spatial and temporal pattern of wildfires in California from 2000 to 2019. Sci Rep 11, 8779 (2021). https://doi.org/10.1038/s41598-021-88131-9

[3] Camilleri, S., Agius, M. R.,; Azzopardi, J. (2020). Analysis of online news coverage on earthquakes through text mining. Frontiers in Earth Science, 8. https://doi.org/10.3389/feart.2020.00141

[4] Chaulagain, B., Shakya, A., Bhatt, B., & Pandey, R.K. (2019). Casualty Information Extraction and Analysis from News. Spain: Universitat Politecnica de Valencia.

[5] S. Camilleri, J. Azzopardi and M. R. Agius, "Investigating the Relationship between Earthquakes and Online News," 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 2019, pp. 203-

210, doi: 10.1109/AIKE.2019.00043.