SCSUG 2016 – Paper AA22

Enabling News Trading by Automatic Categorization of News Articles

Praveen Kumar Kotekal, Oklahoma State University

Vishwanath Kolar Bhaskara, Oklahoma State University

ABSTRACT

Traders making decisions based on news developments is nothing new. Any big market announcement 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 Categorization studio and Text Rule Builder that categorize a published news article into positive or negative categories that eventually indicates direction of stock price. Dataset was 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 prices 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

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 analyze 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.

PROJECT CONSIDERATIONS

IDENTIFICATION OF POTENTIAL BENEFACTORS

This study particularly helps traders as a tool to help making better trading decisions. With this kind of system, it is easy to predict movement of stock prices in future. Thus this system helps in correcting actions immediately and act properly in making trading decisions to gain more profits and prevent losses

DATA PREPARATION

The textual data for our analysis is prepared using following steps. These steps include data extraction from the news websites, importing the textual data in SAS environment to create a SAS dataset, parsing the textual data to identify the term-document matrix and identify the linguistic terms, Text filtering to check for spelling errors using the dictionary. The detailed flow of the data preparation is explained in following steps.

Data extraction

Process in Figure B shows the Python tool used for Extracting text data from News Website:

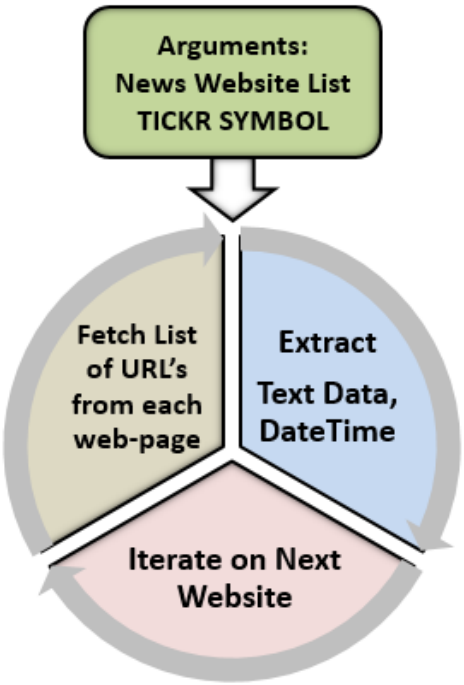


FIGURE B

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 URL's 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

DATA DICTIONARY

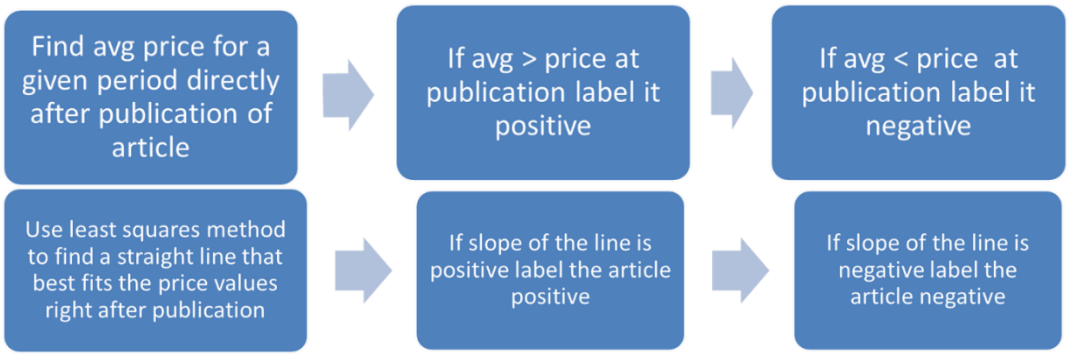
Variable	Description
ID	This variable describes unique article number
TEXT NEWS	This variable represents the actual news article posted on a news website
NEWS LABEL	This variable denotes movement of stock price after the outbreak of news on the website.

Table 1. Data Dictionary

NEWS LABELLING

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.



Text parsing

The textual dataset generated by SAS® Text import node is parsed to enumerate the terms contained in the document. It identifies the word terms based on various parts of speech present in the document. Following properties are altered in properties panel of Text Parsing node.

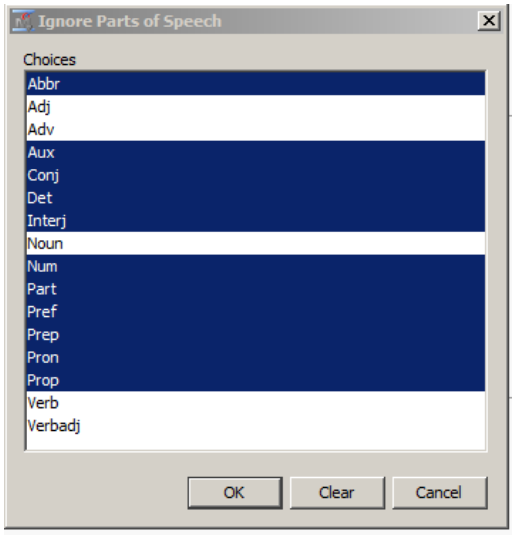
“Detect different parts of speech” is turned off which limits the terms with same parts of speech.

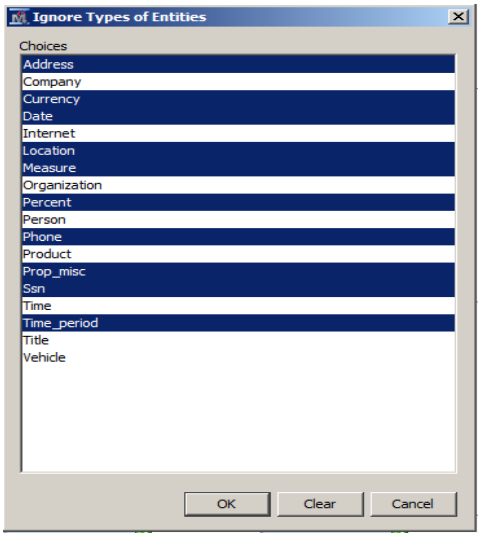
“Find Entities” is set to “Standard”.

Following parts of speech were ignored which filter prepositions, determinants, auxiliary verbs etc. which generally contains very less information. More information on ignored parts of speech shown in the below diagram.

We have ignored Numeric and Punctuation attributes.

Also entities such as Currency, Internet, Measure, Person etc. are ignored. More information on ignored types of entities shown in the below diagram’s





The text parsing node also generates the term by frequency document matrix which is used to understand the most frequently occurring term and the number of documents it has occurred in. It is

also used to analyze the terms that are rarely used. Ideally the terms that are used moderately are the ones that are the most helpful in exploration and modeling.

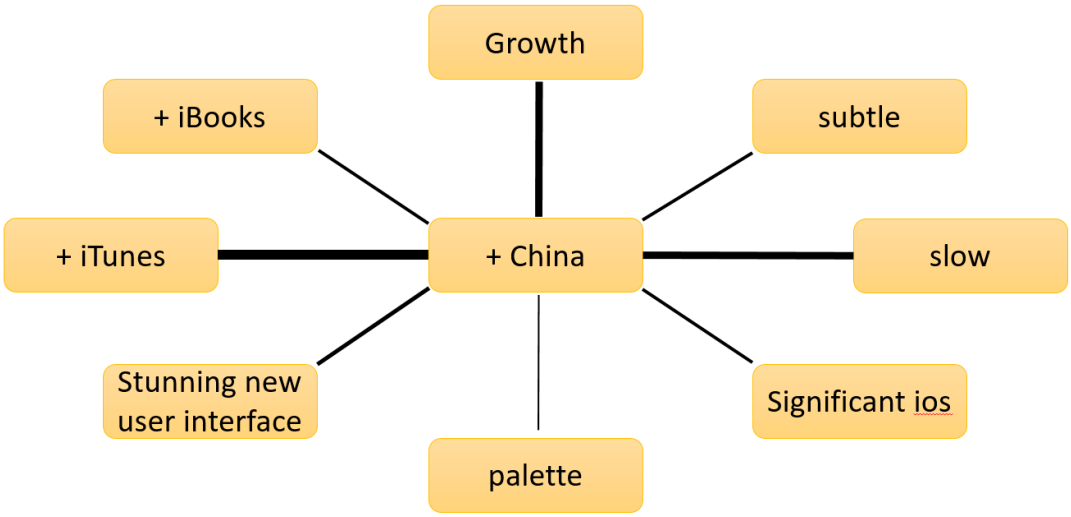
Text Filtering

To reduce the number of terms used in the documents, text filter node is used. We have used English dictionary to identify and correct the spell check errors; if not handled, will result in keeping similar words as different terms expanding the term document matrix. Using filter viewer, we can view which all documents contain a specific term and also create concept links based on those terms. We have used Text filtering node with term weight property as “Inverse Document Frequency”.

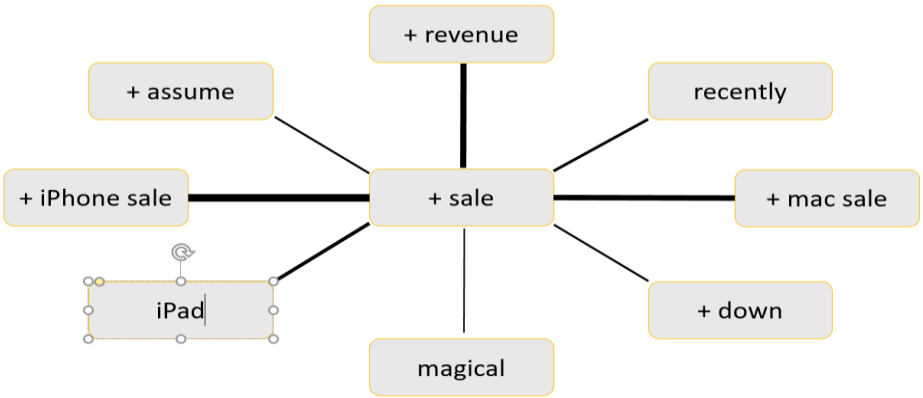
The “Check Spelling” option corrected wrong spellings of the word “accessories” as shown below.

The tables listed above gives the number of rides at borough level for the next one week. The count is obtained by taking the weather metrics for the same duration and performing scenario analysis with those values. The resulting output is log transformed, actual counts are obtained by the exponentiation of the output.

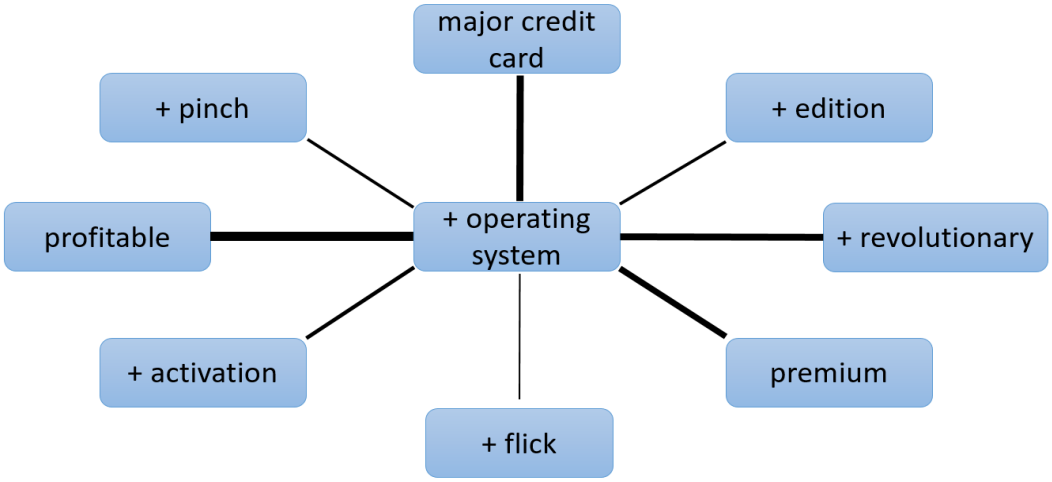
Exploratory Analysis – Concept Links:



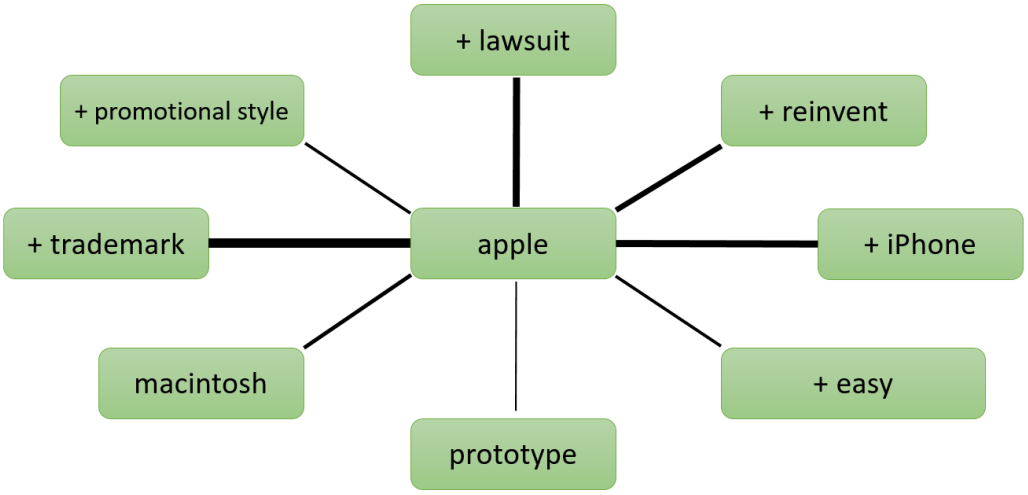
As it can be seen China has a strong link with growth and slow. This information along with text profile results was used to enhance rules in content categorization studio. For example, China was added to negative rules as it is strongly linked with growth which is in negative profile. Also China has a strong link with iTunes as Apple’s iTunes and iBooks were banned in China during the period 2015.



IPhone and Mac sales are strongly linked to the word sale. Interestingly Ipad has a a weaker link compared to Iphone sale and mac sale.



Operating system is strongly linked to profitable and this is added in the rules as profitable features in positive profile. Also Mac's operating system has always been considered revolutionary which can be seen by a strong association.



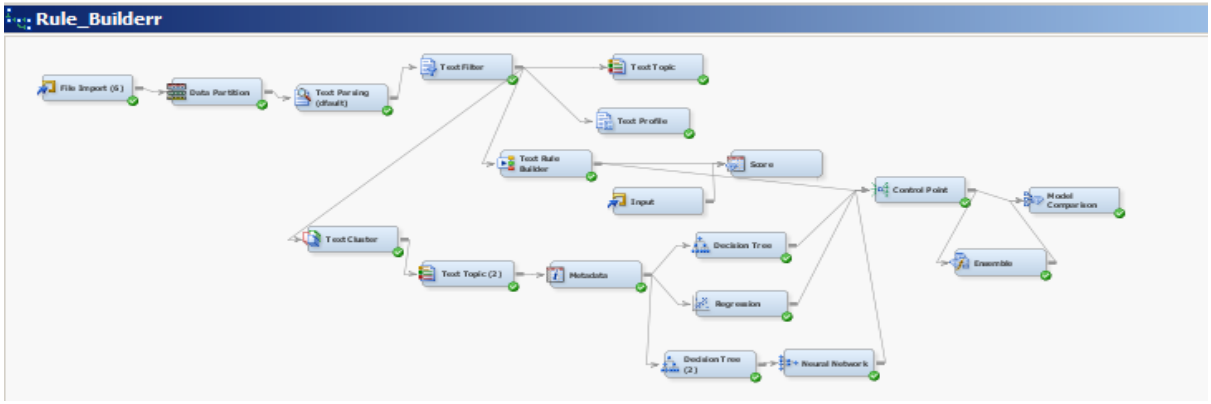
Apple has a strong association with the words trademark and iPhone.

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

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

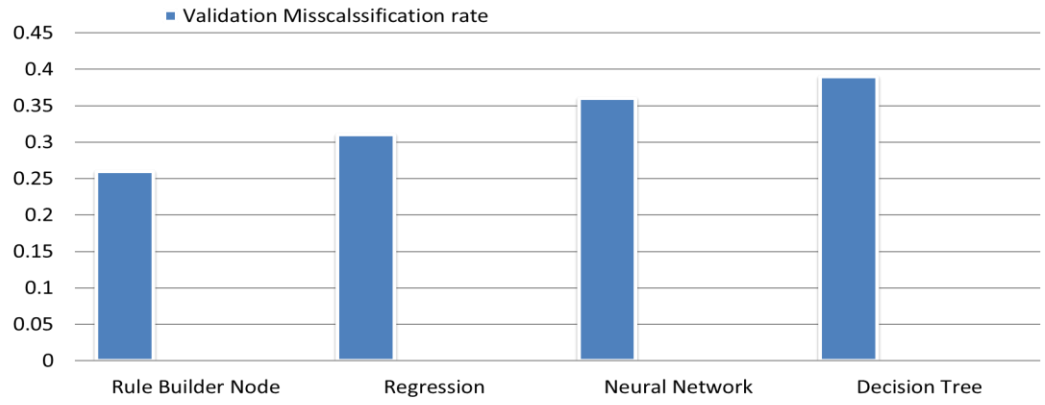


Model 2: Predictive Model with input variables from Text cluster node and Text topic node:

With Text clusters and Text topics as inputs, we used Regression, Decision tree and neural network nodes with below selection criteria.

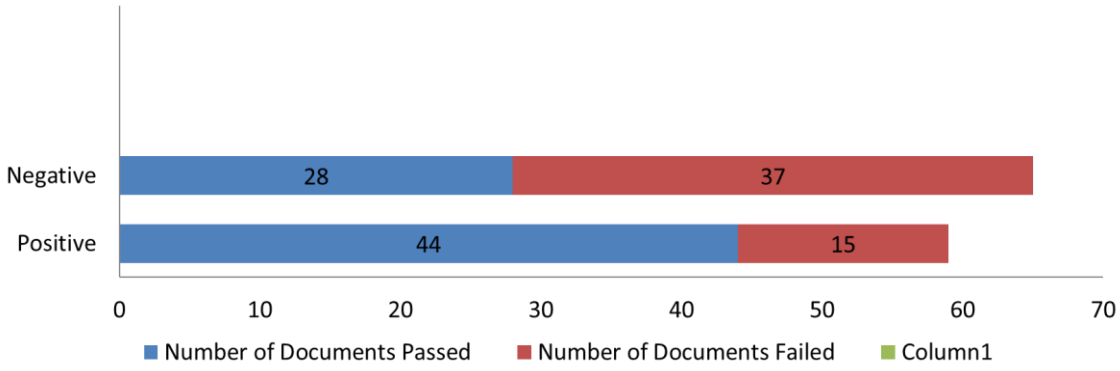
Regression	-Stepwise selection criteria- Average square error as assessment measure.
Decision Tree	Largest and Decision as assessment measure
Neural network	Average square error as assessment measure

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.



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.



Category	Sensitivity
Positive	75%
Negative	43%

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, “devaluation”, “devalution”, devalue”)) has been added to negative rules as it’s been identified as a reptitive 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

Category	Sensitivity
Positive	83%
Negative	87%

CONCLUSION

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

FUTURE WORK

The scope of the project can be extended to hourly analysis. Hourly analysis can bring more insights in terms of optimization. Analyzing at least a year worth of data will bring further insights about seasonality. Demand can be more accurately predicted if the actual number of rides requested information is available along with the live number of rides.

REFERENCES

http://toddwschneider.com/posts/a-tale-of-twenty-two-million-citi-bikes-analyzing-the-nyc-bike-share-system/%20-%20citibike-weather

https://newsroom.uber.com/uberdata-uber-for-style-and-comfort/

ACKNOWLEDGEMENT

We thank Dr. Goutam Chakraborty, Professor, Department of Marketing, Director of Business Analytics program- Oklahoma State University for his continuous support and guidance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Praveen Kumar Kotekal

Masters in Business Analytics Program

Oklahoma State University, Stillwater, OK, 74075

Phone: 405-712-4399

Email: praveen.kotekal@okstate.edu

Vishwanath Kolar Bhaskara,

Masters in Business Analytics Program

Oklahoma State University, Stillwater, OK, 74075

Phone: 405-712-2934

Oklahoma State University Stillwater, OK, 74075

Email: kolarbh@ostatemail.okstate.edu