# Optimized Stock Price Prediction from News

## Bryan Cheong, Christopher Yuan, Stephen Ou

### Abstract

We use several supervised learning algorithms in conjunction with a ‘bag-of-words’ representation of reputable news articles to predict stock price movement. With model selection and parameter tuning, we were able to achieve 60% accuracy on our test set.

### Classification Pipeline

1. **Preprocessing**
   - The stock market took a major hit today as the Spring CS229 class presented their final projects. Supervised learning methods either did not work or caused several companies including Apple, BP, and Honda to go bankrupt. President Obama issued a plea to students.

2. **Bag of Words**
   - The Wall Street Journal

3. **Labels (prices)**
   - **Date**: 5/11
   - **BA**: ↑
   - **GM**: ↓
   - **GE**: ↑
   - **BP**: ↓
   - **FB**: ↑

4. **TF-IDF term weighting**
   - Decreases weights of common words
   - $\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$
   - $\text{tf}(t, d) = 1 + \log(\text{tf}_{t,d})$
   - $\text{idf}(t, D) = \log\left(\frac{N}{\sum_{d \in D} \text{df}_t}ight)$

5. **Optimizations**
   - Stemming
     - (discouraging, discouraged, discourage) $\rightarrow$ discourag
     - (diversified, diversify, diversifying) $\rightarrow$ diversifi

6. **Parameter tuning**
   - Naive Bayes: Laplace smoothing
   - SVM: kernel functions, type of regularization
   - Boosting: base estimator, number of estimators
   - Perceptron: loss functions, type of penalty

### Test Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.45</td>
</tr>
<tr>
<td>Stemming</td>
<td>0.49</td>
</tr>
<tr>
<td>Stemming + TF-IDF</td>
<td>0.53</td>
</tr>
<tr>
<td>Stemming + TF-IDF + Best Params</td>
<td>0.61</td>
</tr>
</tbody>
</table>

### Most Relevant Terms

#### Positive
- suggest: 0.54
- rumor: 0.51
- high: 0.51
- report: 0.48
- rise: 0.45
- express: 0.44
- staff: 0.43
- soar: 0.42
- diversified: 0.40
- reorganization: 0.40

#### Negative
- worry: -0.51
- discourage: -0.48
- misinform: -0.48
- reprehensible: -0.46
- abusive: -0.45
- dispute: -0.45
- denounced: -0.45
- cheapen: -0.43
- rough: -0.41
- takeover: -0.40

(Used unstemmed words for readability)

### Analysis & Future Work

- We can predict changes in companies’ stock value using words from newspaper articles
- Refinements using model selection tools significantly improve the predictive power of the baseline bag-of-words model
- Future applications could focus on simulating the performance of making stock purchase and sale decisions using our optimized model

### References:

3. Timmons and Ryan. “Predicting the stock market with news articles.”