Building an NLP-Powered Repository for Cyber Risk Literature

David An, Linfeng Zhang, Zhiyu (Frank) Quan, PhD
dzan2@Illinois.edu, lzhang18@Illinois.edu, zquan@Illinois.edu

ABSTRACT

With the large and growing body of cyber risk literature, we see three major challenges faced by the actuarial research community: there is no context aware tool for finding cyber literature, no central repository of cyber risk resources, and a lack of accounting of literature trends. To address the abovementioned challenges, we propose to build a repository of cyber-risk articles with an NLP powered search tool that can easily be used by researchers to find relevant materials.

INTRODUCTION

As the world has become more reliant on information technology than ever, especially after the pandemic, cyber risk has received a tremendous amount of attention from practitioners and scholars. Aside from studies on mitigating cyber risk from the engineering and technical perspective, since the time when cyber insurance was first introduced to the market, there has been a rapidly expanding volume of literature that focuses on other aspects of cyber risk, such as the legal and financial consequences of cyber incidents, and they are closely related to the development of the cyber insurance industry.

Goals

- Apply natural language processing (NLP) techniques to classify and group cyber-risk and cybersecurity related academic literature. Using this information, we want to construct a tool for academics and researchers to use to identify trends and new technologies in cyber-risk and cyber security.
- For example, given a text query, a researcher would be able to obtain relevant pieces of literature and topics to use as well as visualize trends and different topic groupings.
- In addition to that, researchers would be able to suggest and add new pieces of literature into the database as well. This would be a growing repository.

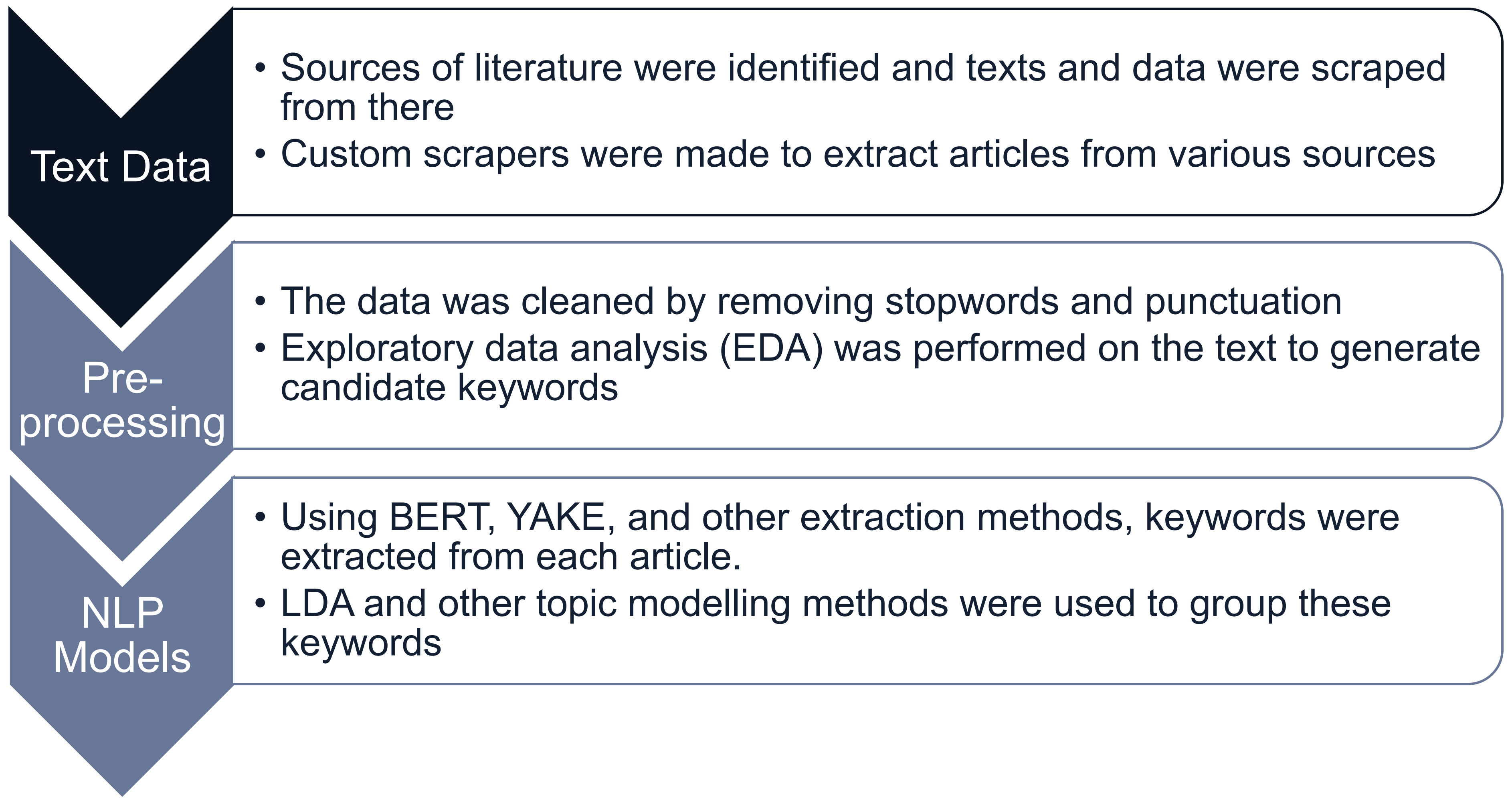
Terminology:

- Natural Language Processing: Natural language processing (NLP) refers to a branch of artificial intelligence relating to giving computers the ability to speak and understand text[1].
- Cybersecurity: Cybersecurity is known as the practic protecting computer systems and networks from information disclosure and data theft.
- Keyword Extraction: Keyword extraction is the process of automatic identification of terms that best capture the meaning of a document.

Tools:

- Libraries: Tensorflow, Scikit-learn, Pandas
- Languages: Python, JavaScript

METHODS

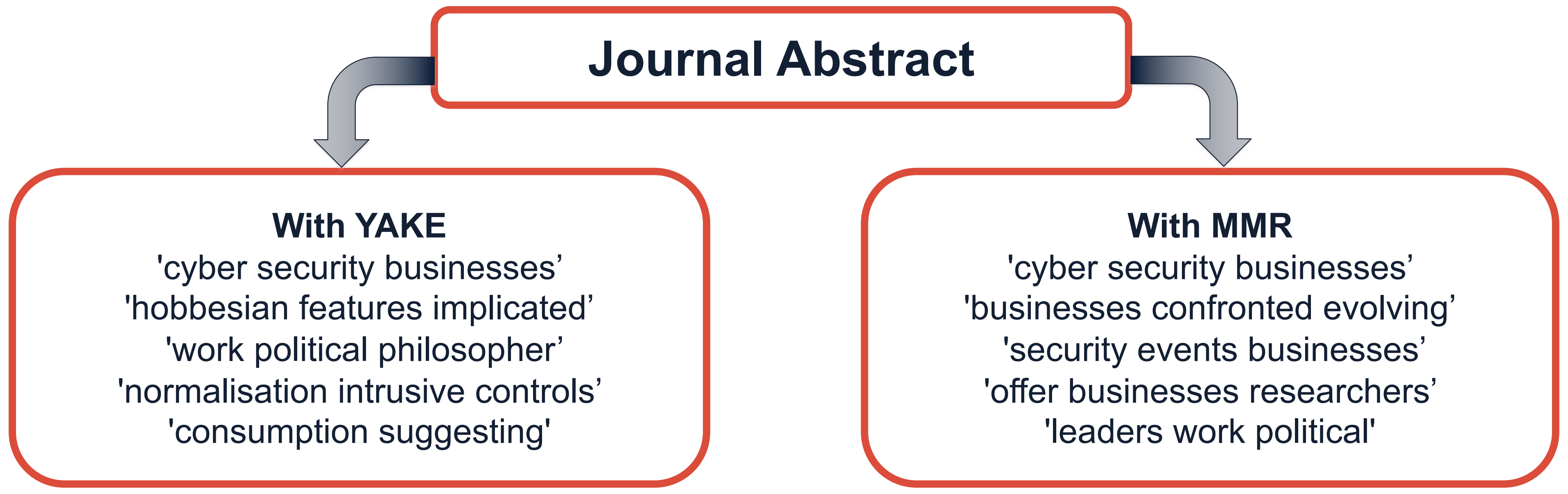


Text Data
- Sources of literature were identified and texts and data were scraped from there
- Custom scrapers were made to extract articles from various sources

Pre-processing
- The data was cleaned by removing stopwords and punctuation
- Exploratory data analysis (EDA) was performed on the text to generate candidate keywords

NLP Models
- Using BERT, YAKE, and other extraction methods, keywords were extracted from each article.
- LDA and other topic modelling methods were used to group these keywords

RESULTS



Journal Abstract

With YAKE
'cyber security businesses'
'hobbesian features implicated'
'work political philosopher'
'normalisation intrusive controls'
'consumption suggesting'

With MMR
'cyber security businesses'
'businesses confronted evolving'
'security events businesses'
'offer businesses researchers'
'leaders work political'

$$MMR = \text{argmax}_{D_i \in R \setminus S}\left[\lambda \text{Sim}_i(D_i, Q) - (1-\lambda)\max_{D_j \in S}\text{Sim}_2(D_i, D_j)\right]$$

D_i: Documents in Collection C
Q: Query
R: Relevant Documents in C
S: Current Result Set

Results: Outcome of the keywords is shown on two different models. The number of keywords generated can be changed with hyperparameter tuning but as the number of keywords increase, the relevancy and quality tends to decrease

DISCUSSION

- The keywords generated by YAKE are much more diverse than the ones generated by Max Marginal Relevance, however both capture the semantic meaning of the article based on the abstract
- The keywords are listed in descending relevancy with the first word being the most relevant. We see that the top keywords generated between the two models are very similar
- The diversity of the keywords generated by Max Marginal Relevancy can be adjusted with the diversity factor (λ)

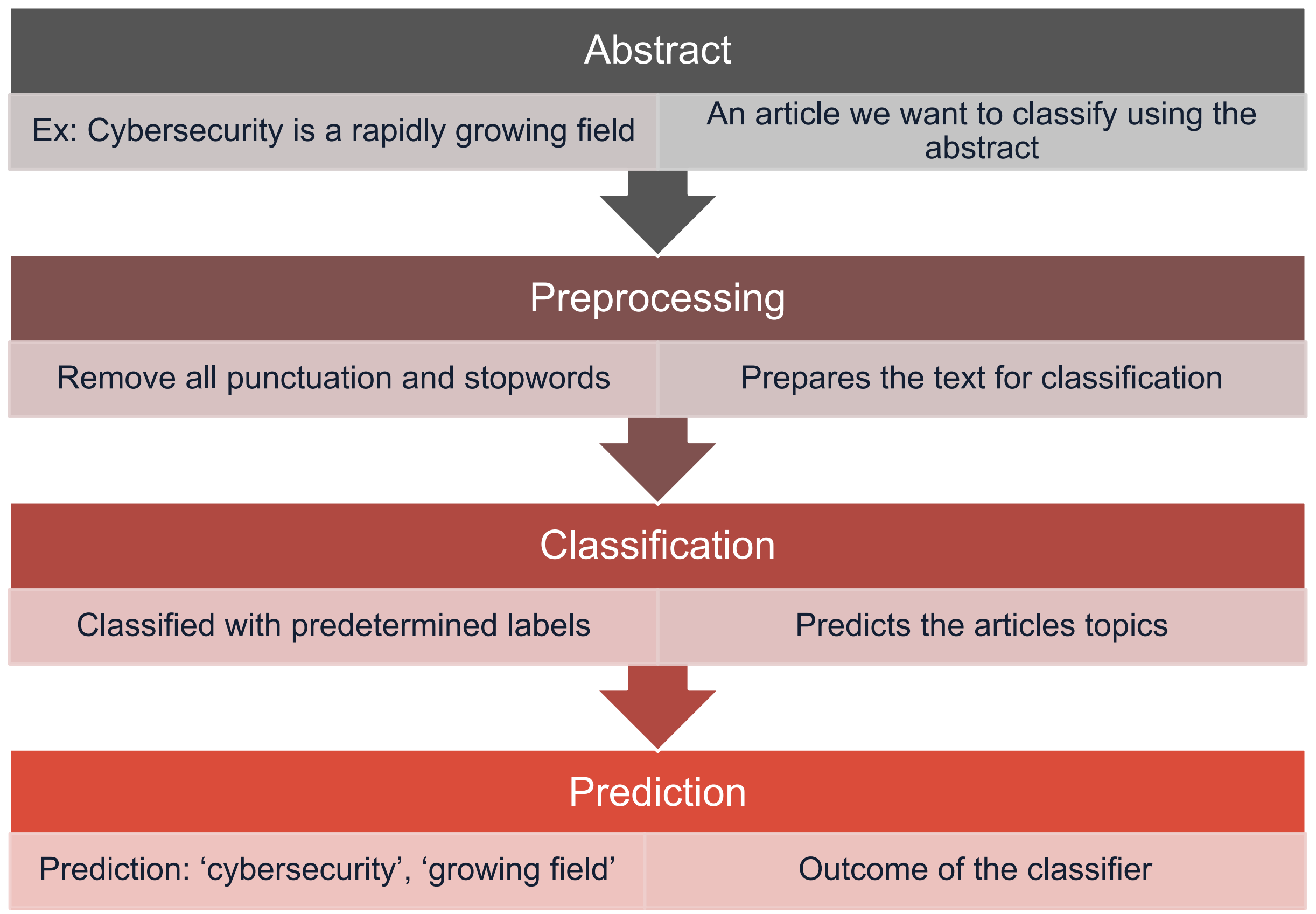
FUTURE DIRECTIONS

Currently, we have only implemented the keyword and topic modelling system. In the future, we propose a system that is able to automatically group these articles based on semantic content.

How the Article Pipeline Works:

1. The article summary/abstract is extracted and preprocessed based on a set of rules.
2. Text classification is done on the summary to group it into the most suitable category and the database is updated.

Proposal for Keyword Modeling and Classification System:



Abstract
Ex: Cybersecurity is a rapidly growing field | An article we want to classify using the abstract

Preprocessing
Remove all punctuation and stopwords | Prepares the text for classification

Classification
Classified with predetermined labels | Predicts the articles topics

Prediction
Prediction: ‘cybersecurity’, ‘growing field’ | Outcome of the classifier

REFERENCES

[1] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, “YAKE! Keyword extraction from single documents using multiple local features,” *Information Sciences*, vol. 509, pp. 257–289, Jan. 2020, doi:10.1016/j.ins.2019.09.013.

[2] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, “A Survey of the State of Explainable AI for Natural Language Processing,” Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.00711

[3] M. Eling, “Cyber risk research in business and actuarial science,” *European Actuarial Journal*, vol. 10, no. 2, pp. 303– 333, Dec. 2020, doi:10.1007/s13385-020-00250-1.

ACKNOWLEDGEMENTS

I would like to thank Dr. Zhiyu (Frank) Quan for welcoming me into his lab as well as his guidance and advice on the project. I would also like to thank Linfeng Zhang for his support throughout this project as well as well as giving me tips on how to refine my data and code. Finally, I thank the Illinois Risk Lab (iRisk Lab) and the Illinois Department of Actuarial Science for giving me this amazing opportunity to conduct research and gain experience which has enhanced my undergraduate studies.