

D-CLASSIFIED NEWS

Sponsored by: Noonum
Team Members: Frank Chen, Apoorva Shetty, V J Kamala, Tharun Sikhinam



PROBLEM STATEMENT:

- Noonum is a fin-tech AI startup that leverages graphs and NLP to be a knowledge engine for business and finance
- Their current news dashboard is seen in fig 1, which contains both relevant and irrelevant news articles
- The aim of the project is to classify news articles as relevant or irrelevant based on their "market-relevance" and to explain why this classification was made.

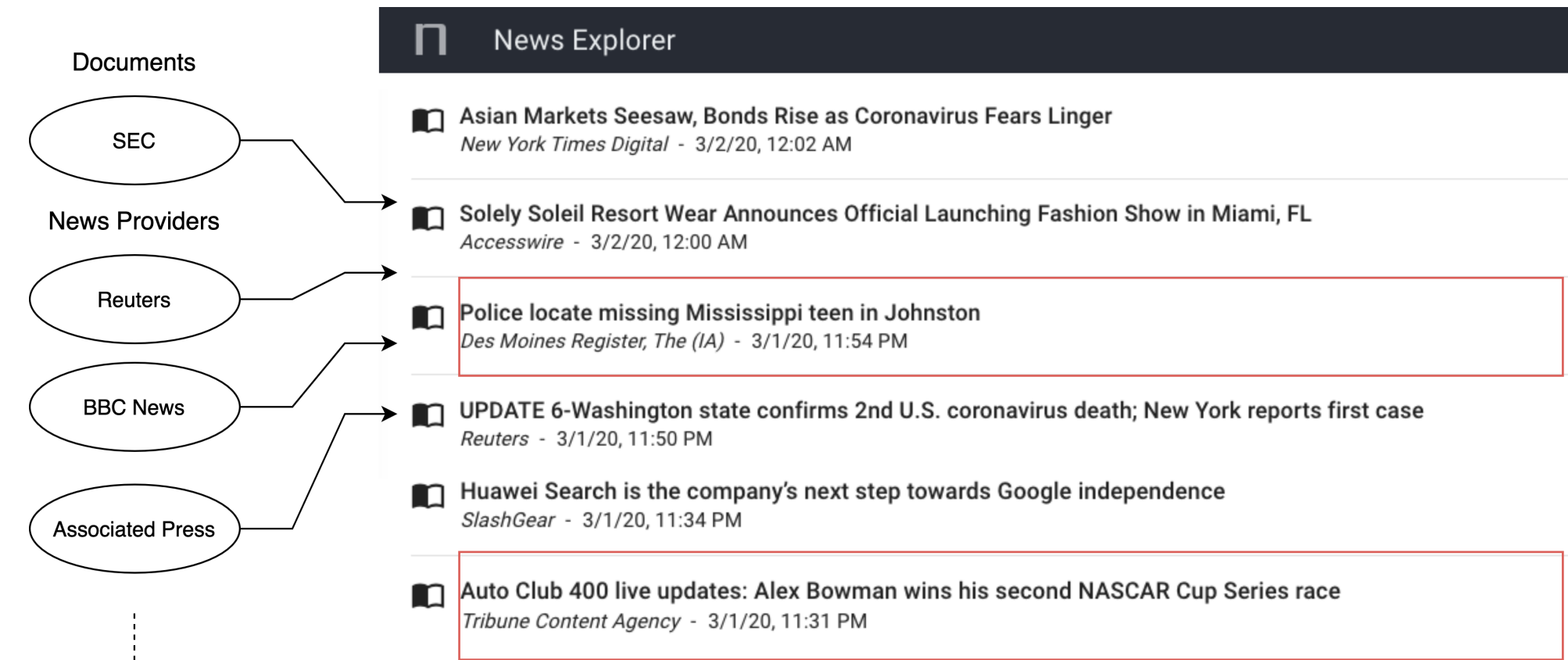
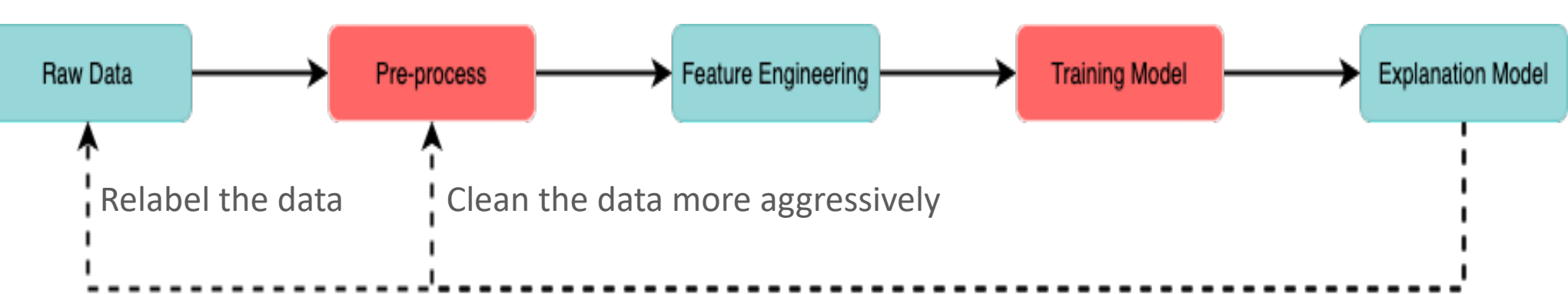


Fig. 1 Noonum dashboard

PROPOSED SOLUTION:



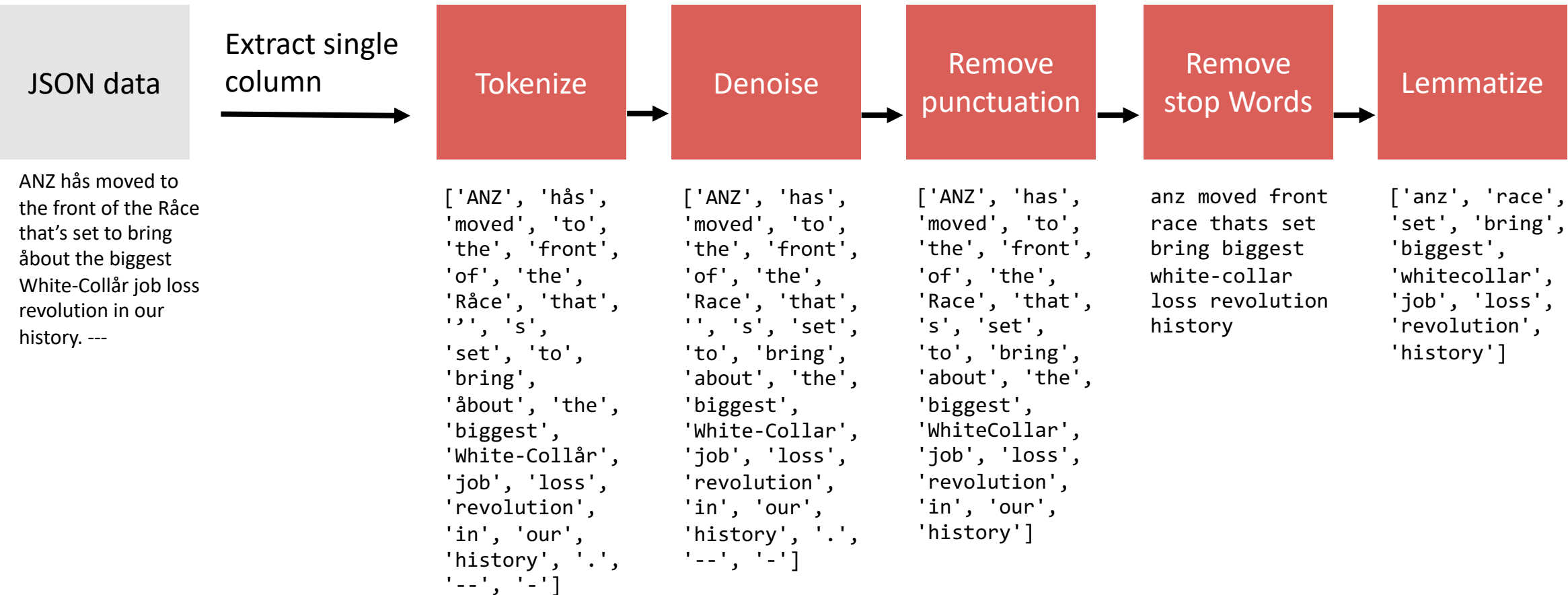
DATA

Irrelevant News Articles (10K records)
Relevant News Articles (10K records)
uid <alpha-numerical id>
source <text>
headline <text>
summary <text>
content <text>



- Dataset consists of 10k relevant and irrelevant articles from various news sources
- The dataset contains source, headline, summary and content of a news article
- The news articles were classified by looking for the presenece of a publicly traded company in the articles

DATA PRE-PROCESSING



FEATURE ENGINEERING

Bag of Words (BOW)	<table><tr><th>banks</th><th>company</th><th>revolution</th><th>market</th><th>fallout</th><th>...</th></tr><tr><td>1</td><td>1</td><td>1</td><td>1</td><td>0</td><td>...</td></tr></table>	banks	company	revolution	market	fallout	...	1	1	1	1	0	...	• Counts of each word in the document
banks	company	revolution	market	fallout	...									
1	1	1	1	0	...									
TF-IDF	<table><tr><th>banks</th><th>company</th><th>revolution</th><th>market</th><th>fallout</th><th>...</th></tr><tr><td>0.03538</td><td>0.06532</td><td>0.00819</td><td>0.0213</td><td>0.0065</td><td>...</td></tr></table>	banks	company	revolution	market	fallout	...	0.03538	0.06532	0.00819	0.0213	0.0065	...	• Frequency of a word in current document times how rare the word is in the corpus
banks	company	revolution	market	fallout	...									
0.03538	0.06532	0.00819	0.0213	0.0065	...									
Named Entity Resolution	<table><tr><th>PERSON</th><th>ORG</th><th>GPE</th><th>LOC</th><th>PRODUCT</th><th>TIME</th></tr><tr><td>1</td><td>3</td><td>1</td><td>1</td><td>0</td><td>0</td></tr></table>	PERSON	ORG	GPE	LOC	PRODUCT	TIME	1	3	1	1	0	0	• Counts of named entity types such as PERSON, LOCATION
PERSON	ORG	GPE	LOC	PRODUCT	TIME									
1	3	1	1	0	0									
Topic Modeling	<table><tr><th>POLITICS</th><th>SPORTS</th><th>FINANCE</th><th>HEALTH</th><th>TECH</th><th>CRIME</th></tr><tr><td>0.432018</td><td>0.00000</td><td>0.56055</td><td>0.00000</td><td>0.000000</td><td>0.00000</td></tr></table>	POLITICS	SPORTS	FINANCE	HEALTH	TECH	CRIME	0.432018	0.00000	0.56055	0.00000	0.000000	0.00000	• Topic Modeling scores after LDA
POLITICS	SPORTS	FINANCE	HEALTH	TECH	CRIME									
0.432018	0.00000	0.56055	0.00000	0.000000	0.00000									
Embedding	<table><tr><th>weight_0</th><th>weight_1</th><th>weight_2</th><th>weight_3</th><th>.....</th><th>weight_k</th></tr><tr><td>-0.466</td><td>0.256</td><td>-0.2795</td><td>0.08231</td><td>.....</td><td>-0.28593</td></tr></table>	weight_0	weight_1	weight_2	weight_3	weight_k	-0.466	0.256	-0.2795	0.08231	-0.28593	• Doc2Vec, BERT embedding map text to k-dim vector
weight_0	weight_1	weight_2	weight_3	weight_k									
-0.466	0.256	-0.2795	0.08231	-0.28593									

TRAINING AND EXPLANATION MODEL

- For the training Model we chose XGBoost, Naïve Bayes and Logistic regression since these classifiers are proven to perform well on text classification from literature survey. To map with these, we had to choose the best model and feature combination, which was found to be:
Feature: **Bag of Words**
Model: **XGBoost**
Classification Accuracy: **74%**
Precision/Recall: **74%/73%**
- For our explanation model we decided to use the local explanation model LIME, since it gives us a clear idea of which specific words are weighted as irrelevant or relevant

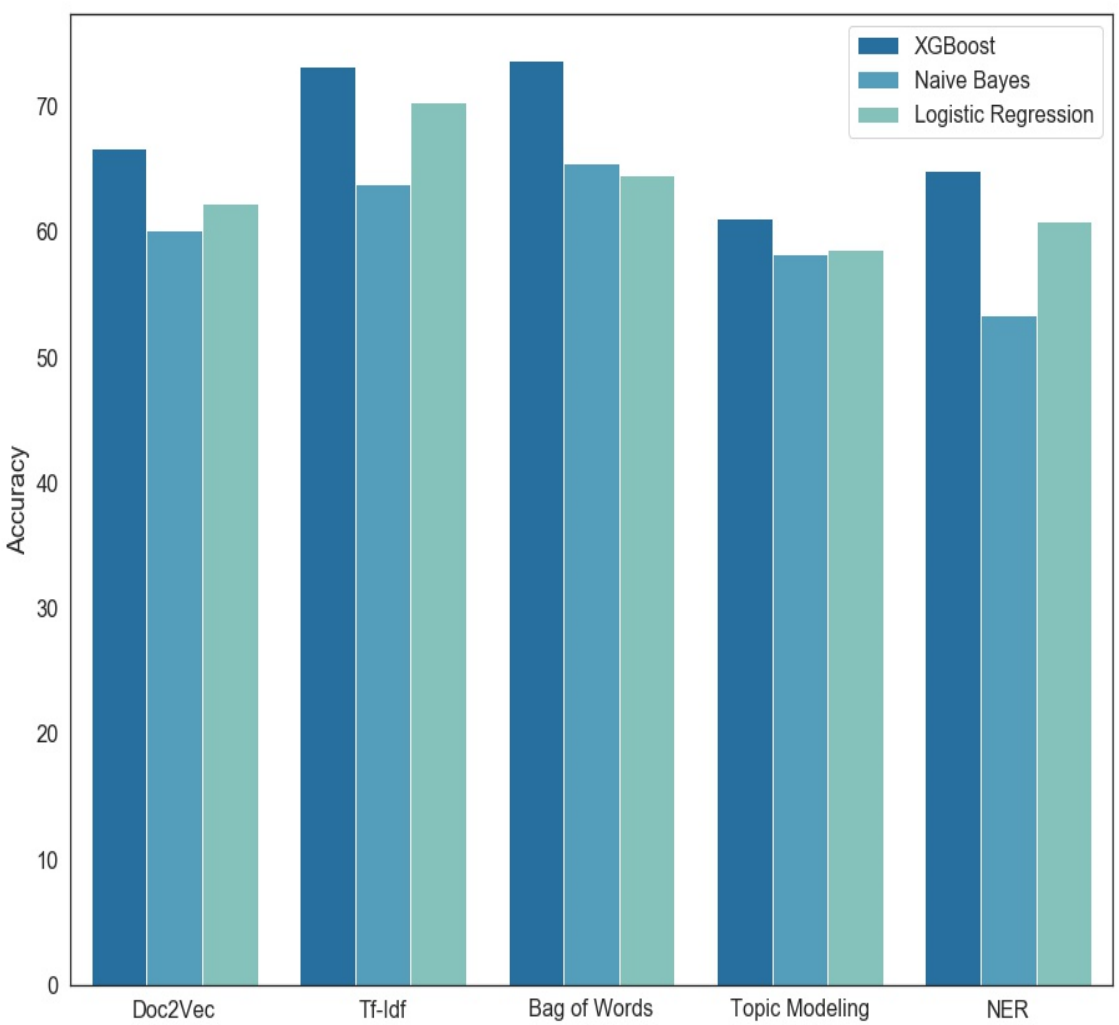
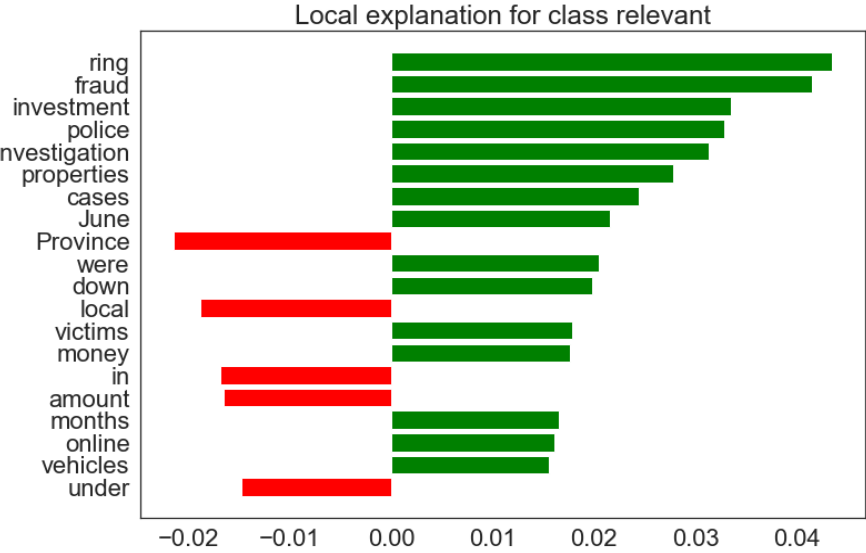


Fig. 2 Model and Feature performance

RESULTS

The following results were generated using LIME on our trained model for a random test point

Text with highlighted words
CHENGDU, July 30 (Xinhua) -- Police in southwest China's Sichuan Province busted an online fraud ring and detained 128 suspects involved in over 5,000 cases.
The amount of money involved in the fraud cases totaled 100 million yuan (14.5 million U.S. dollars).
In March 2019, a victim reported to the local police in the city of Pengzhou that he had been cheated out about 870,000 yuan by online investment fraud in five months.
By tampering with the data of the online platforms they built, the suspects have tricked victims to double down on their amount of investment and appropriated their money with the excuse of failed investment, according to the local police.
Through three months' investigation, local police gradually confirmed the criminal gang's location, staff structure, fraud process and fund flow information.
In June, 45 core members of the ring were arrested, and a total of 128 suspects' identities were later confirmed, with nine vehicles, over 200 computers, two properties and bank accounts containing over 5 million yuan were frozen.
The case is under further investigation, Enditem



INSIGHTS

- Running LIME on 100 random test points, we were able to get the frequency of most weighted words in both classes
- Using these insights we were able to clean the data further to get a higher accuracy.

Fig. 3 Iteration 1: Accuracy ~60%

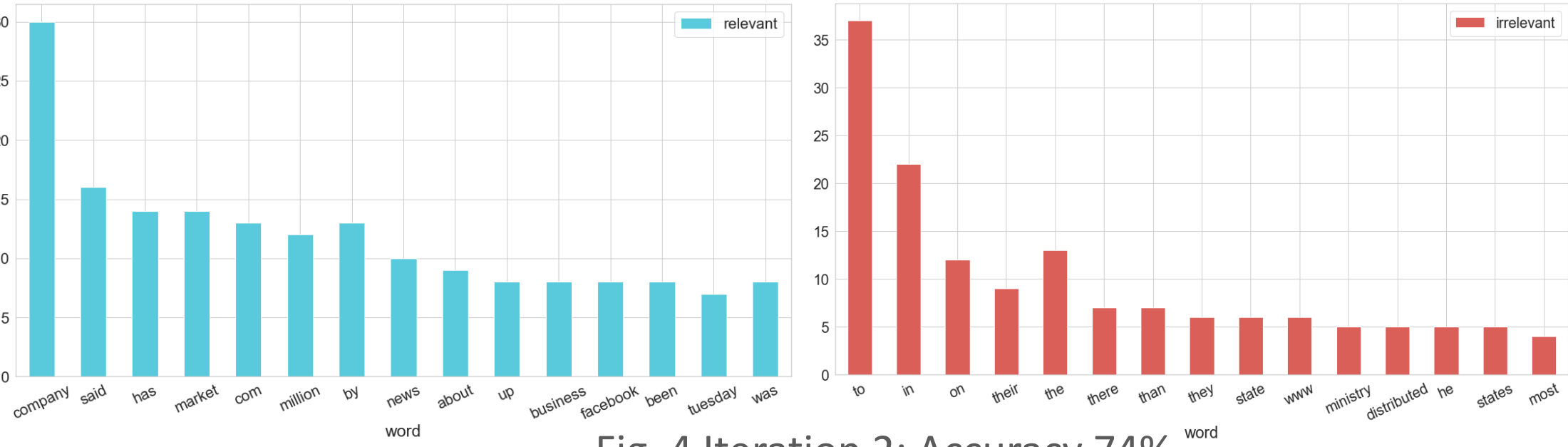
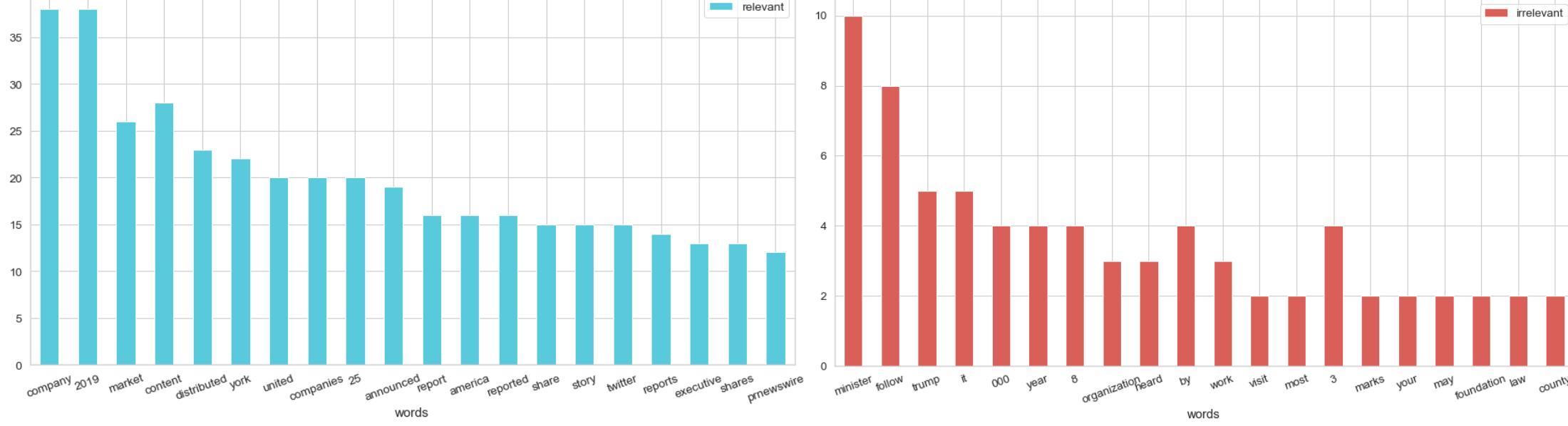


Fig. 4 Iteration 2: Accuracy 74%



CONCLUSIONS & FUTURE WORK

- Simpler feature sets such as BoW and TF-IDF performed well. Aggressive cleaning & pre-processing w.r.t to the context of the application improved the accuracy of the model.
- LIME gave us a clearer picture of what our model was truly learning; indicating that the market relevant terms were being captured by the models.
- The task lying ahead is to provide for a feedback loop to re-incorporate what we learn from our explanation into labelling the data, and also into our data pre-processing and feature-engineering steps.