Ashraf Wan
June 16, 2022
IST 736

Project #1

Introduction:

This project will be focused on getting the sentiment analysis as well as topic modeling for financial news headlines from 2014. The purpose of this is to find the sentiment analysis is to find out the sentiment for headlines or topics to help investors make the most profit in investments.

Methods:

The dataset was procured from Kaggle. The dataset contained 2 columns, headlines and sentiment with 1500 rows.

Preprocessing steps included checking to see if there's any missing values, which there wasn't any. Since the number for the three sentiments are unbalance, the same number of rows were randomly selected for each sentiment to make the data more balance and unbiased. From there, a train-test split was done with 70% for train set and 30% for test set.

Three different modeling examples ran a few times with different parameters to see the difference in performance. For each modeling algorithm, we ran 10-fold cross validation to get the mean of the training accuracy and ran prediction to get the testing accuracy as well as the classification reports to find out how each sentiment performed.

First modeling algorithm was Bernoulli Naïve Bayes which ran four times with different parameters. The different parameters are listed below.

- Test #1: Minimum document frequency at 5, remove stop words
- Test #2: Minimum document frequency at 5
- Test #3: Minimum document frequency at 10, remove stop words
- Test #4: Minimum document frequency at 10

Second modeling algorithm was Multinomial Naïve Bayes which ran twice with different parameters. The different parameters are listed below.

- Test #1: Minimum document frequency at 5, remove stop words
- Test #2: Minimum document frequency at 5

The third modeling algorithm was SVM which ran four times with different parameters. The different parameters are listed below.

- Test #1: Minimum document frequency at 5, remove stop words
- Test #2: Minimum document frequency at 5
- Test #3: Minimum document frequency at 5, remove stop words, bigram

- Test #4: Minimum document frequency at 5, remove stop words, bigram

The final modeling algorithm is the topic modeling. LDA was used to run two tests with different parameters. The different parameters are listed below.

- Test #1: Minimum document frequency at 5, remove stop words
- Test #2: Minimum document frequency at 5, remove stop words, bigram

Results:

Bernoulli Naïve Bayes:

Test #1

Mean of 10-fold cross validation accuracy: 60.19%

Test accuracy: 56%

	precision	recall	f1-score	support
neutral	0.65	0.60	0.62	157
positive	0.51	0.76	0.61	146
negative	0.53	0.32	0.40	147
accuracy			0.56	450
macro avg	0.57	0.56	0.54	450
weighted avg	0.57	0.56	0.55	450

Test #2

Mean of 10-fold cross validation accuracy: 61.62%

Test accuracy: 59.34%

	precision	recall	f1-score	support
neutral	0.68	0.60	0.64	157
positive	0.53	0.79	0.64	146
negative	0.61	0.39	0.48	147
accuracy			0.59	450
macro avg	0.61	0.59	0.58	450
weighted avg	0.61	0.59	0.58	450

Test #3

Mean of 10-fold cross validation accuracy: 56.76%

Test accuracy: 51.33%

Ashraf Wan
June 16, 2022
IST 736

	precision	recall	f1-score	support
neutral	0.62	0.50	0.56	157
positive	0.46	0.77	0.57	146
negative	0.51	0.27	0.36	147
accuracy			0.51	450
macro avg	0.53	0.51	0.49	450
weighted avg	0.53	0.51	0.50	450

Test #4

Mean of 10-fold cross validation accuracy: 59.05%

Test accuracy: 55.78%

	precision	recall	f1-score	support
neutral	0.65	0.54	0.59	157
positive	0.50	0.79	0.61	146
negative	0.59	0.35	0.44	147
accuracy			0.56	450
macro avg	0.58	0.56	0.54	450
weighted avg	0.58	0.56	0.55	450

Multinomial Naïve Bayes:

Test #1

Mean of 10-fold cross validation accuracy: 60%

Test accuracy: 56.44%

	precision	recall	f1-score	support
neutral	0.63	0.65	0.64	157
positive	0.53	0.63	0.57	146
negative	0.53	0.41	0.46	147
accuracy			0.56	450
macro avg	0.56	0.56	0.56	450
weighted avg	0.56	0.56	0.56	450

Test #2

Mean of 10-fold cross validation accuracy: 58.44%

Test accuracy: 60.38%

Ashraf Wan
June 16, 2022
IST 736

	precision	recall	f1-score	support
neutral	0.64	0.64	0.64	157
positive	0.54	0.70	0.61	146
negative	0.58	0.41	0.48	147
accuracy			0.58	450
macro avg	0.59	0.58	0.58	450
weighted avg	0.59	0.58	0.58	450

SVM:

Test #1

Mean of 10-fold cross validation accuracy: 60%

Test accuracy: 61.55%

	precision	recall	f1-score	support
neutral	0.71	0.61	0.66	157
positive	0.55	0.63	0.59	146
negative	0.61	0.61	0.61	147
accuracy			0.62	450
macro avg	0.62	0.62	0.62	450
weighted avg	0.62	0.62	0.62	450

Test #2

Mean of 10-fold cross validation accuracy: 63.71%

Test accuracy: 62.45%

	precision	recall	f1-score	support
neutral	0.72	0.67	0.70	157
positive	0.57	0.62	0.59	146
negative	0.58	0.59	0.58	147
accuracy			0.62	450
macro avg	0.63	0.62	0.62	450
weighted avg	0.63	0.62	0.63	450

Test #3

Mean of 10-fold cross validation accuracy: 61.43%

Test accuracy: 59.78%

Ashraf Wan
June 16, 2022
IST 736

	precision	recall	f1-score	support
neutral	0.69	0.58	0.63	157
positive	0.55	0.62	0.58	146
negative	0.58	0.60	0.59	147
accuracy			0.60	450
macro avg	0.60	0.60	0.60	450
weighted avg	0.61	0.60	0.60	450

Test #4

Mean of 10-fold cross validation accuracy: 64.19%

Test accuracy: 62%

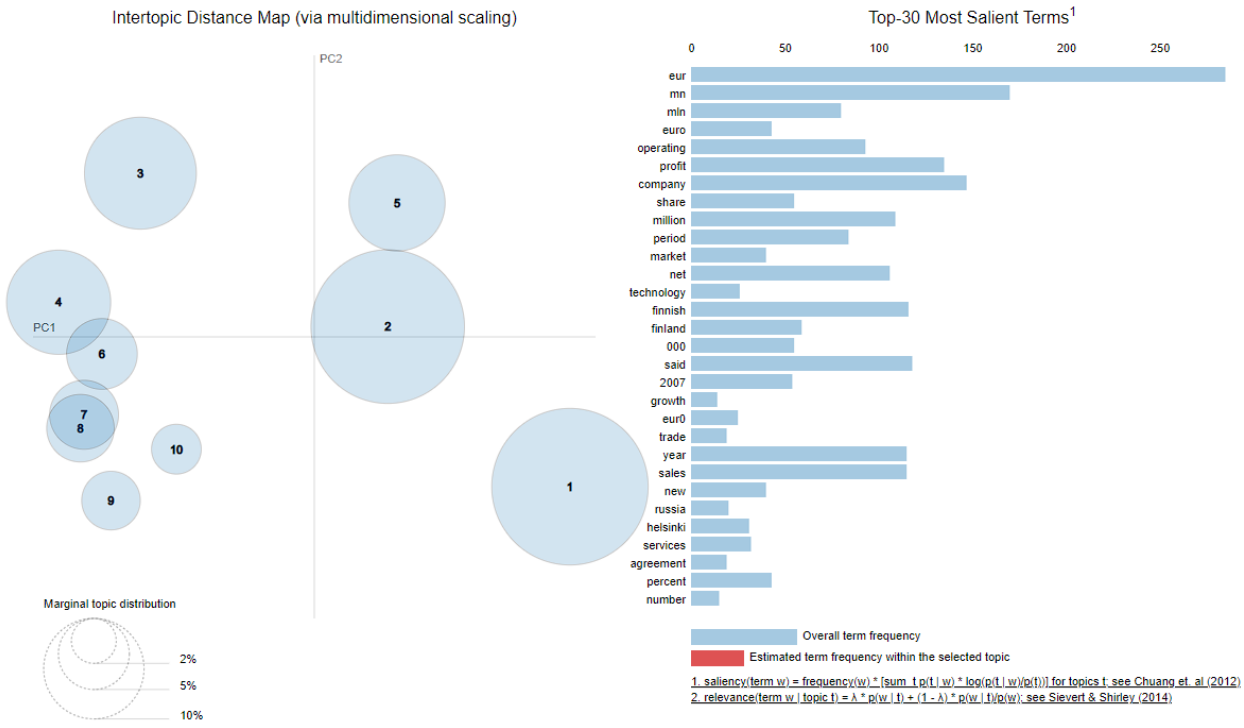
	precision	recall	f1-score	support
neutral	0.69	0.62	0.65	157
positive	0.57	0.62	0.59	146
negative	0.61	0.61	0.61	147
accuracy			0.62	450
macro avg	0.62	0.62	0.62	450
weighted avg	0.62	0.62	0.62	450

Topic Modeling:

Test #1

```
Topic 0:
company business according equipment building 000 mobile software addition construction
Topic 1:
eur mn profit operating sales period net quarter compared million
Topic 2:
market services products company including based customers property portfolio nordic
Topic 3:
growth lay target temporary current temporarily period end store 20
Topic 4:
company finnish finland said group corporation media bank hel 2008
Topic 5:
number employees new finland elcoteq group outotec electronics personnel 10
Topic 6:
mln euro share eur0 2007 net 000 capital finnish 2006
Topic 7:
technology agreement production oil plant power lines parties supply said
Topic 8:
trade russia long term pleased financial head helsinki known said
Topic 9:
year million said sales finnish company percent net 2009 oyj
```

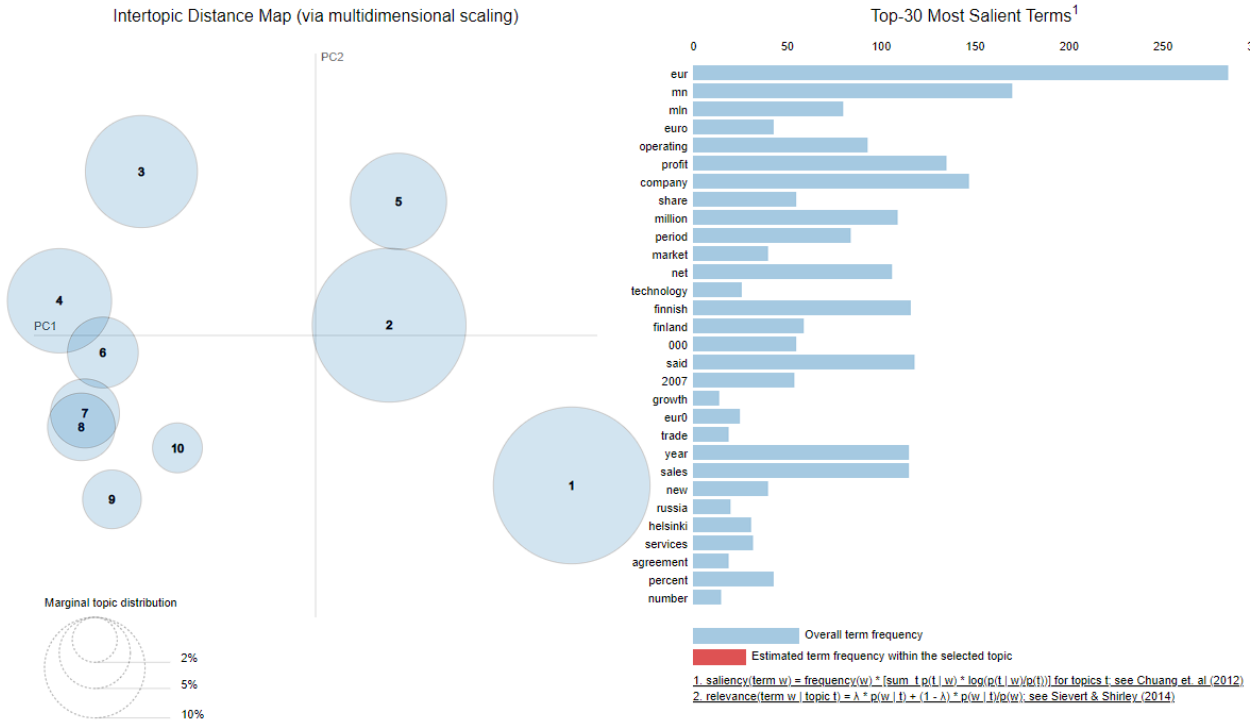
Ashraf Wan
June 16, 2022
IST 736



Test #2

```
Topic 0:
cut building area total cash costs laid additional 000 plan
Topic 1:
nokia said software pleased costs long term mobile china long term
Topic 2:
share eur0 capital eps price helsinki share capital earnings shares eur
Topic 3:
mln euro company period net said profit mln euro year finnish
Topic 4:
billion russia known trade total year nordic 30 staff 40
Topic 5:
eur mn million year finnish 2009 profit net 2008 group
Topic 6:
eur mn operating sales profit operating profit percent eur mn non mn eur
Topic 7:
company market finland business services bank production share customers new
Topic 8:
company new construction units products personnel said sales employees 2011
Topic 9:
service company finland elcoteq services director addition manufacturing financial kone
```

Ashraf Wan
June 16, 2022
IST 736



Conclusions:

The results show that the accuracies are generally higher for when stopwords are included compared to when stopwords are removed. In terms of financial news headlines, it is better to leave the stopwords when performing modeling algorithm because they carry some weight when it comes to getting sentiment from news headlines.

Even with the same numbers of data for each sentiment, it seems the modeling algorithm had trouble identifying positive sentiments while it had an easier time identifying negative and neutral sentiments. However, the positive sentiments have a higher recall percentage compared to the other sentiments in all the modeling algorithm except for Multinomial Naïve Bayes Test #1 and SVM Test #2. Even though the positive sentiments are harder to identify, the ones that are identified are more likely to be identified again.

According to the results, the best modeling algorithm is the Bernoulli Naïve Bayes Test #2 because it has a high training and testing accuracy as well as a high precision and recall percentage.

For topic modeling, it seems that most frequent topics between the two tests are the same but with different top 10 phrases in it. According to the graph, Topic #7 and #8 seem to be similar since the distance between the two are very close. Topic #1 and #3

are the farthest away from other topics. All the other topics are distant but in close proximity of one another.

Data Source:

https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news