Datarama.

BAMBOO PIPE

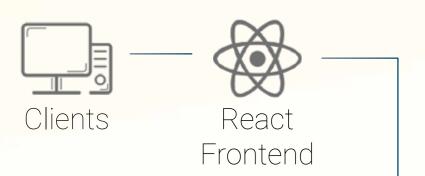
Datarama is a business intelligence company specializing in **regulatory compliance**. Combining advanced technology and expert human analysis, analysts at Datarama conduct due diligence on companies and provide clients with information on whether they conform to laws and standards in the business world.

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?



SYSTEM ARCHITECTURE





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.



TOPIC MODELLING (TM)

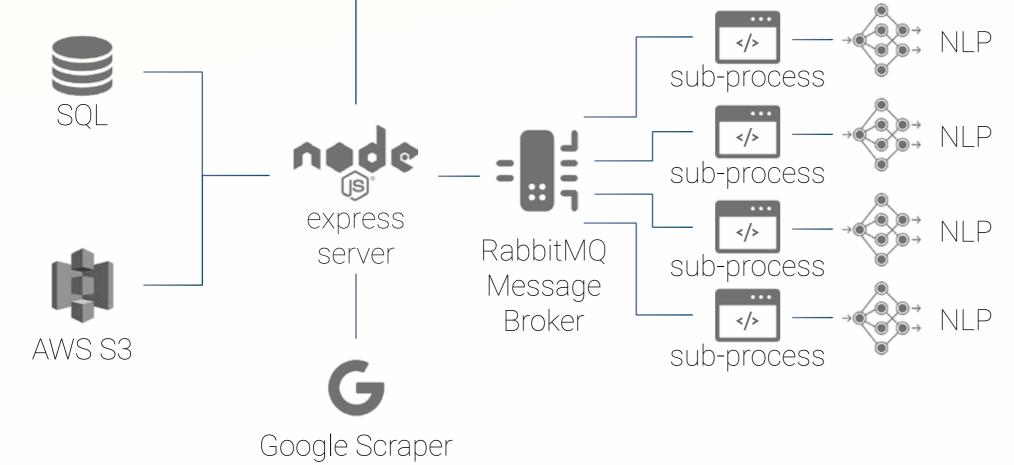
We used a graphical model, GraphSAGE (Hamilton et al., 2017), to identify topics in articles. It leverages node features to learn an embedding function that generalizes to unseen documents.

Each paragraph was treated as an individual document to predict multiple

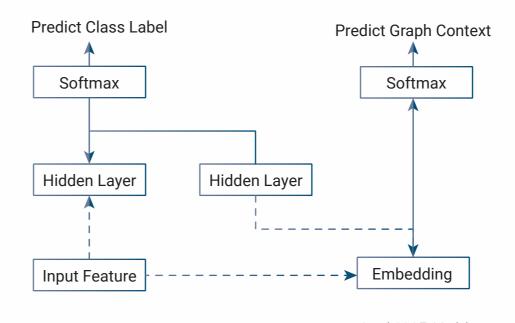
ENTITY-SENTIMENT ANALYSIS (ESA)

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

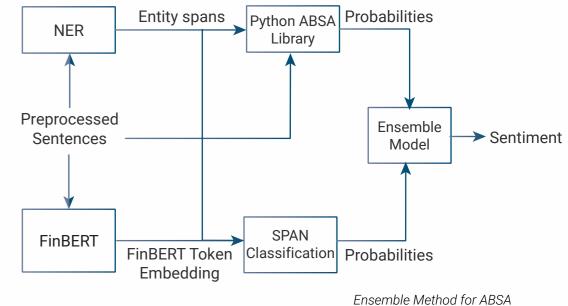
We used BERT (Bidirectional Encoder Representations) (Devlin et. al., 2019) vectors for the NER task.



topics for each article. The top three most probable topics will be returned for each paragraph and displayed on the frontend web-app.

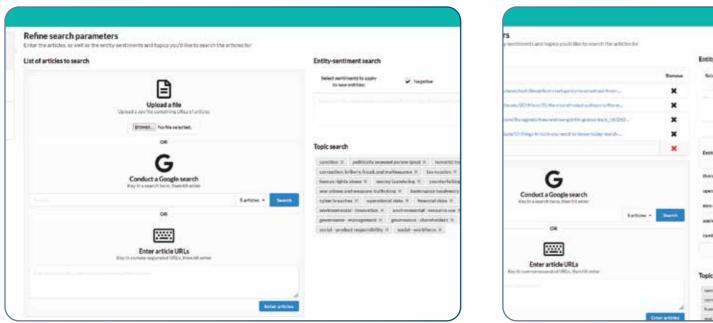


An ensemble method of two models was used to obtain the aspect-based sentiment. FinBERT embeddings (Araci, 2019) were used to obtain sentence vectors for the ABSA task, as FinBERT is specially trained on financial news.



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.

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.

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.

TOPIC MODELLING

	Precision	Recall	F1
--	-----------	--------	----

NER (BERT)

	Precision	Recall	F1
PERSON	0.72	0.65	0.68
СОМ	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

ABSA (FINBERT)

Precision	Recall	F1

GraphSAGE Model

Article evaluation program			
	Entity sentiment	Topis	Shortlist article
-to-crawl-out-tion-the-chadlow-of-theran	~	~	
bot suffices in the writing on the wall fun	~	~	
53-gross-back_160282/profile=3444	×	~	~
know today march 25-2019-3	~	~	

The rise of robot authors: is the writing on the wall for human novelists? Will androids write novels about electric sheep ? The dream, or nig Entity type PERSON med to has come one step closer with the recent announcement of an artificial interinge stories or fiction. It was the brainchild of Openal - a nonprofit lab backed by new and other tech entrepreneur which styly alarmed the literati by announcing that the AI (called GPT2) was too dangerous for them to release into the wild, because it could be employed to create " deepfakes for text "." Due to our concerns about malicious applications of the technology, "they said, " we are not releasing the trained model." Are machine - learning entities going to be the new we of information terrorism, or will they just put humble midlist novelists out of business ? Let's first take a step back. All has been the next big thing for so long that it's easy to assume " artificial intelligence " now exists . It does n't, if by " intelligence " we mean what we sometimes encounter in our fellow humans . GPT2 is just using methods of

statistical analysis, trained on huge amounts of human - written text = 40 GB of web pages, in this case, that received recommendations from <u>Reddit</u> readers - to predict what ought to come next. This probabilistic approach is how <u>Goople</u> <u>Translate</u> works , and also the method behind <u>Gmail's</u> automatic replies (* OK. ** See you then. ** That 's fine !*) It can be eerily good, but it is not as intelligent as . say, a bee

Right now, novelists do n't seem to have much to fear . Fed the opening line of George Onvell 's Nineteen Eighty - Four - " It was a bright cold day in April , and the clocks were striking thirteen * - the machine continued the narrative as follows :* I was in my car on my way to a new job in Seattle . I put the gas in , put the key in , and then I let it run . I just imagined what the day would be hundred years from now. In 2045, I was a teacher in some school in a poor part of rural China. I sta

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 their and corresponding analyses.

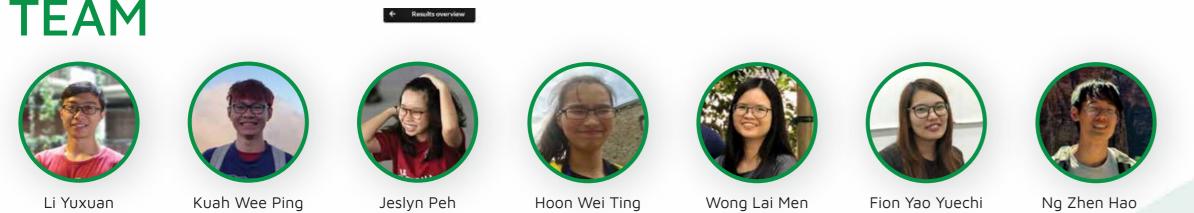
	Average	0.68	0.61	0.63	
	Best Topic	0.81	0.63	0.71	
	Worst Topic	0.67	0.47	0.55	

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!



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**