# IntelliNews

An "Intelligent" Exploration of News Events and the Possible Associative Influence on the Stock Market

Name: Karnaz Obaidullah
Student # 501000900
Supervisor: Tamer Abdou, PhD

Ryerson University

# Table of Contents



Ryerson University

# Executive Summary

1. **Motivation and Purpose:**
    1. Deciphering news impact on financial markets
    2. Enabling event-driven investment strategies
2. **Methodology:**
    1. Simple Statistical Approach
    2. Semantics-based extraction
    3. LDA for thematic analysis
    4. Latent Semantic Analysis
    5. Sentiment analysis
    6. Applying Machine Learning
3. **Key Findings:**
    1. Semantics-based extraction challenges
    2. LDA insights on thematic structure
    3. Sentiment analysis aids market anticipation
    4. Text mining identifies crucial terms



All code files and figures are available on https://github.com/karnazko27/CIND-820-Big-Data-Analytics-Project

# Introduction

**1. Dataset:**

- Utilizing Global News dataset from Kaggle.
- Majority text data, collected via NewsAPI, termed unstructured.

**2. Methodology:**

- NLP and ML algorithms: Naïve Bayes, Clustering, Topic Modeling, etc.
- Objective: Classify news articles as events.

**3. Research Questions:**

- Impact of financial news on market prices.
- Criteria for classifying events.
- Standardization of text for ML/NLP algorithms.
- Role of event classification in investment strategies.
- Influence of events on market prices.

**Table 1: Research Questions Overview**

| Research Question | Tools and Techniques | Limitations |
|---|---|---|
| What is an Event? How do we classify an event from news articles?<br><br>Is it just something that has occurred or is it something substantial that has occurred and is repeated throughout different news articles? | Python libraries such as spaCy, NLTK, Gensim, scikit-learn | Event classification from News can be difficult since there is no clear-cut solution. |
| What makes News "*Event-worthy*"?<br>How do we differentiate between relevant and irrelevant news? For example, Apple beating its expected Q1 Profits will be relevant to Apple Inc. but not so much for a completely unrelated entity such as a Gold Mining Company like Barrick Gold Corporation. | Python libraries such as spaCy, NLTK, Gensim, scikit-learn | Determining worthiness of events is subjective. |
| What is the impact of events on market prices? What is the degree of effect of events on stock market?<br><br>How do financial news affect the intraday financial market prices of major stock market indices and select individual companies mentioned in the news? How does classifying events help us in formulating an event-driven investment strategy? | Python libraries such as pandas, spaCy, NLTK, Yahoo Finance API, statsmodels | Causality and Dependency are not the same. A News Event appearing the same day as a change in the market price may not mean causation.<br><br>We need to be able to differentiate between Causality and Dependency. |

Ryerson University

# Literature Review

# Literature Review

- **Event Studies:**
  - Application of Issue Attention Cycle framework (Downs, 1972) to financial news analysis.
  - Event windows defined for market reaction analysis.
  - Abnormal returns calculated to estimate event impact (MacKinlay, 1997).
  - Statistical significance assessed using two-tailed t-tests.
- **Distribution of Financial News:**
  - Strauss and Smith (2019) analyzed media source distribution.
  - Example: Distribution of financial news on new Tesla battery (Figure 1.1).
- **Intraday Event Study:**
  - Strauss and Smith's findings: Abnormal returns spiked post-event, declined over time.
  - Example: Intraday event study for Tesla shares (Figure 1.2).
- **Limitations:**
  - Follow-up reporting impact not considered due to resource constraints.
  - Internal communications' impact not analyzed, contrary to previous findings.

# Literature Review

**Figure 1.1: Distribution of financial news on new Tesla battery**

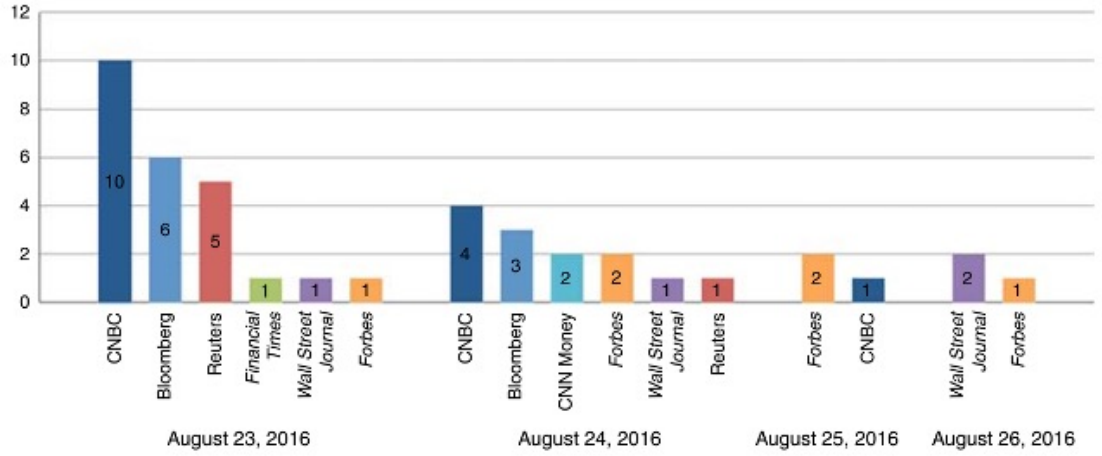



Fig 1.1: Distribution of financial news on new Tesla battery from day of announcement (August 23, 2016) until 3 days after (Figure 1 from Strauss and Smith (2019))

# Literature Review

**Figure 1.2: Intraday event study for Tesla shares**

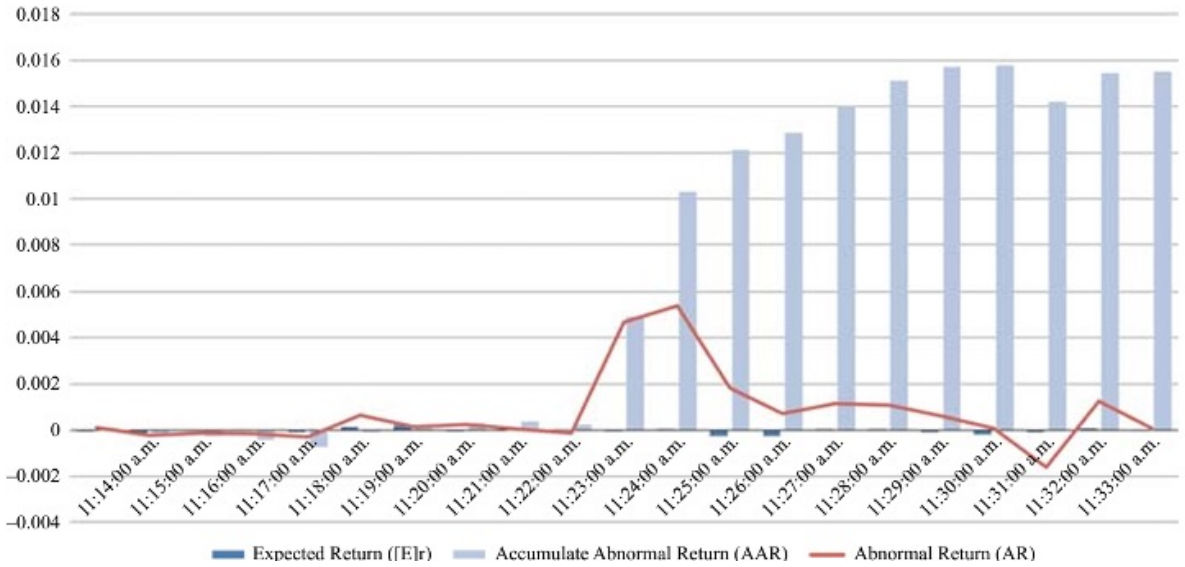



Fig 1.2: Intraday event study for Tesla shares at the moment when the tweet about the new product announcement released by Elon Musk on August 23, 2016 at 11:23 am EST (Figure 3 from Strauss and Smith (2019))

$$AR_{it} = R_{it} - E(R_{it} | X_t),$$

- $AR_{it}$ = *abnormal return for stock i at time point t (trading minute)*
- $R_{it}$ = *actual return*
- $E(R_{it} | X_t)$ = *expected return condition on market portfolio X at time point t*

# Data Description

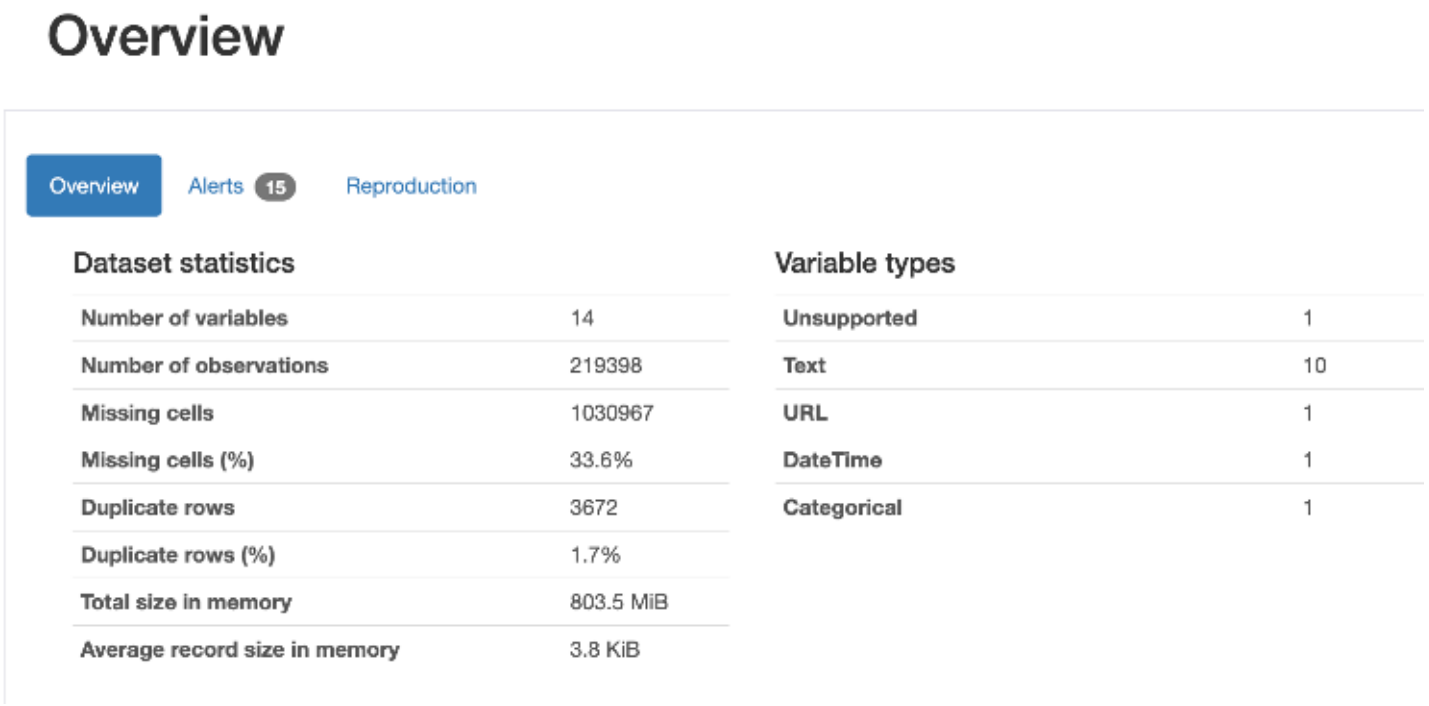
# Data Description

**Table 2: Features Overview**

| Feature Name | Data Type | Data Description |
|---|---|---|
| article_id | Integer | Unique Identification of Article |
| source_id | Text | Source identifier |
| source_name | Text | Source Name |
| author | Text | The author of the article |
| title | Text | The headline or title of the article |
| description | Text | A description or snippet from the article |
| url | Text | The direct URL to the article |
| url_to_image | Text | The URL to a relevant image for the article |
| published_at | Datetime | The date and time that the article was published in UTC timezone |
| content | Text | The unformatted content of the article, where available. This is truncated to 200 chars |
| category | Text | Category of News article |
| full_content | Text | Article Extracted from its respected URL |
| article | Text | Text of artticle |
| title_sentiment | Text | Sentiment of article |

# Data Description

**Fig 1.3: y-data profiling report OVERVIEW tab**

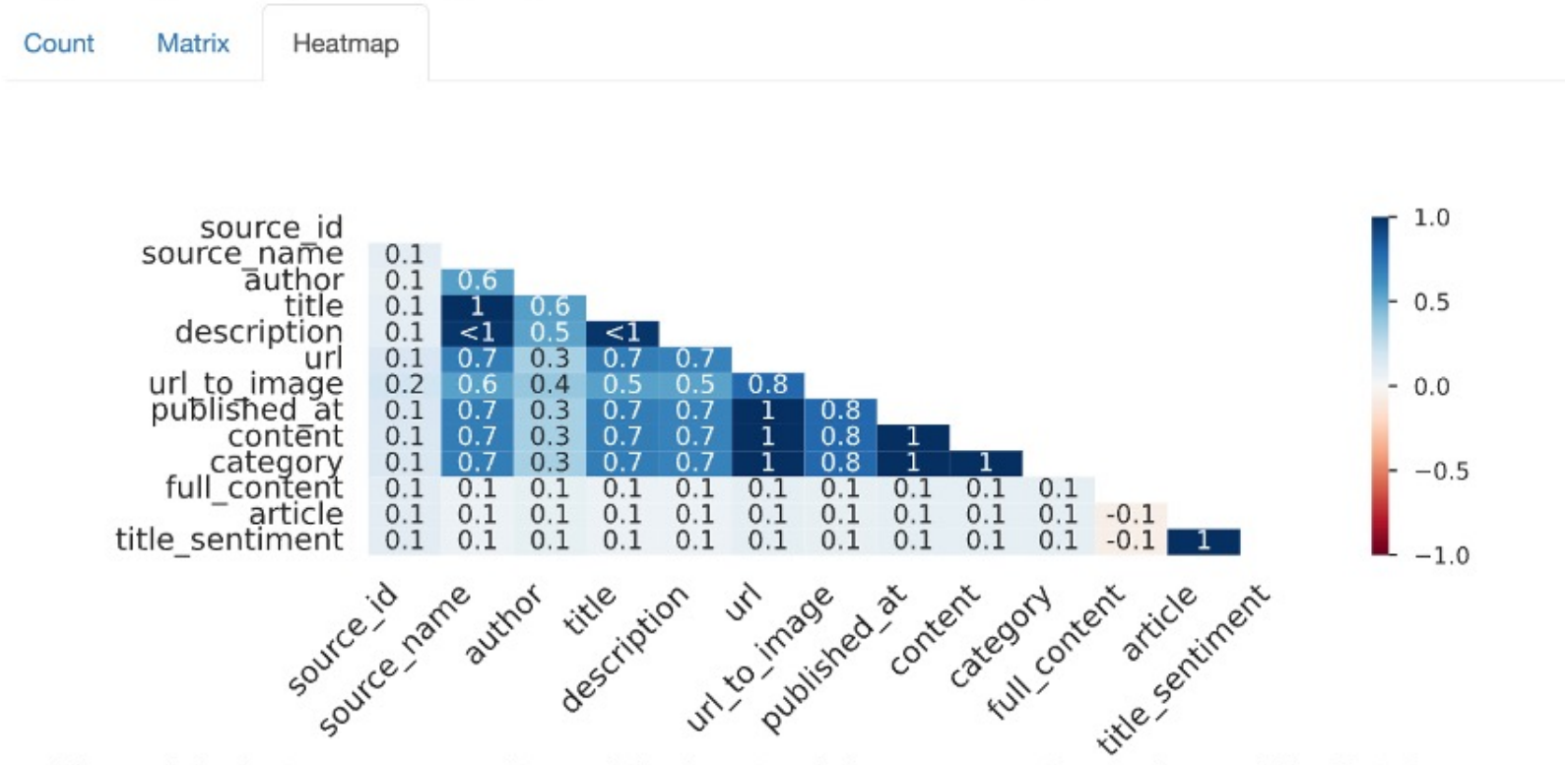


# Data Description

**Fig 1.4: y-data profiling report ALERT tab**

# Data Description

**Fig 1.8: y-data profiling report Correlation Heatmap**



# Data Description:
## Data Cleaning and Wrangling

Filer the master dataframe to only include the relevant columns ('Stock', 'Finance') we are interested in for our research. There are now 33,029 data instances and 14 features.

We perform the following data cleaning/wrangling tasks before applying machine learning techniques:

- Converting 'published_at' to pandas datetime object
- There are no duplicate articles looking at article_id column so there is no need to remove duplicates
- There are no rows containing all missing values
- Concatenate all text attributes into one feature 'all_text'. The text attributes have NaNs that appear as floats and need to be filtered and converted to string datatype.

# Approaches/Methodologies Used with Analysis and Findings

# Approaches/Methodologies Used with Analysis and Findings

**Method 1: Simple Text Mining using Statistical Approach**

**Method 2: Semantics-based Information Extraction**

**Method 3: Topic Modelling using Latent Dirichlet Allocation (LDA)**

**Method 4: Using Semantic Approach with Latent Semantic Analysis (LSA)**

**Method 5: Sentiment Analysis of News**

**Method 6: Applying Machine Learning to assess Possible Associative Impact on Dow Jones Index Price Change**

# Approaches/Methodologies Used with Analysis and Findings

**Method 1: Simple Text Mining using Statistical Approach**

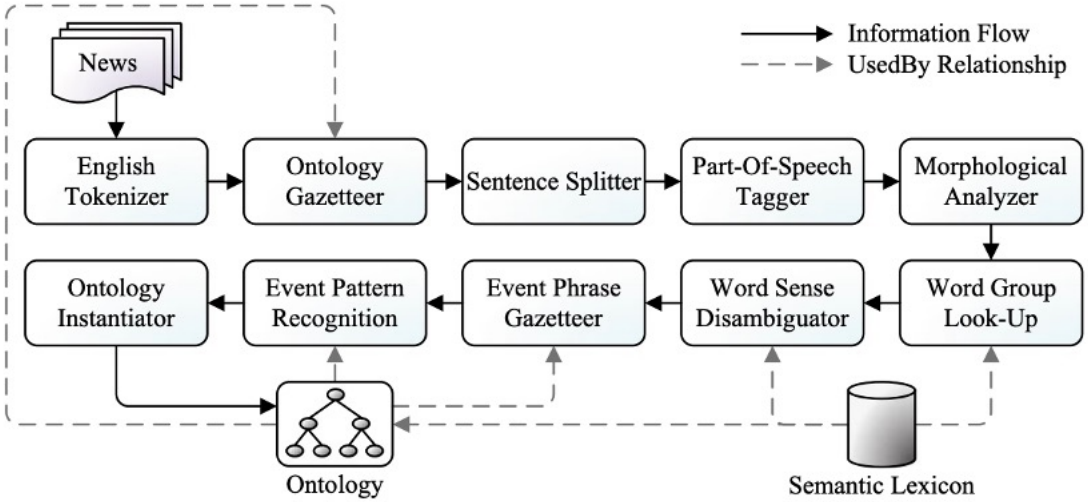
- **Theoretical Basis:**
  - Utilization of Unigrams, Bigrams, and N-grams.
  - Application of TF-IDF and Bag of Words models.
  - Goal: Identifying patterns, important terms, and relationships in text data.
- **Practical Findings:**
  - Foundational work conducted in notebook "2_Text_Analysis.ipynb".
  - Method served as a launching pad for subsequent approaches.
- **Limitations:**
  - Limited Scope: Potential oversight of nuanced semantic meanings and context.
  - Dimensionality Issues: Challenges in computational efficiency and interpretation with large datasets.
  - Dependency on Preprocessing: Effectiveness reliant on preprocessing steps, impacting result quality.
- **Conclusion:**
  - Despite limitations, statistical text mining approach lays groundwork for advanced analyses.
  - Foundation for extracting meaningful information from textual data.

# Approaches/Methodologies Used with Analysis and Findings

## Method 2: Semantics-based Information Extraction

- **Theoretical Basis:**
  - Hogenboom et al. (2013) proposed Semantics-Based Pipeline for Economic Event Detection (SPEED)
  - Focus on extracting financial events from news articles using semantic annotations
  - Pipeline comprises various steps including ontology-based gazetteer and event pattern recognition

**Fig 1.9: SPEED design (From Fig 1 Hogenboom et al., 2013)**



# Approaches/Methodologies Used with Analysis and Findings

**Method 2: Semantics-based Information Extraction**

- **Practical Findings:**
  - Pipeline tested on sample text to extract event phrases and patterns.
  - Result showed empty event phrases and patterns due to ontology limitations.
- **Limitations of Approach:**
  - Dependency on Ontology and Semantic Lexicons.
  - Limited Coverage of Events.
  - Difficulty Handling Noisy Text.
  - Scalability and Speed.
  - Limited Adaptability to New Domains.
  - Complexity of Event Pattern Recognition.
  - Evaluation on Limited Dataset.
  - Handling Ambiguity.
- **Addressing Limitations:**
  - Further research needed to improve data resources, algorithms, and evaluation methodologies.

# Approaches/Methodologies Used with Analysis and Findings

**Method 3: Topic Modelling using Latent Dirichlet Allocation (LDA)**

•**Theoretical Basis:**

- Utilization of Topic Modeling, specifically Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003).
- LDA represents documents as mixtures of topics, each characterized by a distribution over words.
- Preprocessing steps include lowercase conversion, noise removal, stop word removal, stemming, lemmatization, and normalization.

•**Performance Metrics:**

- PMI (Pointwise Mutual Information) used as a performance metric.
- UMass Coherence Score measures semantic similarity between words in the same topic.
- C_v Coherence Score evaluates coherence of topics by pairwise similarity between top words.

# Approaches/Methodologies Used with Analysis and Findings

**Method 3: Topic Modelling using Latent Dirichlet Allocation (LDA)**

```
[(90,
  '0.156*"group" + 0.031*"opened" + 0.025*"gap" + 0.025*"river" + 0.021*"l…" + '
  '0.018*"gapped" + 0.015*"peak" + 0.015*"prior" + 0.013*"ma" + '
  '0.012*"guidance"'),
 (77,
  '0.017*"chars" + 0.014*"time" + 0.013*"like" + 0.011*"best" + 0.009*"good" + '
  '0.009*"need" + 0.008*"start" + 0.008*"season" + 0.008*"money" + '
  '0.008*"market"'),
 (72,
  '0.049*"health" + 0.043*"union" + 0.023*"hunt" + 0.020*"battery" + '
  '0.016*"popular" + 0.012*"promised" + 0.011*"lagos" + 0.011*"kit" + '
  '0.011*"governor" + 0.010*"away"'),
 (65,
  '0.052*"rating" + 0.049*"report" + 0.041*"target" + 0.039*"price" + '
  '0.032*"research" + 0.028*"stock" + 0.026*"free" + 0.025*"buy" + 0.025*"—" + '
  '0.024*"reports"'),
 (21,
  '0.141*"des" + 0.036*"der" + 0.036*"die" + 0.031*"chars" + 0.025*"web3" + '
  '0.023*"cloud" + 0.022*"ag" + 0.017*"hong" + 0.015*"100" + 0.013*"kong"'),
 (71,
  '0.043*"—" + 0.042*"quarter" + 0.042*"free" + 0.041*"according" + '
  '0.040*"report" + 0.040*"recent" + 0.038*"shares" + 0.031*"holdings" + '
  '0.028*"filing" + 0.027*"llc"'),
 (50,
  '0.035*"chars" + 0.021*"car" + 0.021*"giant" + 0.019*"new" + 0.018*"deal" + '
  '0.017*"sales" + 0.015*"playstation" + 0.013*"year" + 0.013*"…" + '
  '0.013*"ps5"'),
 (9,
```

# Approaches/Methodologies Used with Analysis and Findings

**Method 3: Topic Modelling using Latent Dirichlet Allocation (LDA)**

- **Practical Findings:**
  - UMass Coherence Score: -0.30226957540121774
  - C_v Coherence Score: 0.9684235864472729
  - Interpretability and coherence of topics evaluated.
- **Limitations of Approach:**
  - Subjectivity in tuning model parameters and selecting optimal number of topics.
  - Difficulty in capturing semantic coherence across topics, especially in noisy or heterogeneous datasets.
  - Diverse topics may result, some of which may not be relevant, especially with larger topic numbers.

# Approaches/Methodologies Used with Analysis and Findings

**Method 4: Using Semantic Approach with Latent Semantic Analysis (LSA)**

- **Theoretical Basis:**
  - Latent Semantic Analysis (LSA) extracts meaning from text passages through statistical computations over document collections (Evangelopoulos, 2013).
  - Grounded in the Vector Space Model (VSM), representing documents as vectors in a term space.
  - Utilizes techniques like TF-IDF and log-entropy transformations to discount frequent terms and promote less frequent ones.
- **Practical Findings:**
  - Abstract results made interpretation challenging.

```
Terms: ['00' '000' '0000' ... '딥브레인ai' 'ｐｏｉｎｔ株式会社' 'ｓｔｏｃｋ']
LSA Matrix: [[ 5.70278071e-01  4.38536797e-01  1.81315096e-03 ...  5.261094
   4.55782413e-02 -3.40515535e-01]
 [ 3.97650354e-01  3.25930322e-01  2.91875106e-03 ...  6.96682871e-01
   1.09193974e-01 -3.08910055e-01]
 [ 6.80599283e-01  2.75654307e-01  1.55537404e-03 ...  2.28934688e-01
  -1.07239610e-01 -3.91214515e-01]
 ...
 [ 3.39249149e-01  1.77825012e-01  4.79131301e-03 ...  6.68427922e-01
   4.15634719e-02 -4.93134302e-01]
 [ 2.79680239e-01  3.81584504e-01  1.66657951e-03 ...  7.83010650e-01
```

# Approaches/Methodologies Used with Analysis and Findings

**Method 4: Using Semantic Approach with Latent Semantic Analysis (LSA)**

```
Cosine Similarity Matrix:
[[1.         0.9534782  0.85704471 ... 0.90671046 0.85991229 0.04027889]
 [0.9534782  1.         0.72648767 ... 0.96738391 0.93004733 0.02452458]
 [0.85704471 0.72648767 1.         ... 0.72463004 0.58136539 0.02854006]
 ...
 [0.90671046 0.96738391 0.72463004 ... 1.         0.87365836 0.03436503]
 [0.85991229 0.93004733 0.58136539 ... 0.87365836 1.         0.02245589]
 [0.04027889 0.02452458 0.02854006 ... 0.03436503 0.02245589 1.        ]]
```

•**Limitations of Approach:**

- Dimensionality Reduction: Loss of information due to dimensionality reduction using SVD.
- Bag-of-Words Representation: Ignores word order and context, limiting nuance capture.
- Semantic Ambiguity: May not distinguish between different word senses accurately.
- Noisy Text: Sparse matrices may lead to noisy representations.
- Scalability: Computationally expensive for large datasets.
- Interpretability: Semantic dimensions may be difficult to interpret, lacking direct correspondence to intuitive concepts.

# Approaches/Methodologies Used with Analysis and Findings

```
Naive Bayes Accuracy: 0.8613973182780522
              precision    recall  f1-score   support

           0       0.81      0.61      0.70      1846
           1       0.87      0.95      0.91      5239

    accuracy                           0.86      7085
   macro avg       0.84      0.78      0.80      7085
weighted avg       0.86      0.86      0.85      7085

Random Forest Accuracy: 0.9115031757233593
              precision    recall  f1-score   support

           0       0.94      0.70      0.81      1846
           1       0.90      0.98      0.94      5239

    accuracy                           0.91      7085
   macro avg       0.92      0.84      0.87      7085
weighted avg       0.91      0.91      0.91      7085

Logistic Regression Accuracy: 0.903316866619619
              precision    recall  f1-score   support

           0       0.88      0.72      0.80      1846
           1       0.91      0.97      0.94      5239

    accuracy                           0.90      7085
   macro avg       0.90      0.85      0.87      7085
weighted avg       0.90      0.90      0.90      7085

AdaBoost Accuracy: 0.8808750882145377
              precision    recall  f1-score   support

           0       0.82      0.69      0.75      1846
           1       0.90      0.95      0.92      5239

    accuracy                           0.88      7085
   macro avg       0.86      0.82      0.84      7085
weighted avg       0.88      0.88      0.88      7085
```

## Method 5: Sentiment Analysis of News

- **Theoretical Basis for Approach:**
  - Sentiment analysis predicts the impact of events on stock market prices using various machine learning algorithms (Lazrig & Humpherys, 2022).
  - Algorithms include Naïve Bayes, Random Forest, Logistic Regression, AdaBoost, and VADER.
- **Practical Findings:**
  - Naive Bayes: Accuracy of 0.86, precision of 0.81 (negative sentiment) and 0.87 (positive sentiment).
  - Random Forest: Accuracy of 0.91, balanced precision and recall for both sentiment classes.
  - Logistic Regression: Accuracy of 0.90, balanced precision and recall for both sentiment classes.
  - AdaBoost: Accuracy of 0.88, lower precision and recall for negative sentiment.
- **Limitations of Approach:**
  - Limited Feature Set
  - Subjectivity in Labeling
  - Assumption of Independence
  - Imbalanced Classes
  - Overfitting

# Approaches/Methodologies Used with Analysis and Findings

**Method 6: Applying Machine Learning to assess Possible Associative Impact on Dow Jones Index Price Change**

**Theoretical Basis for Approach:**

- **Efficient Market Hypothesis (EMH):** Asserts that asset prices reflect all available information, suggesting swift incorporation of news into stock prices.
- **Information Theory:** New information reduces uncertainty; news articles disseminate information, prompting reassessment of expectations and trading strategies.
- **Behavioral Finance:** Investors may not always act rationally; news articles can evoke emotional responses, influencing trading behavior.
- **Event Studies:** Analyze the impact of specific events on stock prices; machine learning assists in identifying relevant events and quantifying their impact.

# Approaches/Methodologies Used with Analysis and Findings

```
Training Linear Regression...

Linear Regression Results:
Mean Squared Error: 0.24481128909218303
Root Mean Squared Error: 0.4947840833052161
Mean Absolute Error: 0.17682357502018473
R-squared: 0.9999999974194661

Training Ridge Regression...

Ridge Regression Results:
Mean Squared Error: 25175.671646892224
Root Mean Squared Error: 158.6684330511026
Mean Absolute Error: 59.54692282845346
R-squared: 0.9997346255025521

Training Lasso Regression...
/usr/local/lib/python3.10/dist-packages/sklearn/line
  model = cd_fast.sparse_enet_coordinate_descent(

Lasso Regression Results:
Mean Squared Error: 13416.480194049083
Root Mean Squared Error: 115.82953075122545
Mean Absolute Error: 10.886683384214882
R-squared: 0.9998585780852661

Training Random Forest Regression...

Random Forest Regression Results:
Mean Squared Error: 7181.884186162952
Root Mean Squared Error: 84.74599805396684
Mean Absolute Error: 1.4387287678634435
R-squared: 0.9999242964027588
```

**Method 6: Applying Machine Learning to assess Possible Associative Impact on Dow Jones Index Price Change**

**Practical Findings:**

•Linear Regression: Exceptional performance with low MSE, RMSE, and MAE, and an almost perfect fit.

•Ridge Regression: Inferior predictive performance compared to Linear Regression, potentially due to regularization constraints.

•Lasso Regression: Moderate performance, reasonable accuracy, but slightly lower than Linear Regression and Random Forest.

•Random Forest Regression: Excellent predictive performance, remarkably low MSE, RMSE, and MAE, with outstanding fit to the data.

# Approaches/Methodologies Used with Analysis and Findings

**Method 6: Applying Machine Learning to assess Possible Associative Impact on Dow Jones Index Price Change**

**Limitations of Approach:**

- Data Quality and Bias
- Linguistic Complexity
- Limited Contextual Understanding
- Market Efficiency and Randomness
- Correlation vs. Causation
- Overfitting and Generalization
- Regulatory and Ethical Considerations

# Conclusion

# Conclusion

**Challenges:**

- Text analysis faces semantic ambiguity and subjective interpretation
- Dependency and causation assessment between news events and stock market movements is complex

**Reflection:**

- Assumptions and biases encountered in the project
- Challenges with text preprocessing and model overloading
- Realization that the project was a journey, not a destination

**Final Thoughts:**

- Exploration of the link between news articles and stock market
- Event extraction as a step to determine relevance of news
- Acknowledgment of the iterative nature of research and learning process

# Thank You!

Ryerson University