Data directory -
https://github.com/guneetsinghmehta/financial-modeling/tree/master/Data/Intrino_news/nivetha/output_files

Training Data:-
https://github.com/guneetsinghmehta/financial-modeling/tree/master/Data/Intrino_news/nivetha/output_files/train

Test Data:
https://github.com/guneetsinghmehta/financial-modeling/tree/master/Data/Intrino_news/nivetha/output_files/test

Directory of all code
https://github.com/guneetsinghmehta/financial-modeling/tree/master/Data/Intrino_news/nivetha

Zipped folder

Team members:
Neha Mittal
Nivetha Singara Vadivelu
Guneet Singh Mehta

We have extracted name of the companies e.g. Apple, Microsoft, Facebook etc

Total number of mentions marked up: 3882

Number of documents in set I : 200
Number of mentions in set I : 2558

Number of documents in set J : 100
Number of mentions in set J : 1324

Classifier selected after first cross-validation: Logistic Regression
Precision: 0.7877
Recall: 0.5202
F1: 0.6266

Classifier selected before rule-based post-processing step: Logistic Regression
On set J
Precision: 0.8716
Recall: 0.5405
F1: 0.6672

We have blacklisted some words such as 'the', 'editors', 'capital', 'trump', 'ajo,', 'lp', 'alzheimers', 'university'

- Rules present in “should_filter” function in “get_tuple_format_for_negative_data.py” file

Classifier selected after rule-based post-processing step: Logistic Regression
On set J
Precision: 0.9254
Recall: 0.5553
F1: 0.6941

All the text files have company names marked up.

The company names are enclosed in the tag <strong>. For example:-
<strong>3M</strong>
<strong>Amazon</strong>

Since many of our features depend on the position of the markup in file, we have the company name marked up in positions in text where it occurs