Enabling News Trading by Automatic Categorisation of News Articles

Vishwanath Kolar Bhaskara, Praveen Kumar Kotekal

MS in Business Analytics, Oklahoma State University

ABSTRACT
Traders making decisions based on news developments is nothing new. Any big market announcements of a company such as annual and quarterly earnings, dividend announcements, acquisitions, mergers, tender offers, stock splits or major management changes are known to have direct impact on the company’s stock prices and news traders have always evinced keen interest to exploit and act on this information. However, in an age where news is novel only for few minutes it is important for traders to identify its underlying message, assess its possible impact on stock trends and take on the spot decisions before the market has had time to adjust itself to this news.

Through this project we predicted the direction of stock price changes immediately after the news article publication and also major factors that influence either rise or fall of stock prices. We built automatic text categorization models using SAS Content Categorisation studio and Text Rule Builder that categorize into various custom topics in any published news article that eventually indicates direction of stock price (positive, negative). Initial data set is created from scraping the news articles of Apple Inc. along with its time stamp from TheEconomist.com, NY Times, Wall Street Journal, Reuters.com using Python web crawler. We have collected stock price of Apple Inc. for corresponding time stamps of news article publication. Final data set consists of 1968 articles with its title, Change in stock price, time stamp and news content since April 2014. We pre-labeled news articles in this dataset as positive and negative based on the changes in stock prices immediately after the publication of the article.

Introduction (Continued)
News articles written about companies influence people either consciously or unconsciously in their decision process when trading in the stock market. News related to Annual and quarterly earnings, dividend announcements, acquisitions, mergers, tender offers, stock splits, and major management changes, and any substantive items of unusual or non-recurrent nature are examples of news items that are useful for traders in their trading decisions. These types of news are usually published immediately as breaking news and are often given to the press directly from the companies. To analyse the data from various websites, Python tool is built which automatically fetches text data from various links. We fetched articles related to AAPL from 2014 April 1st to 2016 July 10th.

Data Fetching Process
Process in Figure B shows the Python tool used for Extracting text data from News Website:
1. First argument takes list of News websites from which data needs to extracted
2. Second argument takes ticker symbol on which data needs to be extracted. For example: Apple company results are extracted using AAPL
3. Third argument takes start date and end date during which published URLs are fetched to extract text data
4. Output of this process is CSV file including date time of published article, title, and text data of the article

Default arguments for this tool include: News sources (TheEconomist.com, NYTimes, Wall Street Journal, Reuters.com), Type of news (Management changes, quarterly earnings, dividend announcements, acquisitions, mergers, tender offers)
Methods and Results

The goal of text classification is to assign some article to either positive or negative news. Below steps are followed for text pre-processing:

- Replace special characters
- Remove duplicate characters
- Replace numbers
- Remove default stop words and user defined stop words
- Set stem words to True

Below steps are followed as Part of Feature selection:

- Feature scoring method – Mutual information (Top 5000 most relevant features with respect to positive/negative label)
- Unigrams TF-IDF feature extraction

We labelled the same dataset using both the methods and created a new data set that consists only of documents that are labeled with the same sentiment in both of the other sets.

Methods and Results (continued)

Categorization has been performed using following methods:

1. SAS Enterprise Miner - Rule builder node
2. SAS Enterprise Miner – Traditional predictive model with input variables from Text cluster node and Text topic node.
3. Content Categorization studio: Statistical Model.
4. Content categorization studio – Rule based model based on rules generated from Rule builder node in enterprise miner and enhancing the rules

Model 1: SAS EM Text Rule Builder node Model:
Text rule builder is a Boolean rule base categorizer that automatically generates an ordered set of rules for describing and predicting a target variable.

Data Partition Node: Set Training Data – 80% and Test Data 20%.


Text Filter Node: Added a new dictionary and set the Term weight to Mutual Information as our data set has a target variable.

Rule Builder node: Remove default stop words and user defined stop words

Specifications in properties panel: Generalization error, Purity of rules, Exhaustiveness – Low

Stock prices for Apple company are fetched from yahoo finance news. Stock prices are fetched for every two hours on a day ranging from April 1st 2014.

We labelled the same dataset using both the methods and created a new data set that consists only of documents that are labeled with the same sentiment in both of the other sets.
Methods and Results (continued)

Text Profile Node: Results

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term1</td>
<td>Value1</td>
</tr>
<tr>
<td>Term2</td>
<td>Value2</td>
</tr>
<tr>
<td>Term3</td>
<td>Value3</td>
</tr>
<tr>
<td>Term4</td>
<td>Value4</td>
</tr>
<tr>
<td>Term5</td>
<td>Value5</td>
</tr>
<tr>
<td>Term6</td>
<td>Value6</td>
</tr>
<tr>
<td>Term7</td>
<td>Value7</td>
</tr>
</tbody>
</table>

Text Topic Node: Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category1</td>
<td>Value1</td>
</tr>
<tr>
<td>Category2</td>
<td>Value2</td>
</tr>
<tr>
<td>Category3</td>
<td>Value3</td>
</tr>
<tr>
<td>Category4</td>
<td>Value4</td>
</tr>
<tr>
<td>Category5</td>
<td>Value5</td>
</tr>
<tr>
<td>Category6</td>
<td>Value6</td>
</tr>
</tbody>
</table>

Rule Builder Node rules:

- Rule1: If term is present then category is positive
- Rule2: If term is not present then category is negative

The text profile and text topic node results are important to understand what terms are mainly influential in categorization. As it can be seen – partnerships, increase in sales, appstore sales, mac sales and dividend are key terms in positive category while slow growth, ios, earnings decline, application etc., are key terms in negative category. This categorization helps in understanding the rules and further modifying them in content cat studio if required.

Methods and Results (continued)

Model 2: Predictive Model with input variables from Text cluster node and Text topic node:

- We used Regression, Decision tree and neural network nodes with below selection criteria.

  - Regression: Stepwise selection criteria- Average square error as assessment measure.
  - Decision Tree: Largest and Decision as assessment measure
  - Neural network: Average square error as assessment measure

We used Model comparison node to compare the results of the above three with Rule builder node, with validation misclassification rate as selection criteria. Text Rule builder model was picked as the best model based on the criteria.

Methods and Results (continued)

Model 3: Content Categorization Studio: Statistical Model:

- It is a black box model which categorizes articles based on internal algorithms. We need a pre-labelled target variable for the statistical model.

Model 4: Content Categorization Studio: Rule based model by modifying rules generated by Text Rule Builder Node in Enterprise Miner:

- We have used the rules obtained in text builder and modified them in content categorization studio to automatically categorize the documents. This doesn’t require a target variable as it classifies the documents based on the rules. We performed an iterative process by studying the applied rules on both passed and failed documents. In failed documents we looked for keywords that identify them into particular category and added them into rules. After this iterative process we built a consistent set of rules that categorize the data into positive or negative categories. We found the accuracy to be highest when using the rule based model in content categorization studio for this dataset.
Based on observing text topics, text profile node results and rules applied in failed documents, we identified certain keyword combinations that should definitely belong to either positive or negative categories and rules were modified accordingly. Some changes made to the rules generated by text rule builder node are mentioned below:

1. Keyword “China” has been added to negative rules, because as per concept links China is strongly associated to “growth” which is a key term in negative category of text profile results.
2. “Operating system” has been added to positive rules as it has been part of the topic along with keyword “profitable” which is in the positive rules.
3. (AND,“yuan”,(OR, “devalution”, “devaluation”, “devalue”)) has been added to negative rules as it’s been identified as a repetitive key word in most of the failed categorizations and it linked negatively to apple’s stock price.
4. (AND,(OR, “increase”, “increases”), (OR, “mac sale”, “mac sales”)) has been added to the positive rule set as it’s found to have positive effect on stock prices.
5. “Promotional Style” has been added to negative keyword list.

<table>
<thead>
<tr>
<th>Category</th>
<th>Sensitivity</th>
<th>Number of Passed Documents</th>
<th>Number of Failed Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>83%</td>
<td>223</td>
<td>48</td>
</tr>
<tr>
<td>Negative</td>
<td>87%</td>
<td>275</td>
<td>41</td>
</tr>
</tbody>
</table>

Best Model: Model -4

- Of the 4 different models that have been built, model 4 which is Content Categorization Studio: Rule based model by modifying rules generated by Text Rule Builder Node in Enterprise Miner is chosen as the final/best model as it categorizes the articles with highest accuracy.
- As the rules are manually changed after observing the produced rules and also the results from text topic and text profile node, this model is the most consistent among others.

Factors influencing stock prices:

Upon text analysis, based on the rules generated and domain knowledge, below are our findings regarding what factors influenced apple stock prices over time:

- Topics that positively influence stock prices:
  - New Partnerships
  - Mac sales
  - Operating systems
  - Appstore sales
  - Increase in sales in China
  - Earnings growth

- Topics that negatively influence stock prices:
  - Earnings Decline
  - Big product hit (This should be other company’s big product)
  - Iphone/mac sales going down
  - Promotional style
  - Yuan’s devaluation

References

1. Chakraborty, Goutam; Pagolu, Murali and Garla, Satish. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®

Acknowledgment

We heartily thank Dr. Goutam Chakraborty, Professor, Department of Marketing and founder of SAS® and OSU Data Mining Certificate Program - Oklahoma State University for his support throughout the research.

Contact

Your comments and questions are valued and encouraged. Contact the authors at Vishwanath Kolar Bhaskara, Praveen Kumar Kotekal, Oklahoma State University, Stillwater, OK, 74075 Email: kolarbh@ostatemail.okstate.edu, praveen.kotekal@okstate.edu

Based on observing text topics, text profile node results and rules applied in failed documents, we identified certain keyword combinations that should definitely belong to either positive or negative categories and rules were modified accordingly. Some changes made to the rules generated by text rule builder node are mentioned below:

1. Keyword “China” has been added to negative rules, because as per concept links China is strongly associated to “growth” which is a key term in negative category of text profile results.
2. “Operating system” has been added to positive rules as it has been part of the topic along with keyword “profitable” which is in the positive rules.
3. (AND,“yuan”,(OR, “devalution”, “devaluation”, “devalue”)) has been added to negative rules as it’s been identified as a repetitive key word in most of the failed categorizations and it linked negatively to apple’s stock price.
4. (AND,(OR, “increase”, “increases”), (OR, “mac sale”, “mac sales”)) has been added to the positive rule set as it’s found to have positive effect on stock prices.
5. “Promotional Style” has been added to negative keyword list.

<table>
<thead>
<tr>
<th>Category</th>
<th>Sensitivity</th>
<th>Number of Passed Documents</th>
<th>Number of Failed Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>83%</td>
<td>223</td>
<td>48</td>
</tr>
<tr>
<td>Negative</td>
<td>87%</td>
<td>275</td>
<td>41</td>
</tr>
</tbody>
</table>

Best Model: Model -4

- Of the 4 different models that have been built, model 4 which is Content Categorization Studio: Rule based model by modifying rules generated by Text Rule Builder Node in Enterprise Miner is chosen as the final/best model as it categorizes the articles with highest accuracy.
- As the rules are manually changed after observing the produced rules and also the results from text topic and text profile node, this model is the most consistent among others.

Factors influencing stock prices:

Upon text analysis, based on the rules generated and domain knowledge, below are our findings regarding what factors influenced apple stock prices over time:

- Topics that positively influence stock prices:
  - New Partnerships
  - Mac sales
  - Operating systems
  - Appstore sales
  - Increase in sales in China
  - Earnings growth

- Topics that negatively influence stock prices:
  - Earnings Decline
  - Big product hit (This should be other company’s big product)
  - Iphone/mac sales going down
  - Promotional style
  - Yuan’s devaluation

References

1. Chakraborty, Goutam; Pagolu, Murali and Garla, Satish. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®

Acknowledgment

We heartily thank Dr. Goutam Chakraborty, Professor, Department of Marketing and founder of SAS® and OSU Data Mining Certificate Program - Oklahoma State University for his support throughout the research.

Contact

Your comments and questions are valued and encouraged. Contact the authors at Vishwanath Kolar Bhaskara, Praveen Kumar Kotekal, Oklahoma State University, Stillwater, OK, 74075 Email: kolarbh@ostatemail.okstate.edu, praveen.kotekal@okstate.edu