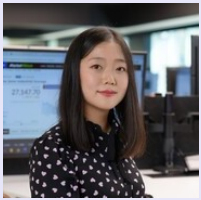
Catastrophe Duration and Loss Prediction via Natural Language Processing

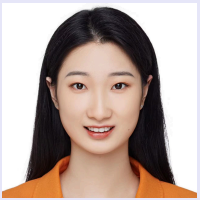
Feng Li

School of Statistics and Mathematics
Central University of Finance and Economics

https://feng.li

Our cactastrophe loss prediction team




Yanfei Kang (Beihang Uni.)
Han Li (Melbourne Uni.)
Han Wang, **Wen Wang**, **Feng Li** (Central University of Finance and Economics)

This research is sponsored by the Casualty Actuarial Society (北美产险精算学会) through the 2022 Individual Grant Competition project "**Catastrophe loss prediction via natural language processing**".

Outline

Catastrophe loss

- Climate change is one of the gravest risks to humanity in the 21st century (Reichstein et al., 2013, *Nature*).
- According to Swiss Re Institute (2019), natural catastrophes caused total economic losses of US$155 billion in 2018, leading to US$76 billion in insurance payouts .
- In 2020, the total economic losses from natural catastrophes rose to US$190 billion, resulting in US$89 billion in insurance payouts (Swiss Re Institute, 2021).

Google

California fires 2023

All | News | Images | Maps | Videos | More | Tools

Past month | Sorted by date | Hide Duplicates | Clear

The Guardian

As Hurricane Hilary prepares to land, California and Mexico ...

Southern California gets first tropical storm warning as conditions could ... extreme heatwaves and severe wildfires in California is increasing due to the...

18 mins ago

City of Malibu

News Flash • Malibu, CA • CivicEngage

The team collected about 200 pounds of trash. staff beach cleanup cm update 8.18.2023. WILDFIRE safety. (NEW) CITY FIRE SAFETY LIAISONS ARE...

1 hour ago

AZCentral

Hurricane Hilary updates: Live coverage as storm hits AZ, CA

Hurricane Hilary is advancing toward California and Arizona, with the ... Fire Station 1: 96 Acoma Blvd. S. (Around north side of the station by the...

LIVE 1 hour ago

Reuters

GM's Cruise robotaxi collides with fire truck in San Francisco

The San Francisco Fire Department did not respond to requests for comment. Advertisement · Scroll to continue. The California Public Utilities Commission (CPUC)...

1 hour ago

Predicting the severity of catastrophes

- Accurately predicting the severity of catastrophes remains a challenge, and requires further scholarly attention and research.
- Natural language processing (NLP), an emerging field in artificial intelligence, provides the ability to automatically read, understand, and derive meaning from text data.
- In this project, we explore the potential of leveraging text data from news articles for predicting the duration and losses of catastrophes.

Multimodal data as a means to understand the severity of catastrophes

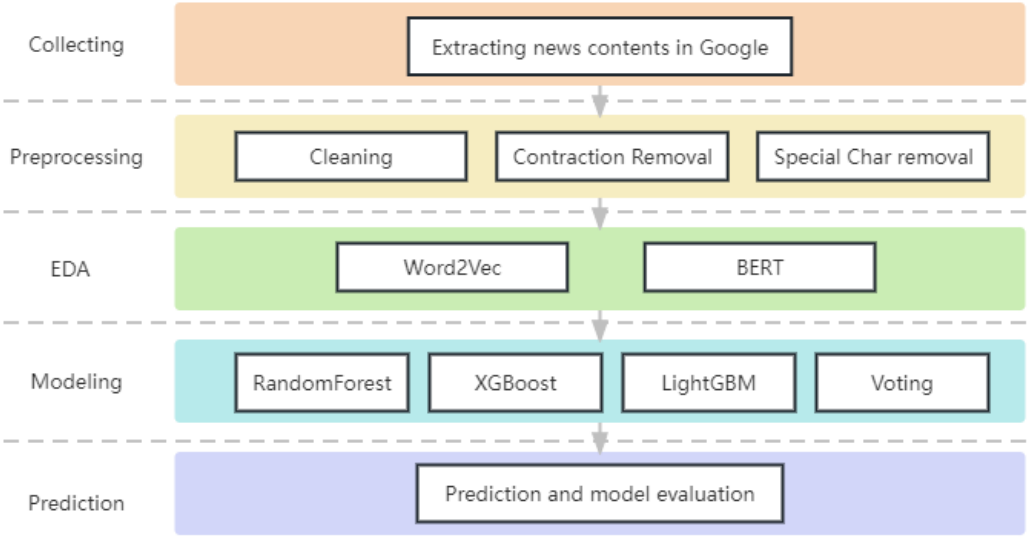
- **Professional evaluations** are always delayed months behind disasters.
- **Satellite data** are not always accessible to all the sectors.
- **Multimodal data** are the new normal.

Primary objective

- Our primary objective is to obtain new insights on how information from online news articles can be utilized to help achieve a **timely estimation of catastrophe duration and insurable economic loss**.
  - Can online news provide an early warning of extreme catastrophe events?
  - What words (or topics) are the most significant indicators of catastrophe severity?
- Answers to these questions will be of particular value to insurance and reinsurance companies with large exposure in catastrophe risk.

Our approach

- **Stage 1: Daily news source for catastrophe events.** We identify and collect relevant news from news aggregator websites such as Google News by searching with a combination of location and catastrophe keywords, and restrict the search to the entire month in which the catastrophe occurred.
- **Stage 2: Full text scraping.** For each news headline displayed in Google, we build a web crawler to extract its news contents with the Python programming language.
- **Stage 3: NLP steps.** We employ standard NLP algorithms to extract features from scraped news contents. We calculate and compute two types of numerical features, namely statistical features and semantic features.
- **Stage 4: Catastrophe duration and loss prediction.** To detect early warnings of severe catastrophe events, one could employ the anomaly detection tools for a given variable of interest from statistical features.


Collecting
Extracting news contents in Google
Preprocessing
Cleaning
Contraction Removal
Special Char removal
EDA
Word2Vec
BERT
Modeling
RandomForest
XGBoost
LightGBM
Voting
Prediction
Prediction and model evaluation

US wildfire and hurricane data

- Our primary data source for catastrophe duration and loss in the US is the Spatial Hazard Events and Losses Database (SHELDUS).
- This database is maintained by the Arizona State University and has already been commonly used by the actuarial community.
- It provides detailed catalogues of all major catastrophe events in the US such as **droughts, earthquakes, floods, hurricanes, thunderstorms, tornadoes, and wildfires**.
- Available information includes the **location of event, duration of event, time of occurrence, type of loss**, as well as the **amount of loss**. We collect catastrophe duration and loss data at the **county-level** for the same time period as the scraped news contents.

US wildfire and hurricane data

Table: Catastrophe data description

Variable	Description
State Name County Name	Catastrophe location
Hazard	Catastrophe type
Year Month	Catastrophe time
CropDmg PropertyDmg	Catastrophe loss of crop and property
Duration_Days	Catastrophe duration
Text	News text information crawled from Google

Category and word frequency

Table: Category and word frequency of keywords for wildfire events

Environment		Actions		Wildfire	
Keyword	Frequency	Keyword	Frequency	Keyword	Frequency
air	1737	evacuation	2545	fire	15018
forest	1590	reported	835	burned	1981
creek	1237	orders	808	wildfire	1896
winds	1101	help	798	burning	1638
valley	1081	service	666	smoke	1637
lake	1071	containment	650	emergency	843
weather	964	evacuate	624	flames	758
climate	631	closed	523	blaze	744
water	605	issued	461	heat	572
canyon	559	control	381	spread	567

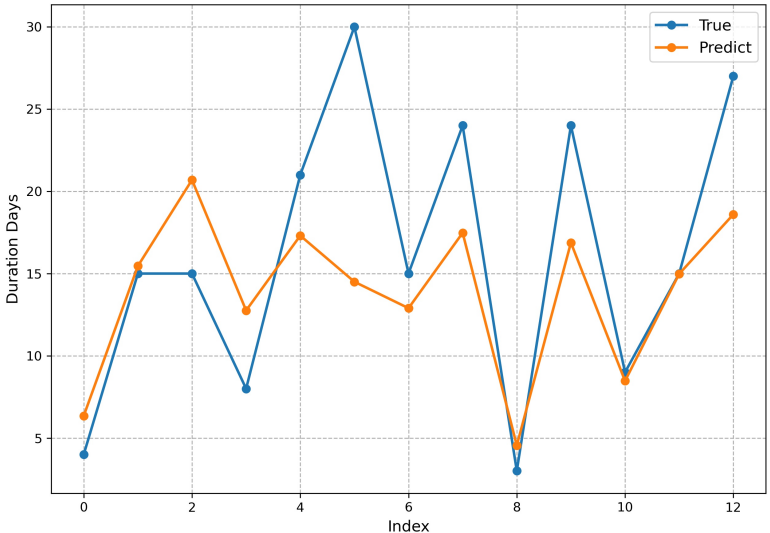
Wildfire duration prediction

- We construct training and test sets through stratified sampling in multiple samples, based on which we fit three classification models using the same method.
- We also utilize the **voting model** (Wang et al., 2022) to combine the prediction results, where the weight of each model is determined by the accuracy of its prediction.
- In terms of evaluation indicators, we selected Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

Table: Prediction accuracy based on Word2Vec and BERT

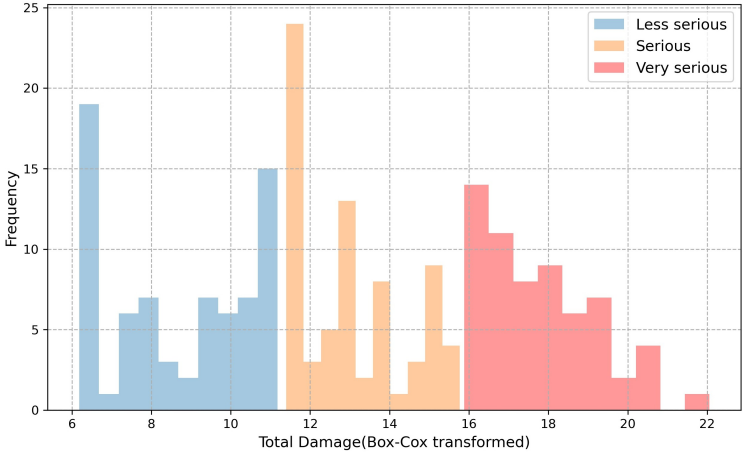
Embedding	Word2Vec		BERT	
Model	RMSE	MAPE	RMSE	MAPE
RandomForest	7.632	0.6552	8.3679	0.6694
XGBoost	7.35	0.4119	6.0531	0.3966
LightGBM	8.275	0.5694	6.8242	0.3577
Voting	**7.3039**	**0.5117**	**6.1121**	**0.2984**

Out of sample voting model prediction effect for BERT



Hurricane loss prediction

Label	Level	Range	Sample size
0	Less serious	$Dmg < 1 \times 10^5$	73
1	Serious	$1 \times 10^5 \leqslant Dmg < 1 \times 10^7$	72
2	Very Serious	$Dmg \geqslant 1 \times 10^7$	62



Hurricane loss prediction

Table: Prediction results for hurricane losses

Model	Actual	Predicted			Precision	Recall	F1-score	Accuracy
		P0	P1	P2				
RandomForest	R0	12	1	2	0.71	0.80	0.75	0.7209
	R1	4	9	2	0.75	0.60	0.67	
	R2	1	2	10	0.71	0.77	0.74	
XGBoost	R0	10	3	2	0.62	0.67	0.65	0.6744
	R1	4	10	1	0.67	0.67	0.67	
	R2	2	2	9	0.75	0.69	0.72	
LightGBM	R0	10	5	0	0.71	0.67	0.69	0.6744
	R1	4	9	2	0.53	0.60	0.56	
	R2	0	3	10	0.83	0.77	0.80	
Voting	R0	11	2	2	0.65	0.73	0.69	0.6977
	R1	5	9	1	0.69	0.60	0.64	
	R2	1	2	10	0.77	0.77	0.77	

Conclusion

- Our results show that utilizing news text information can effectively improve the accuracy of predicting the severity of catastrophes.
- Both the Word2Vec and BERT methods are effective in extracting information from news text, while the use of machine learning models such as Random Forest, XGBoost, and LightGBM can accurately predict catastrophe duration and loss.
- Both XGBoost and LightGBM models perform well in predicting the severity of catastrophes when using BERT-generated text vectors.
- The combining voting model provides the best prediction performance on wildfire duration, and an accuracy of nearly 70% on hurricane losses.
- The code for this article is based on Python. It is available at: https://github.com/feng-li/catastrophe-loss-prediction-with-NLP.

Thank you!

https://kllab.org
feng.li@cufe.edu.cn

References

- Wang, X., Hyndman, R. J., Li, F. & Kang, Y. (2022). "Forecast Combinations: An over 50-Year Review". *International Journal of Forecasting*.
- Swiss Re Institute (2021). *Sigma 1/2021: Natural catastrophes in 2020*.
- Swiss Re Institute (2019). *Sigma 2/2019: Secondary natural catastrophe risks on the front line*.
- Reichstein, M., Bahn, M., Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., Zscheischler, J., Beer, C., Buchmann, N., Frank, D. C., et al. (2013). "Climate extremes and the carbon cycle". *Nature* 500.(7462), p. 287.