Identifying and Predicting Product Launch Articles with Machine Learning Algorithms

Fangfang Lee (NYU), Raghav Sawhney (VT), Eirik Iversen (VT), Luke Kim (VT)

SDAL: Ian Crandell, Gizem Korkmaz, and Stephanie Shipp

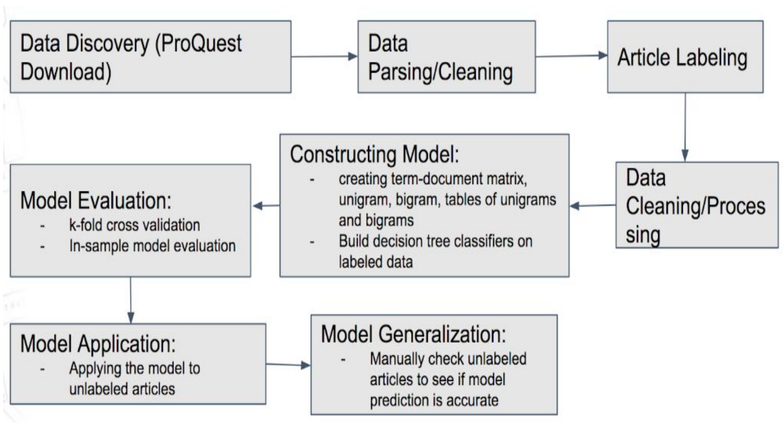
Sponsor: Gary Anderson, The National Center for Science & Engineering Statistics (NCSES) at the National Science Foundation

Introduction

- Product innovation is traditionally measured through surveys of selected companies. NCSES is interested in exploring the possibility of using non-traditional data sources to measure product innovation in lieu of surveys.
- This project assesses the feasibility of detecting news articles that are about product launches using an automated machine learning model.
- Specifically, we test out whether machine learning algorithms can automate the process of predicting whether an article is about product innovation.

Data and Methods

- We collected our data from ProQuest, a publication aggregator. It is an excellent asset for collecting information on business innovation because it compiles a wide variety of articles in all sectors, which includes comprehensive news articles on product announcement and FDA approval.

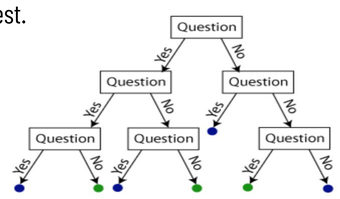


- **The "Dictionary Approach":**
  - Without using any machine learning model, can we manually pick out the words or phrases that predict a product launch article?
  - Given the appearance of a word, what is the probability of the article being about a product launch?
  - Using manually labeled articles, the conditional probability, or positive predictive value (PPV) of an article being a launch can be calculated with Bayes Theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

  - We can then use the PPV to predict which articles are most likely to be launch based on the words.
  - Can we compare this naive approach to the machine learning models?

- **Classification and Regression Tree (CART) [1]:**
  - Decision trees stratify or segment the predictor space into a number of sub regions.
  - A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails).
  - Each branch represents the outcome of the test.
  - Each leaf node represents a class label
  - The paths from root to leaf represent classification rules.



- **Random Forest (RF):**
  - Random forest is an ensemble algorithm that aggregates a series of decision trees.
  - RF lowers the variances in predictions that decision trees otherwise have.

Model Construction

- Data transformation
  - The articles collected from ProQuest are turned into term-document matrices--where each column is a word and each row is a document.
  - From each article we also extract unigrams (individual words) and bigrams (pairs of consecutive words).
  - The unigrams and bigrams are then filtered:
    - Unigrams must show up in at least 40 documents
    - Bigrams must show up in at least 5 documents
  - We also perform the following cleaning procedures:
    - Stemming
    - White-space removal
    - Punctuation removal
    - Stop words removal

- We examined a total of six models for classification:
  - Two class + Unigram
  - Two class + Bigrams
  - Two Class + Mixed grams
  - Three class + unigram
  - Three class + Bigram
  - Three class + mixed grams

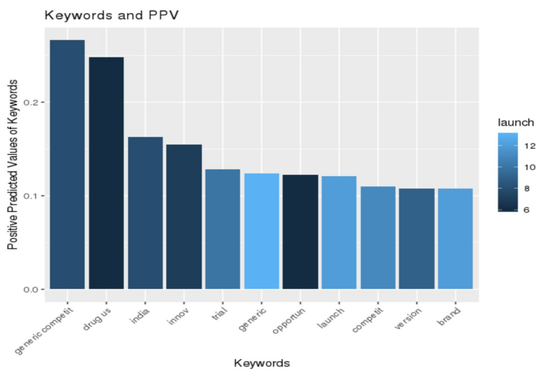
★ Two class classification--is this article about a **product launch** or **not**?
★ Three class classification--is this article about an **FDA approval, a product launch**, or **neither**?
★ Cleaned unigrams, bigrams, and mixed grams serve as predictors in the modeling process.

Model Performance and Evaluation

- Dictionary Approach Model Overview
  - We used Bayes Theorem to calculate the positive predicted value of a given word

$$\mathbb{P}(\text{Launch}|\text{Word}) = \frac{\mathbb{P}(\text{Word}|\text{Launch})\mathbb{P}(\text{Launch})}{\mathbb{P}(\text{Word}|\text{Not Launch})\mathbb{P}(\text{Not Launch}) + \mathbb{P}(\text{Word}|\text{Launch})\mathbb{P}(\text{Launch})}$$

  - We have found a series of words that are likely to be predictive of whether an article is about a launch:

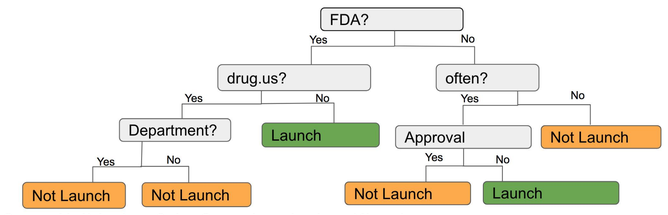


    - The words chosen above are determined by both PPV and discretion of the researchers.
  - We will filter the unlabeled articles by how many times these distinct words have showed up and validate our models.

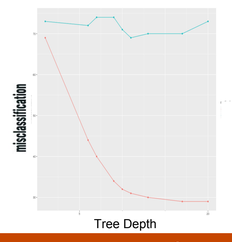
- Tree Modeling Performance Overview
  - **Accuracy rate**: total number of correct prediction/total Prediction.

Class/n-gram	Two-Class(about product/not about product)	Three-class(FDA approval/product launch/neither)
Unigram	92%	89%
Bigram	90%	91%
Mixed	95%	94%

- Decision Tree Diagram (Best performing tree--two class with mixed grams)



- Cross Validation & In-Sample misclassification rate



  - X axis represents tree dept, Y axis represents numbers of misclassified cases
  - As tree grows deeper, number of in-sample misclassification rate (orange) always decreases due to overfitting
  - Number of cross-validated misclass rates (blue) is lowest at 11.

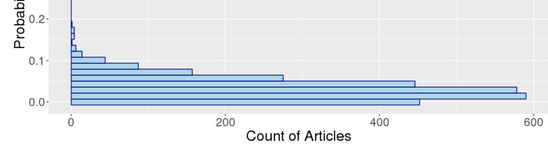
Model Validation and Application

- Dictionary Approach
  - Applying the dictionary to 2880 unlabeled articles, we have obtained the following table:

Number of keywords showing up in an article	0	1	2	3	4	5	6	7	8
Number of unlabeled articles	629	822	500	232	111	56	14	3	1

  - However, after students manually validated the articles with the highest keywords appearance (8 and 7 times), we found out that those articles aren't about a product launch.
  - This is due to the structure of the data:
    - PPV of the keywords is low, even for "generic competit", a 0.22 indicates that given it appears, there is a 22% the article is about launch.
    - We cannot aggregate the PPV of multiple words because we cannot assume their appearances are independent.

- Decision Tree & Random Forest
  - We assembled our previously trained decision trees and used random forest to compute the predicted probability of an article being about a launch.
  - The diagram shows the distribution of the probability of unlabeled articles being about product launch.



  - After manually validating the article with the highest probability, we found out that it is indeed about launch.
  - The machine learning models performed better and more reliably.

References

[1] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.

Tree graphics: https://shapeofdata.wordpress.com/2013/07/09/random-forests/

Challenge and Next Steps

- Challenges
  - Lack of training Data
    - Out of 3600 articles downloaded, only 484 are manually labeled.
  - Imbalance of training data
    - Out of the 484 labeled data, only 21 are about product launch. The imbalance in true positives makes it difficult for the model to detect signal from noise.
    - Without proper domain knowledge, the ability to manually select important keywords is difficult.
  - Tree and forests performed well, but the results aren't necessarily generalizable.

- Next steps
  - Label more articles.
  - Improve the quality of the labeling.
  - Improve domain expertise and repeat the iterative process of finding keywords.
  - Validate more unlabeled articles to check for the optimal threshold of the dictionary approach.
  - Validate using new data sources, e.g., Factiva.