Hackathon: Sentiment analysis and stock performance

Bhupinder Singh
Pratap Dangeti
Dhanvanthri M

Problem Statement

- How can we assess the financial stress in companies using company's financial results and stock market data?

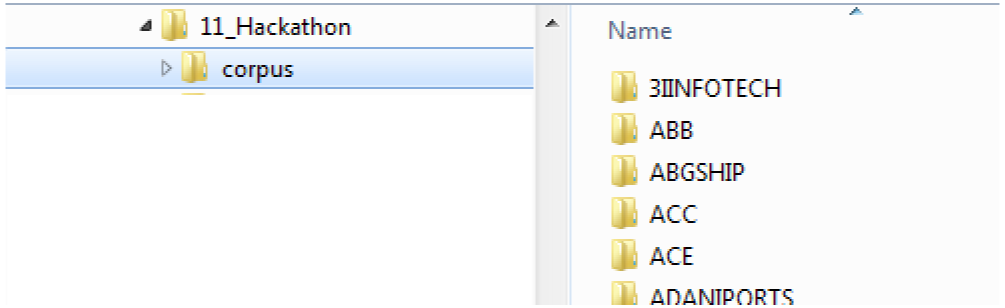
- Can we identify forward looking sentences from the Management Discussion and Analysis section of the annual reports ?

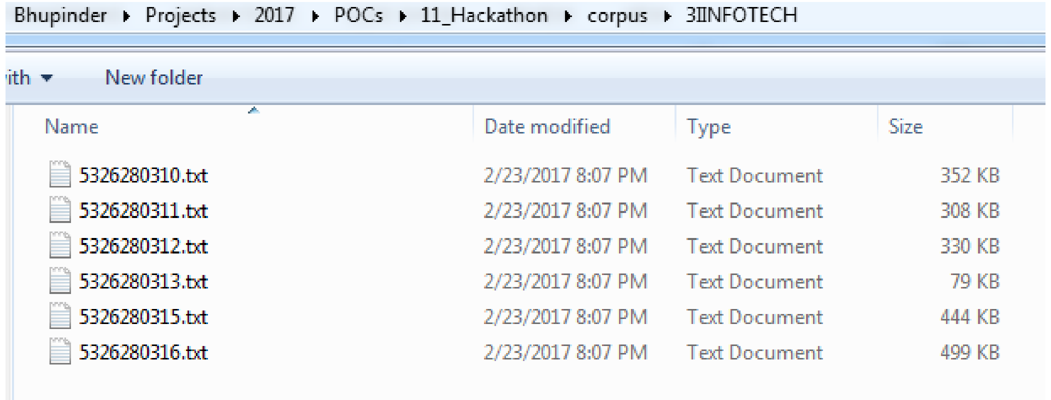
- How are these parameters in the prediction models different across different industries/sectors for India market?

Data Sources

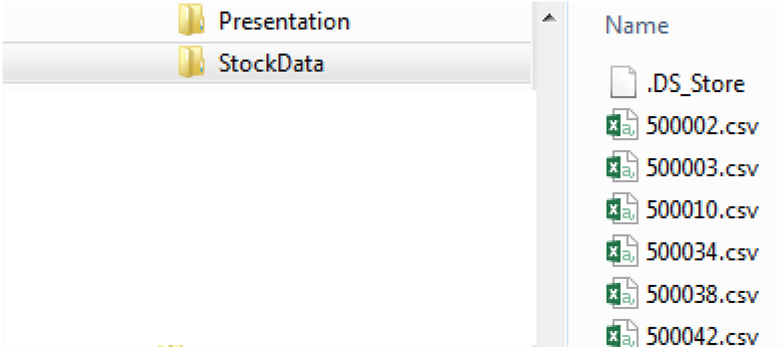
- Annual Reports for companies – **186 companies (Mar 2010 – Mar 2016)**
- Corresponding stock market data for the companies – **147 files (Daily Stock Data Jan 2010-Dec 2015)**

Annual Reports





Stock Data



Date	Open Pric	High Price	Low Price	Close Pric	WAP	No.of Sha	No. of Tra	Total Turn	Deliverab	% Deli. Qt	Spread Hi	Spread Clo
31-Dec-15	1118.85	1123	1111.05	1118.1	1118.216	3541	435	3959604	1903	53.74	11.95	-0.75
30-Dec-15	1120.75	1120.75	1108.55	1110.75	1114.509	2448	301	2728319	1088	44.44	12.2	-10
29-Dec-15	1125	1125	1112.7	1117.1	1118.364	2698	347	3017345	741	27.46	12.3	-7.9
28-Dec-15	1123.3	1132	1116	1119.35	1121.738	3127	355	3507676	1399	44.74	16	-3.95
24-Dec-15	1135.9	1139.3	1107.5	1116.8	1122.349	5261	675	5904676	2769	52.63	31.8	-19.1
23-Dec-15	1140	1147	1130	1130.85	1137.37	4895	611	5567427	2279	46.56	17	-9.15

AnnualReport - Sample

5326280310.txt

Script ID

Annual Report Release Date: Mar 2010

StockData - Sample

520002.csv

Script ID

TATA CONSULTANCY SERVICES

High-Level Solution Approach

Unstructured Annual Reports

PDF

Apache Tika
Pattern
Sentiment Analyzer

Industry performance

Economic Outlook

Balance Sheet Ratios

Attribute-wise Segmentation & Sentiment Analysis

Natural Language Analyses with NLTK

+

Stock Performance Reports

Aggregations on stock performance for 1 month, 6 month, 12 month durations

Model Training & Validation

Machine Learning & Statistical Modeling

Stock Price Predictions

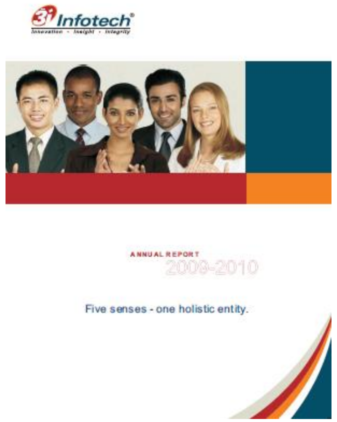
Selection of Tools

- Python was our preferred choice for natural language processing (NLP), since R does not always scale well for NLP processing
- Extraction of content from PDF was done using Apache Tika
- We used Python NLTK for NLP in python (sentence tokenization)
- We used CLIPS Pattern sentiment analyzer (lexicon based) – Gives polarity & subjectivity scores
  - Other sentiment analyzers: Affin, Sentiwordnet, Sentistrength, vadersentiment, sentiment package in R, etc.
  - Ref: SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods, https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0085-1
- We also tried extracting tables separately using pyTables (However, the results were poor)
- We explored scikit-learn in python and regression & classification modeling in R to fit our classifiers on the dataset

Natural Language Analyses with NLTK

Solution Approach: Detailed View

Annual Reports (PDF)



AnnualReport - Sample

Annual Reports (TXT)

3i_AR_Final AW 18-06-10_SAP_Print CTP.cdr
Five senses - one holistic entity.

A N N U A L R E P O R T

2009-2010

2009-2010

AA N N U A L R E P O R T

2009-2010

Company Details

Board of Directors:
Principal Bankers:

AnnualReport - Sample

Filtering Texts by Category

Developed Ontology

Ontology

Sentence Tokenization

SentenceTokenization - Sample

Filtering sentences
Category = Margin

Filtered Sentences - Sample

Sentiment Analysis

Used Pattern for sentiment analysis

Pol, Subj = sentiment(sentence)

Create Sentiment Classes from polarities:
Pos: +0.3 to 1
Neu: -0.3 to +0.3
Neg: -1 to -0.3

Create score for each category

Scores = (pos_count+1)/(neg_count+1)

Filtered Sentences - Sample

Structured Data

Calculate Perc Stock Price change for 1 month, 6 month, 12 month durations

Merge with Stock Data

Preprocessed data v1

Preprocessing

Sentiment scores range from 0 – 35

Decide cut points for senti scores
0, 5, 10, 15, 20, 25

Target variables range
Range: 4.08 to 7.7

Decide cut points for target
0, 4.9, 5.1, 5.3, 8

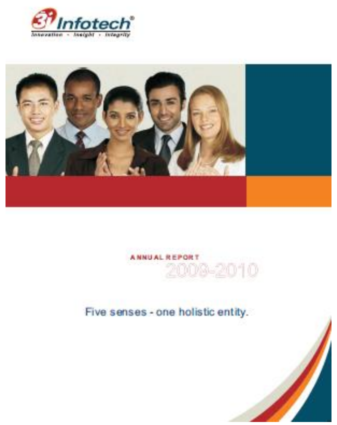
Preprocessed Data v2

Model Train & Validation

```
library("nnet")
m2 <-
multinom(binned_y_target12  ~ .,data =
train_data)
summary(m2)
library(ROCR)
pred =
predict(m2,test_data)
table(pred,
test_data$binned_y_target12)
```

Solution Approach: Analysis of Annual Reports

Annual Reports (PDF) → Annual Reports (TXT) → Sentence Tokenization → Filtering Sentences by Category → Sentiment Analysis → Get Sentiment Classes → Create Category Scores → Preprocessing



3i_AR_Final AW 18-06-10_SAP_Print CTP.cdr
Five senses - one holistic entity.

A N N U A L R E P O R T

2009-2010

2009-2010

AA N N U A L R E P O R T

2009-2010

Company Details

Board of Directors:
Principal Bankers:

	sentence	sentiment_classes	sentiment_polarity
0	During the last year notable contributions have been made, not only in terms of revenue generation, but also by conserving costs, thereby enabling us to maintain our margins as we increased our volume of business.	neutral	0.166666667
1	Operating profit is at Rs 503.14, a growth of 11% over the previous year and operating margins improved to 20.4%.	neutral	-0.166666667
2	The North America geography continued to be the largest contributor to our revenue and profits, with a 55% share of our global revenue, followed by South Asia geography at 26%.	neutral	0

Thresholds:
Neg: -1 to -0.3
Neu: -0.3 to 0.3
Pos: +0.3 to 1

neutral	60
positive	2

idx	senti_scoi	senti_score_econc	senti_score_exper	senti_score_gc
0	2	1.333333333	0.333333333	1.8
1	1.666667	2.5	0.333333333	4.5
2	2	2.5	1	3.75

Create score for each category

Scores = (pos_count+1)/(neg_count+1)

Score = 3/1

AnnualReport - Sample

AnnualReport - Sample

SentenceTokenization - Sample

All Sentences

Filtered Sentences - Sample

Filtered Sentences

Filtered Sentences - Sample

Filtered Sentences - Sample

Preprocessed data v1

Preprocessed Data v2

Solution Approach: Analyzing Stock Data

Stock Data → Calculate Change in Stock Price → Calculate for 1, 6 and 12 months → Transforming Stock Price Change → Binning

Date, Closing Price

Date	Open Pric	High Price	Low Price	Close Pric
31-Dec-15	1118.85	1123	1111.05	1118.1
30-Dec-15	1120.75	1120.75	1108.55	1110.75
29-Dec-15	1125	1125	1112.7	1117.1
28-Dec-15	1123.3	1132	1116	1119.35
24-Dec-15	1135.9	1139.3	1107.5	1116.8
23-Dec-15	1140	1147	1130	1130.85

Change in Closing Price

id	yr	Base_Pric	one_mnth	six_mnth	one_year	one_mnth	six_mnth	one_year
500002	2015	1395.2	1393.65	1226.9	1118.1	-0.1	-12.1	-19.9
500002	2014	716.25	814.05	982.65	1395.2	13.7	37.2	94.8
500002	2013	570.85	474.9	449.9	716.25	-16.8	-21.2	25.5
500002	2012	822.35	829.45	715.35	570.85	0.9	-13	-30.6

Create score for each category

Y target
= ln(stock price change + 150)

ID: 500002
Y_target1 = 5.009
Y_target6 = 4.926
Y_target12 = 4.868

Target variables range
Range: 4.08 to 7.7

Decide cut points
0, 4.9, 5.1, 5.3, 8

binned_y_	binned_y_	binned_y_target12
(0.0741,5.:	(0.0741,5.:	(0.0741,5.28]
(0.0741,5.:	(0.0741,5.:	(0.0741,5.28]
(0.0741,5.:	(0.0741,5.:	(0.0741,5.28]
(0.0741,5.:	(0.0741,5.:	(0.0741,5.28]

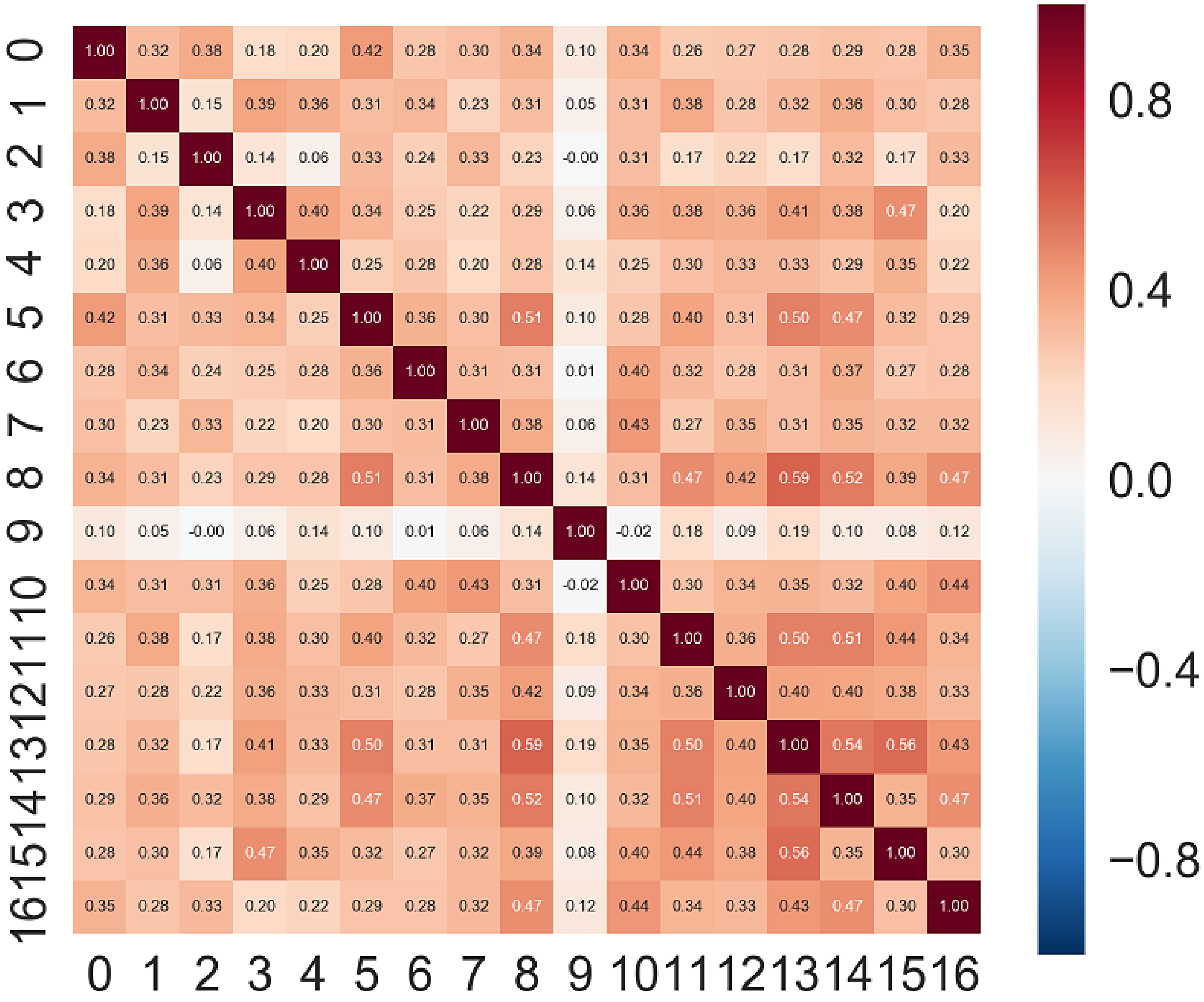
StockData - Sample

StockSummary - Sample

Preprocessed data v1

Preprocessed Data v2

Exploratory Data Analysis

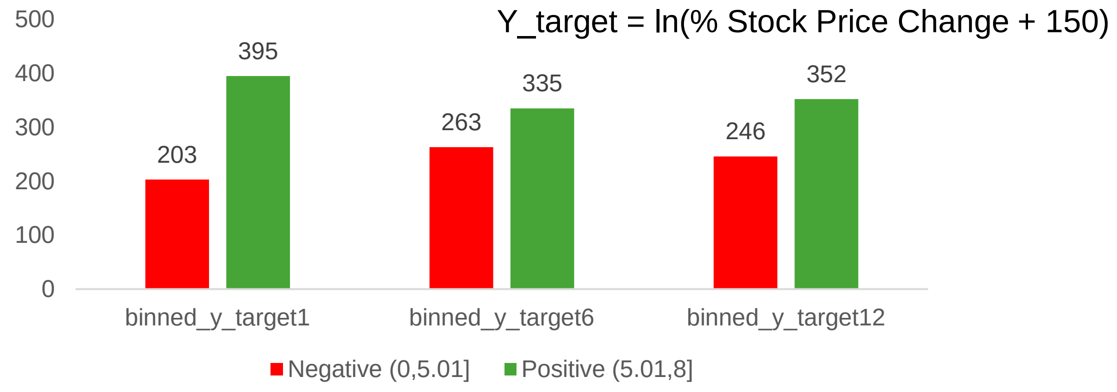


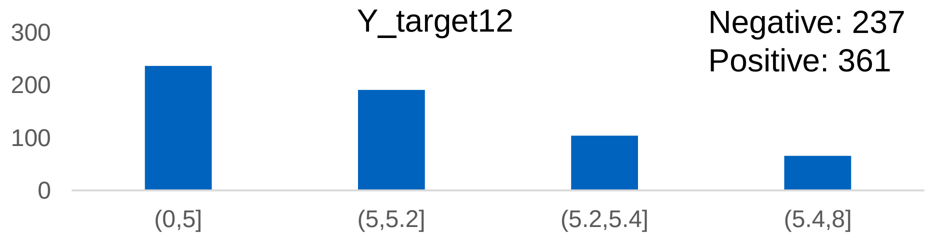
- Mergers & acquisitions & recent trends
  - Variables 13, 8 - Corr Coeff 0.59
- Recent trends & Revenue
  - Variables 13, 15 – Corr Coeff 0.56

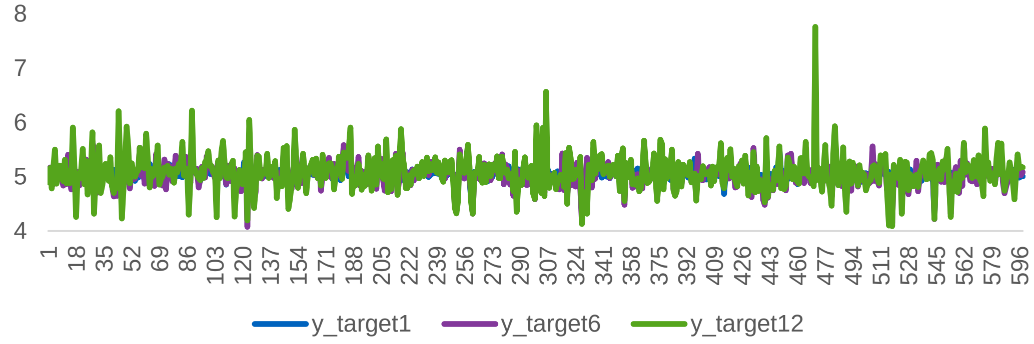
0	senti_score_assets
1	senti_score_economy
2	senti_score_expenses
3	senti_score_govt.policy
4	senti_score_industry
5	senti_score_innovation
6	senti_score_liquidity
7	senti_score_margins
8	senti_score_merges_acquisitions
9	indicators_ratio
10	senti_score_pricing
11	senti_score_product.portfolio
12	senti_score_ratio
13	senti_score_recent.trends
14	senti_score_recognition_awards
15	senti_score_revenue
16	senti_score_stocks

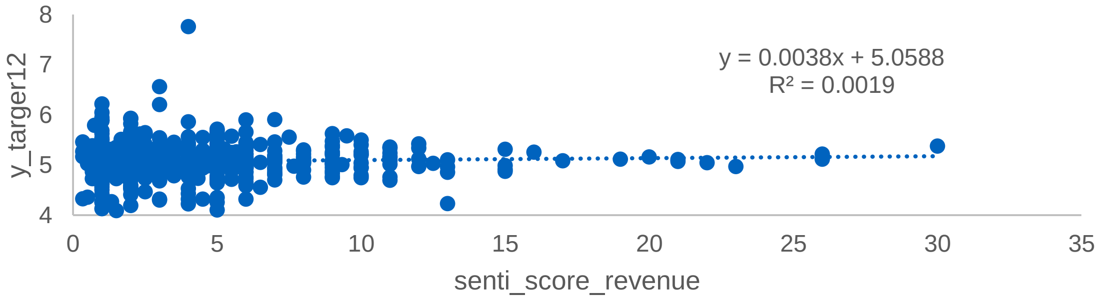
Exploratory Data Analysis

Target variables corresponding to Stock Price change
for 1 month, 6 month, 12 month durations









Model Data:
598 rows for 113 unique companies

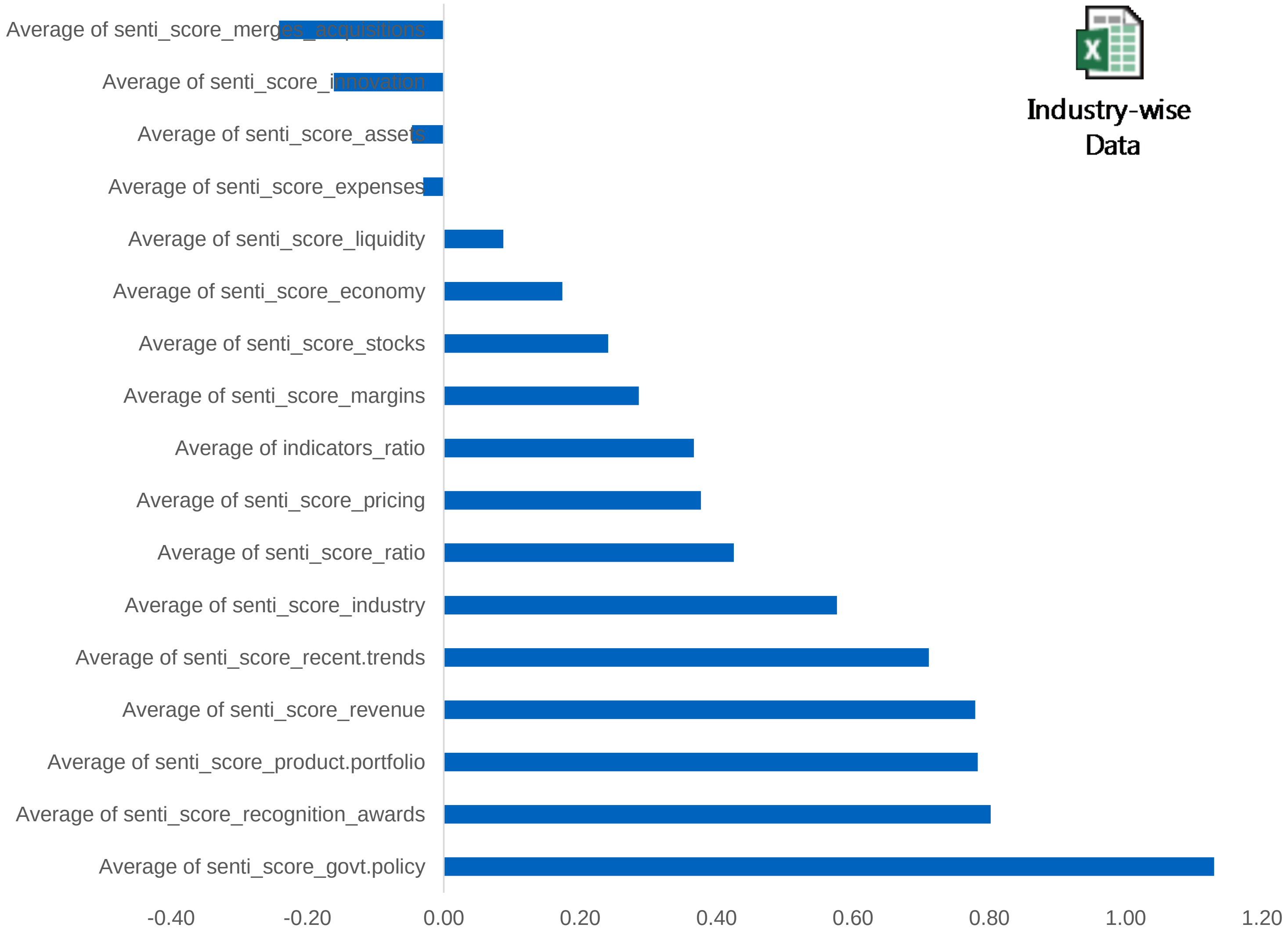
Y_target12

Binned_senti_score_revenue

		Positive			
	Neg	Low	Medium	High	
Row Labels	(0,5]	(5,5.2]	(5.2,5.4]	(5.4,8]	Grand Total
(0,5]	190	136	75	52	453
(10,15]	9	7	5	1	22
(15,20]		3	1		4
(5,10]	37	41	21	13	112
NA	1	4	2		7
Grand Total	237	191	104	66	598

Exploratory Data Analysis

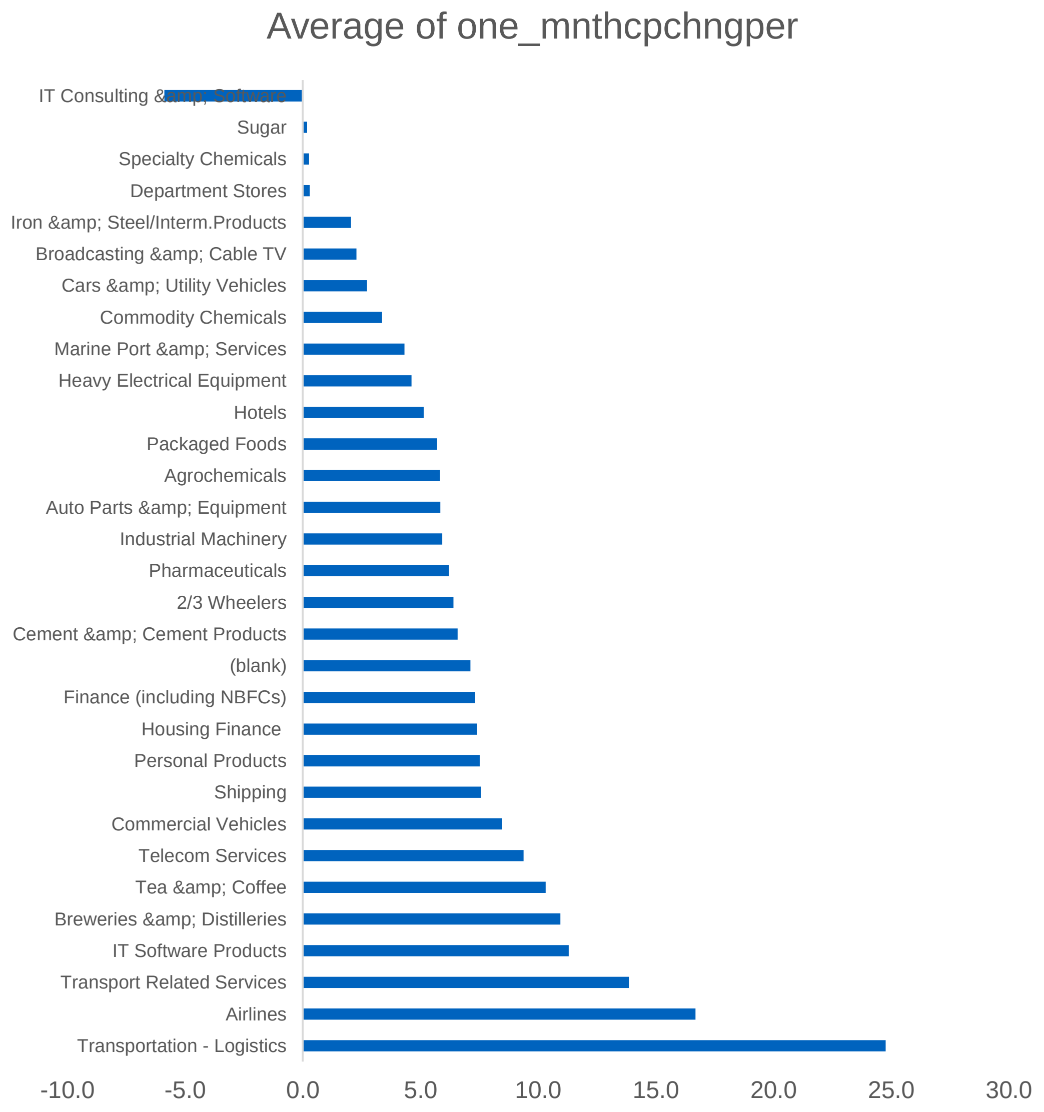
Variables	Difference	Negative (0,5.01]	Positive (5.01,8]
Average of senti_score_govt.policy	1.13	7.5	8.6
Average of senti_score_recognition_awards	0.80	4.4	5.2
Average of senti_score_product.portfolio	0.78	5.2	6.0
Average of senti_score_revenue	0.78	3.7	4.5
Average of senti_score_recent.trends	0.71	4.3	5.0
Average of senti_score_industry	0.58	5.1	5.6
Average of senti_score_ratio	0.43	5.9	6.3
Average of senti_score_pricing	0.38	4.3	4.6
Average of indicators_ratio	0.37	2.7	3.1
Average of senti_score_margins	0.29	2.7	3.0
Average of senti_score_stocks	0.24	2.9	3.1
Average of senti_score_economy	0.17	4.3	4.5
Average of senti_score_liquidity	0.09	3.0	3.1
Average of senti_score_expenses	-0.03	2.0	2.0
Average of senti_score_assets	-0.05	3.9	3.9
Average of senti_score_innovation	-0.16	7.0	6.9
Average of senti_score_merges_acquisitions	-0.24	4.5	4.3



Industry-wise Data

Exploratory Data Analysis

Industry	Average of one_mnthcpchngp
Transportation - Logistics	24.8
Airlines	16.7
Transport Related Services	13.9
IT Software Products	11.3
Breweries & Distilleries	11.0
Tea & Coffee	10.3
Telecom Services	9.4
Commercial Vehicles	8.5
Shipping	7.6
Personal Products	7.5
Housing Finance	7.4
Finance (including NBFCs)	7.3
(blank)	7.1
Cement & Cement Products	6.6
2/3 Wheelers	6.4
Pharmaceuticals	6.2
Industrial Machinery	5.9
Auto Parts & Equipment	5.9
Agrochemicals	5.8
Packaged Foods	5.7
Hotels	5.1
Heavy Electrical Equipment	4.6
Marine Port & Services	4.3
Commodity Chemicals	3.4
Cars & Utility Vehicles	2.7
Broadcasting & Cable TV	2.3
Iron & Steel/Interm.Products	2.1
Department Stores	0.3
Specialty Chemicals	0.3
Sugar	0.2
IT Consulting & Software	-5.9



Model Training & Validation – Random Forest

Data preprocessing: 598 rows for 113 unique companies after removing observations with NA values, we have 510 rows. Split has been performed by 70% (357 observations) - 30% (153 observations) as Training & Testing data respectively

Model Building:

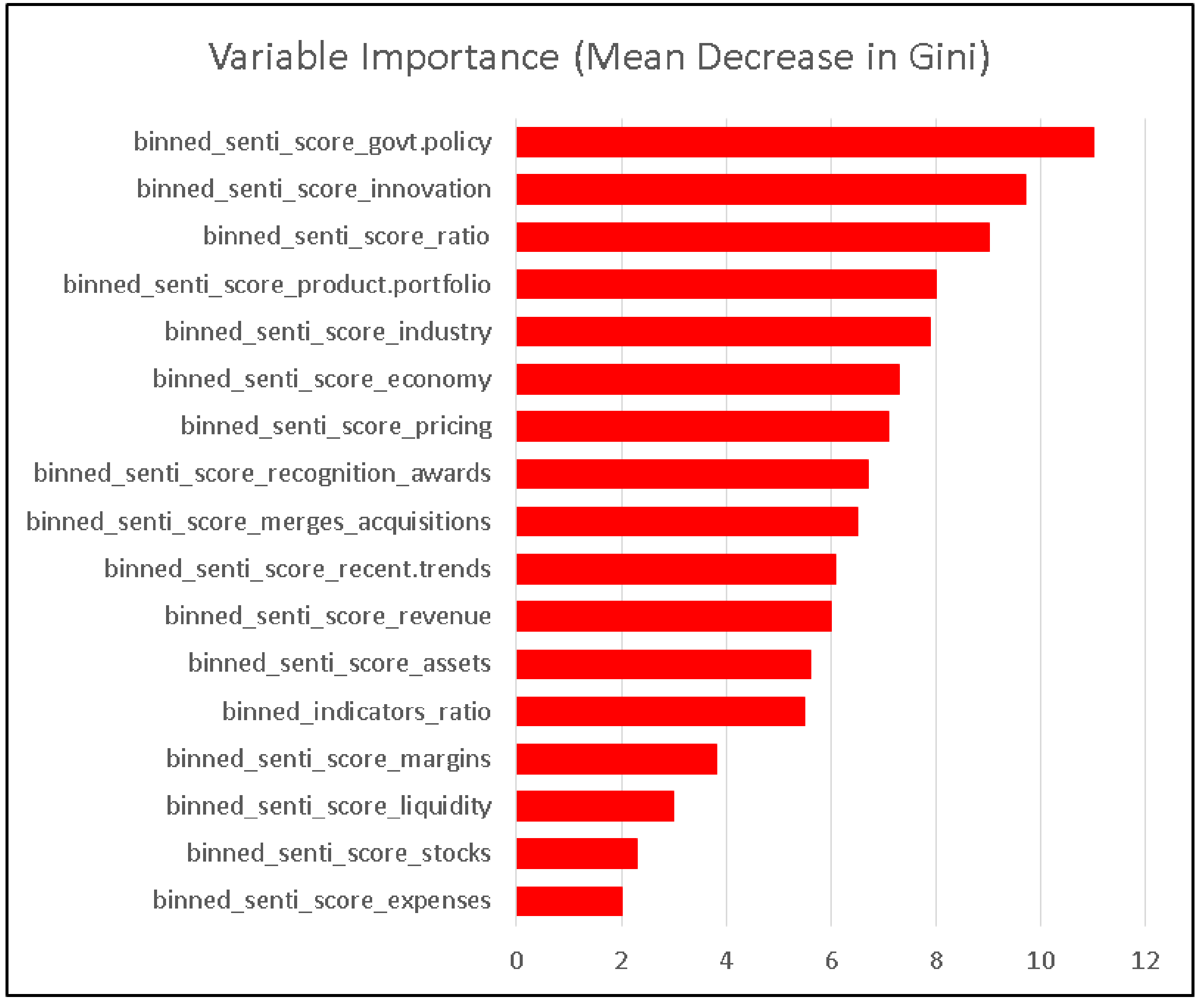
- Random forest is applied using Grid Search to tune parameters of No. of columns selected, Maximum nodes, Number of trees
- For each parameters combination, probability cut-off selected based on sensitivity-specificity chart to create confusion matrix
- Finally classified predicted probabilities of test data based on cut-off generated by sensitivity-specificity plot
- Computed Confusion matrix and calculated Accuracy, precision, recall & f1-scores on test data

Grid Search:

- During grid search we have explored on following hyper parameters
  - No.of columns selected – [2,3,4,5,6,7,8,9]
  - Max nodes – [8,16,32,64,128]
  - Number of Trees – [1000,1500,2000,5000,10000]

Best parameter combination from Grid Search:

Best parameter after performing entire space, best pairs are, No.of columns – 8, Max nodes – 64, Number of trees – 1000 with test accuracy as 61.43%



Model Training & Validation – Random Forest

- **Results:** Accuracy, Precision & Recall are calculated to measure the performance on test data with 153 observations

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

$$Precision = \frac{(TP)}{(TP + FP)}$$

$$Recall = \frac{(TP)}{(TP + FN)}$$

$$f1\ score = \frac{2 * P * R}{(P + R)}$$

Predicted Stock Price Change

Actual Stock Price Change

	Negative (0,5.01]	Positive (5.01,8]
Negative (0,5.01]	50	21
Positive (5.01,8]	38	44

	Precision	Recall	f1
Negative (0,5.01]	56.8 %	70.4%	62.87%
Positive (5.01,8]	67.6%	53.65%	59.85%
Overall	63.4%	61.2%	61.3%

Results – Results for % stock price change for 12 month duration (Y_target12)

Model	Description - binning	Accuracy	Prec	Recall	f1
Nnet (multinomial)	cuts1 = 5 # bins with equal ranges	0.702	0.56	0.67	0.59
Random Forest	cuts1 = 5 # bins with equal ranges	0.755	0.51	0.71	0.59
Nnet (multinomial)	Manually selected cuts2 = c(0,5.0,5.2,5.4,8)	0.3377	0.31	0.39	0.34
Random Forest	Manually selected cuts2 = c(0,5.0,5.2,5.4,8)	0.4238	0.33	0.47	0.34
Nnet (multinomial)	cuts3 = c(0,5.2,5.4,8)	0.6887	0.52	0.67	0.59
Random Forest	cuts3 = c(0,5.2,5.4,8)	0.7152	0.52	0.72	0.60
Nnet (multinomial)	cuts4 = c(0,log(150-10),log(150),5.4,8)	0.2781	0.32	0.38	0.32
Random Forest	cuts4 = c(0,log(150-10),log(150),5.4,8)	0.4371	0.36	0.43	0.35
Nnet (multinomial)	cuts5 = c(0,log(150-25),log(150-10), log(150),5.4,8)	0.298	0.23	0.35	0.25
Random Forest	cuts5 = c(0,log(150-25),log(150-10), log(150),5.4,8)	0.298	0.21	0.39	0.24
Nnet (multinomial)	cuts6 = c(0,log(150),8)	0.4901	0.29	0.54	0.37
Random Forest	cuts6 = c(0,log(150),8)	0.6087 (at conf threshold 0.71)	63.4%	61.2%	61.3%
SVM	cuts6 = c(0,log(150),8)	0.58			

Insights

- A classifier was built to predict stock price change (negative, positive) using Random Forest (~61% F1)
- Additional variables such as actual profits, revenues, etc. may improve the performance of the model
  - These may be parsed from the structured data tables or extracted from unstructured text by leveraging the high frequency grammatical patterns in text
- Inclusion of industry type, as a variable could enhance the performance of the model

TATA CONSULTANCY SERVICES
Experience certainty.

Thank You

References

- Ref: SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods, https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0085-1
- Sentiment Analysis using pattern
  - http://www.clips.ua.ac.be/pattern
- Important financial ratios
  - https://en.wikipedia.org/wiki/Financial_ratio
- PDF to text convertion using Apache Tika
  - https://tika.apache.org/

Software Installation and Steps of Execution

- Pip install pandas
- Pip install pattern
- Pip install tika

- Open and execute extract_pdf.py

- Open sentiment_analysis.py and modify the following lines:
  - inFile = input.csv'
  - outFile = 'result.csv'
  - corpus_folder = ‘corpus‘ #folder location where the extracted txt files for annual reports are located
  - attr_filters = ['growth', 'revenue', 'profit'] #update the list for filtering relevant sentences from the unstructured data
- Running the code
  - Python sentiment_analysis.py

- Run R codes for model training and execution
  - Run data_prep.R for preprocessing
  - Run Hack_Model_v6.R for model training and validation

Solution Approach

- Unstructured data was extracted using apache tika
- We segmented the documents into sentences
- We filtered the sentences by relevant keywords
- We ran the sentiment analyzer to extract sentiments for each sentence
- We created an aggregation score to get an overall sentiment for the relevant section of the sentiment
- We aggregated the stock market data by 1 month, 6 months & 12 months
- We merged this sentiment scores with the aggregated stock market data and fit machine learning classifiers on it
- The raw sentiment scores were binned at the following cut points (0,5,10,15,20)
- The aggregated stock market data for 1, 6 and 12 months were also transformed ln(%price change + 150) and then binned into 5 equal partitions – 150 was added in order to be able to perform ln transformation for negative stock price changes as well