BAMBOO PIPE

Datarama is a business intelligence company specializing in **regulatory compliance**. Combining advanced technology and expert human analysis, analysts at Datarama conduct due diligence on companies and provide clients with information on whether they conform to laws and standards in the business world.

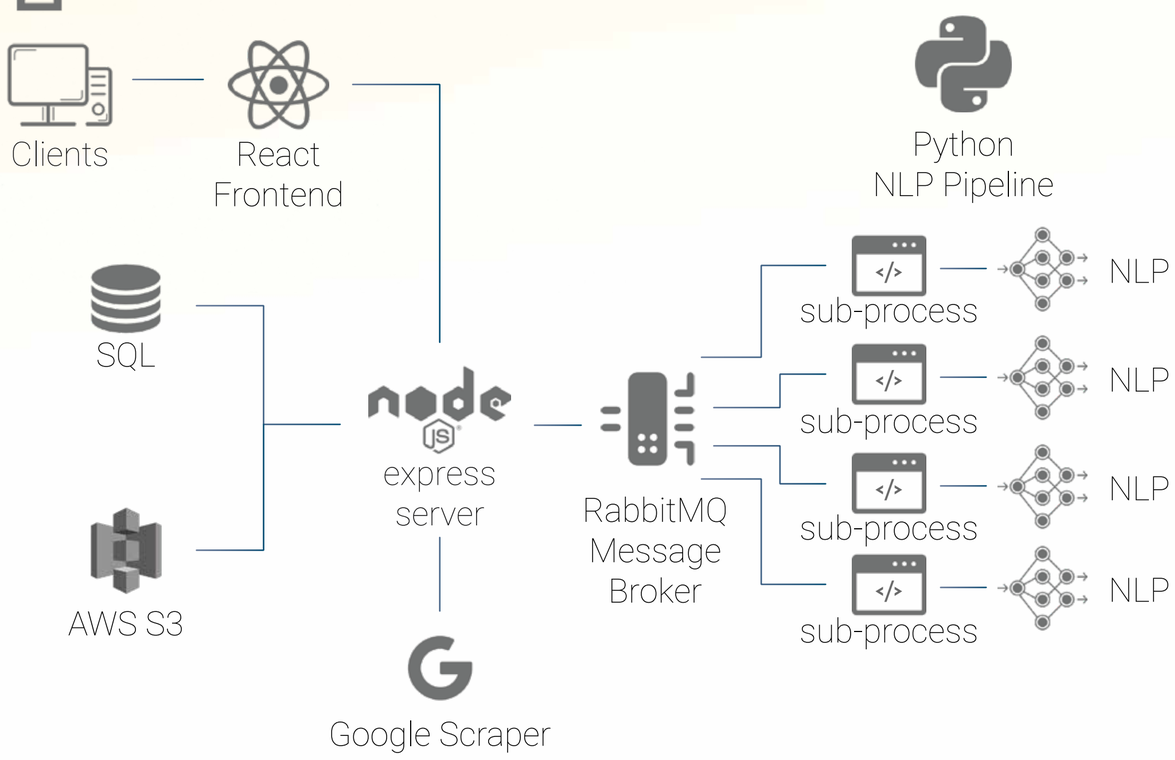
PROBLEM STATEMENT

The research work performed by the analysts in Datarama is time-consuming, repetitive, manual, and has an undesirable dependency on external paid services. How might we design a system that **streamlines the evaluation process for textual data** while **easing Datarama's information ingestion process** in order to **reduce the analysts routine work**?

OUR SOLUTION

An integrated NLP pipeline with GUI to provide a one-stop solution for **entity-sentiment analysis** and **topic identification**, tuned towards making compliance risk analysis more efficient.

SYSTEM ARCHITECTURE



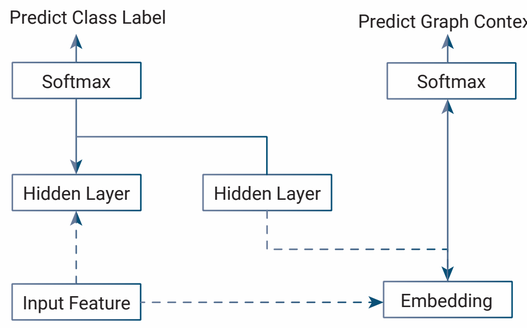
Our solution comprises a frontend, backend, database, and an NLP pipeline, linked up by the RabbitMQ message broker to form the entire functioning system. Parallel processing is applied to the NLP pipeline to improve computation efficiency.

NLP MODELS

TOPIC MODELLING (TM)

We used a graphical model, GraphSAGE (Hamilton et al., 2017), to identify topics in articles. It leverages node features to learn an embedding function that generalizes to unseen documents.

Each paragraph was treated as an individual document to predict multiple topics for each article. The top three most probable topics will be returned for each paragraph and displayed on the frontend web-app.



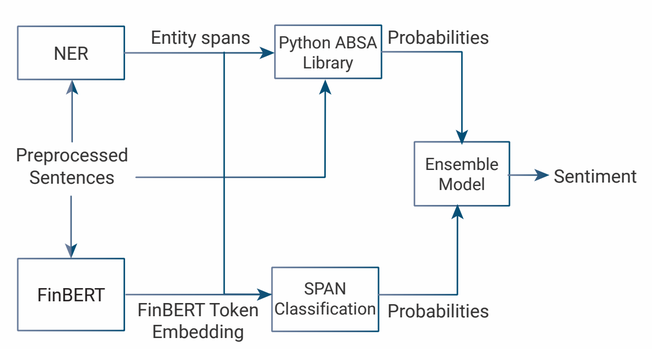
GraphSAGE Model

ENTITY-SENTIMENT ANALYSIS (ESA)

The ESA module comprises a Named Entity Recognition (NER) module and an Aspect-Based Sentiment Analysis module (ABSA).

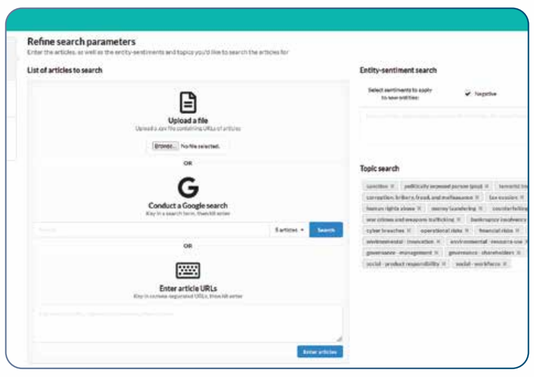
We used BERT (Bidirectional Encoder Representations) (Devlin et. al., 2019) vectors for the NER task.

An ensemble method of two models was used to obtain the aspect-based sentiment. FinBERT embeddings (Araci, 2019) were used to obtain sentence vectors for the ABSA task, as FinBERT is specially trained on financial news.

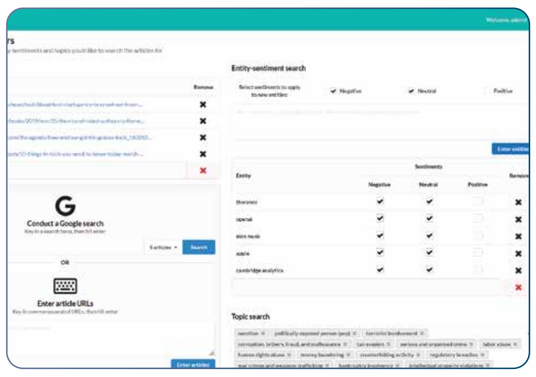


Ensemble Method for ABSA

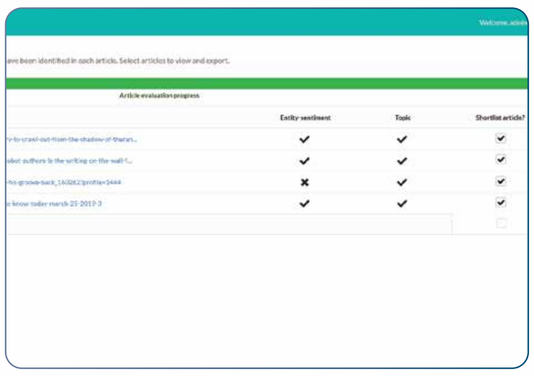
FRONTEND WEB-APP



1. Users will input web links to articles by uploading a .csv file, keying in a search term, or by manually keying in URLs.



2. Next, they will key in the entities and corresponding sentiments, as well as topics they are interested in.



3. After the articles are processed, users will shortlist articles to read. Articles that match at least one query will be automatically shortlisted.

4. Finally, users may read and download the articles and their corresponding analyses.

NLP MODEL PERFORMANCE

The metrics used for evaluating the TM, NER, and ABSA models are precision, recall, and F1-score.

The reason why the NER and ABSA models were better at detecting certain classes can be attributed to the uneven class distributions in the dataset. Hence, we chose to use F1-score instead of other metrics as it is more robust for this type of dataset.

NER (BERT)

	Precision	Recall	F1
PERSON	0.72	0.65	0.68
COM	0.53	0.51	0.52
GOV	0.33	0.29	0.31
FAM	0.22	0.28	0.24
OTHER	0.21	0.15	0.18

TOPIC MODELLING

	Precision	Recall	F1
Average	0.68	0.61	0.63
Best Topic	0.81	0.63	0.71
Worst Topic	0.67	0.47	0.55

ABSA (FINBERT)

	Precision	Recall	F1
Positive	0.34	0.55	0.42
Neutral	0.96	0.90	0.93
Negative	0.27	0.47	0.35

USER FEEDBACK

> The flow of the NLP process is quite clear. It will help us streamline the research!

> Having the flexibility to input articles from various sources will help to reduce the time spent during research!

TEAM

Li Yuxuan · Kuah Wee Ping · Jeslyn Peh · Hoon Wei Ting · Wong Lai Men · Fion Yao Yuechi · Ng Zhen Hao

References:
1. Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, US.
2. Dogu Araci. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
4. Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, & Yiwei Lv. (2019). Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification.

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN