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

Executive Summary

The exponential growth of unstructured data on the internet has created new opportunities for leveraging Natural Language Processing (NLP) techniques in various domains. This project focuses on harnessing NLP and machine learning algorithms to classify news articles into relevant events within the context of financial markets and the broader economy.

Motivation and Purpose: The motivation behind this project stems from the need to decipher the impact of news events on financial market sentiment and share prices. By discerning meaningful events from noise, investors can potentially capitalize on market inefficiencies and formulate event-driven investment strategies. The project aims to address key research questions surrounding the classification of events, the impact of events on market prices, and the standardization of text for NLP algorithms.

Methodology: Several methodological approaches are employed, including semantics-based information extraction, topic modeling using Latent Dirichlet Allocation (LDA), sentiment analysis, and simple text mining techniques. Each approach offers unique insights into the analysis of financial news data and its implications for market sentiment and prices.

Key Findings:

- Semantics-based information extraction holds promise for event detection and annotation, but it is subject to limitations such as dependency on ontology and scalability issues.
- Topic modeling techniques like LDA provide valuable insights into the thematic structure of financial news, aiding in the identification of coherent topics.
- Sentiment analysis reveals the sentiment polarity of news articles, enabling investors to gauge market sentiment and anticipate potential market movements.
- Text mining techniques contribute to foundational text analysis, identifying important terms and relationships within financial news data.

Implications and Conclusion: By discerning meaningful events and sentiments from the noise in financial news data, investors can make more informed decisions and potentially gain a competitive edge in the financial markets. Ongoing research and innovation in text mining and NLP offer opportunities for further enhancing the understanding and utilization of textual data in financial analysis and decision-making. In conclusion, this project underscores the significance of leveraging NLP and machine learning techniques for extracting insights from financial news data, ultimately empowering investors to navigate the complexities of financial markets more effectively.

1. Project Abstract

The proliferation of unstructured data has boomed with the ubiquitous nature of the internet and out measures the availability of structured and semi-structured data on the internet. This presents new opportunities for the field of Natural Language Processing (NLP) which had been mostly in the shadows with the only visible use exhibited publicly in AI voice-generated robots.

Theme and Context

One such opportunity lies in text mining and analyses of myriad data sources. This project will delineate the utilization of Machine Learning Algorithms and Natural Language Processing (NLP) in the classification of news articles into relevant events in the context of financial markets and the greater economy. This falls into the realm of *Text Mining and Sentiment Analysis.* These actions aid in creating an information system or application by which the end-user can see categorized upcoming news as events quickly to implement a type of investment strategy called an event-driven strategy (Kenton, 2022)**.**

An event-driven investment strategy takes advantage of a temporary stock mispricing, which can occur before or after a corporate event occurs. The investor attempts to take advantage of the market mispricing before the information is fully incorporated into the market price as theorized by the efficient market hypothesis (Maverick, 2023). So, the question remains, what makes a particular piece of news an "event"? This does not have to be limited to events communicated by corporations only. This can and should encompass all the publicly information available that can influence the share price.

The purpose of this study is to research the effect news events have on financial market sentiment and the formation of share prices, as well as discovering the nature of "events" that possess market moving information by separating "events" from the noise. To conduct this analysis, we will use qualitative text analysis of online news and quantitative event studies of major stock market indices and select companies mentioned in the news.

Chosen Dataset

For this project, I have chosen the Global News dataset (Saksham, 2023) from Kaggle (https://www.kaggle.com/datasets/everydaycodings/global-news-dataset).

This is a collection of news scraped using the NewsAPI. Most of the data is text data in collections of documents which is referred to as unstructured data.

To achieve the objective of the project which is to classify news articles as *events*, using NLP and Machine Learning algorithms such as Naïve Bayes, Clustering, Topic Modeling, Event Extraction, Named Entity Recognition and Syntactic Dependency Parsing.

Research Questions

We will answer the following research questions in this paper:

1. How do financial news affect the intraday financial market prices of major stock market indices and select individual companies mentioned in the news?
2. How do we classify an event? Is it just something that has occurred or is it something substantial that has occurred and is repeated throughout different news articles? What makes something *event-worthy*?
3. Text has several different authors that have different forms of writing. How do we standardize text into a form that can be easily digested by Machine Learning and Natural Language Processing (NLP) algorithms?
4. How does classifying events help us in formulating an event-driven investment strategy?
5. How do events have an impact on market prices?

Table on Research Questions given in Next page.

Table 1: Research Questions Overview

Research Question	Tools and Techniques	Limitations
What is an Event? How do we classify an event from news articles? 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*"? 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? 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. We need to be able to differentiate between Causality and Dependency.

2. Literature Review

In Strauss and Smith's 2019 paper "*Buying on rumors: How financial news flows affect the share price of Tesla*", a similar study was conducted that study focused on the social media impact on the market of Elon Musk's tweets in addition to corporate communications and financial online news. This study is different from theirs in terms of scope as mine focuses on the wider financial market composed of major stock indices as well as some select individual companies only when they are mentioned in the news. The focus is on the nature of "events" and what makes a piece of news "event-worthy" in that it possesses market-moving information and then looking at the intraday market price movement. The financial markets are highly sensitive to breaking news on economic events like acquisitions, stock splits, etc. (Hogenboom et al., 2013).
There is a complex dynamic between event-driven news and its effect on stock market prices. Therefore, when analyzing business events and their effects on the stock market it is important to distinguish between sources of information (Strauss & Smith, 2019).

Event Studies

In Strauss and Smith's paper they argue that the Issue Attention Cycle theorized by Downs (Downs, 1972) can be applied to the realm of financial news to see if the market will react to the news. The Issue Attention cycle framework is made up of five stages: pre-problem state, alarmed discovery, and euphoric enthusiasm, realizing the costs of significant progress, gradual decline of public interest and the post-problem state (Downs, 1972). We will keep this in mind when we conduct the intraday event studies. Event studies are used to measure impact of events. In our event studies we will use news and stock market data.

Steps of Analysis

In conducting event studies, we look at the distribution of financial news according to the media sources, like the study by Strauss and Smith (2019) below:

Figure 1.1: Distribution of financial news on new Tesla battery

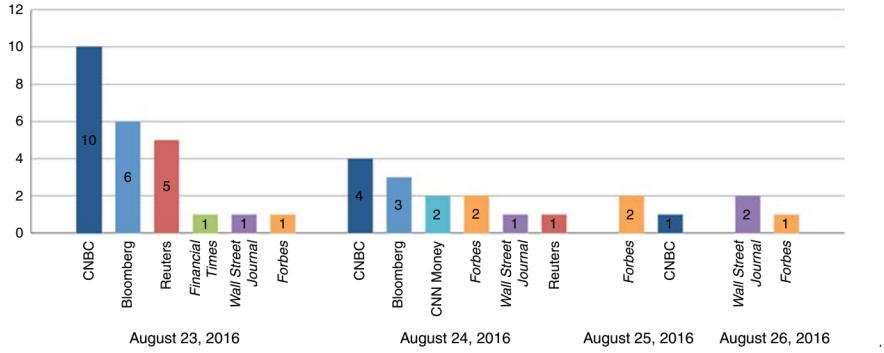


Fig 1.1: Distribution of financial news on new Tesla battery from day of announcement (August 23, 2016) until 3 days after (Figure 1 from Strauss and Smith (2019))

The second step is to define event windows that make up the periods in which security prices of the market are inspected for market reactions to the event (Strauss & Smith, 2019). Consequently, the abnormal returns of the share prices are calculated to estimate impact of events (MacKinlay, 1997):

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

- 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*

We can test whether the abnormal returns are statistically significant from the expected returns prior to the timestamp the news was published and after the news by conducting two-tailed t-tests (Strauss & Smith, 2019).

In the figure below we see that Strauss and Smith's (2019) findings were that the abnormal returns spiked for a particular event then petered off and subsequent events did not have the same degree of impact.

Figure 1.2: Intraday event study for Tesla shares

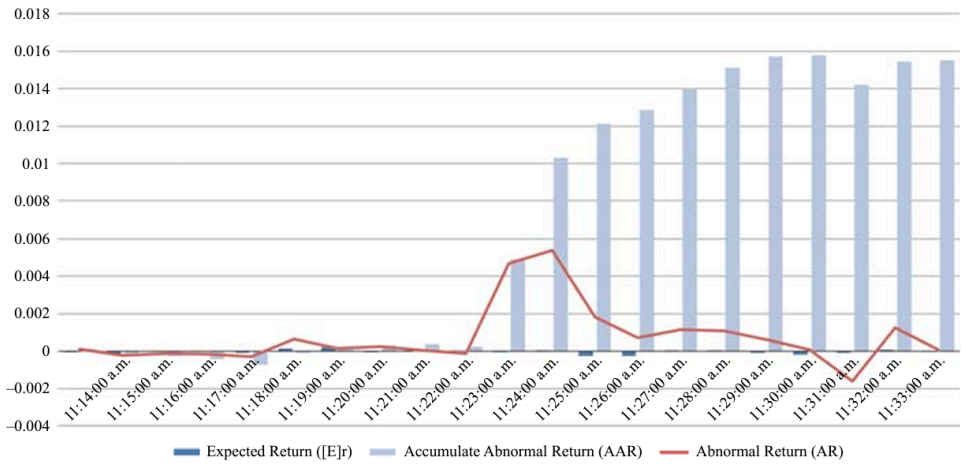


Fig 1.2: Intraday event study for Tesla shares at the moment when the tweet about the new product announcement released by Elon Musk on August 23, 2016 at 11:23 am EST (Figure 3 from Strauss and Smith (2019))

After the news drops there is follow-up reporting in the upcoming days which Strauss and Smith (2019) purported to have a positive impact in their case study of Tesla. Therefore, it may not be enough to look at simply intraday event studies but events and the market reactions over time. That is a limitation of our study which only focuses on intraday due to time and resource constraints. The other limitation of our study is that we only look at external news sources and not internal communications of companies as related market information. This goes against the findings of Strauss and Smith (2019) that it was internal communications (e.g. corporation Twitter, corporate press releases) that had an impact on the market rather than news articles reporting on the event.

3. Data Description[1]

The Global News dataset is composed of three sub-datasets: data.csv, rating.csv, and raw-data.csv. These datasets are merged into a master DataFrame to conduct the analysis for this project. This dataset contains news articles from September 30, 2023, to October 10, 2023 collection via NewsAPI.

This dataset appears to contain information related to various articles, including their titles, authors, publication sources, publication dates, URLs, and content summaries. Each article seems to pertain to different topics, as indicated by the category column. The dataset includes articles from diverse sources such as International Business Times, VOA News, The Indian Express, and The Times of Israel, suggesting a broad range of coverage. Additionally, there are missing values in certain columns like source_id and author, which might require further preprocessing if necessary for analysis.

The master dataset is made up of 1,096,988 data instances and 14 features. A brief description of the features is given below:

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
Feature Name (cont'd)	**Data Type (cont'd)**	**Data Description (cont'd)**
published_at	Datetime	The date and time that the article was published in UTC timezone
content	Text	The unformatted content of the article, where

[1] All relevant code and reports can be found on Github: https://github.com/karnazko27/CIND-820-Big-Data-Analytics-Project

		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

From the above table we can divide up the dataset according to the attribute type:

- Textual attributes are:
  - source_name, author, title, description, url, url_to_image, content, category, full_content, article, title_sentiment, source_id
- Non-textual attributes are:
  - article_id, published_at

Using y-data profiling python library we can get an EDA report of the dataset. Note that for this EDA report we used 20% of the dataset as a sample to prevent crashing the kernel.

The full report can be found here: https://github.com/karnazko27/CIND-820-Big-Data-Analytics-Project

The **Overview** section consists of an Overview tab of the dataset and its statistics, an Alert tab which provides information on the correlated variables and unique values, and the Reproduction tab which contains information about the report generation such as software version, start and end time of report generation, etc.

Fig 1.3: y-data profiling report OVERVIEW tab

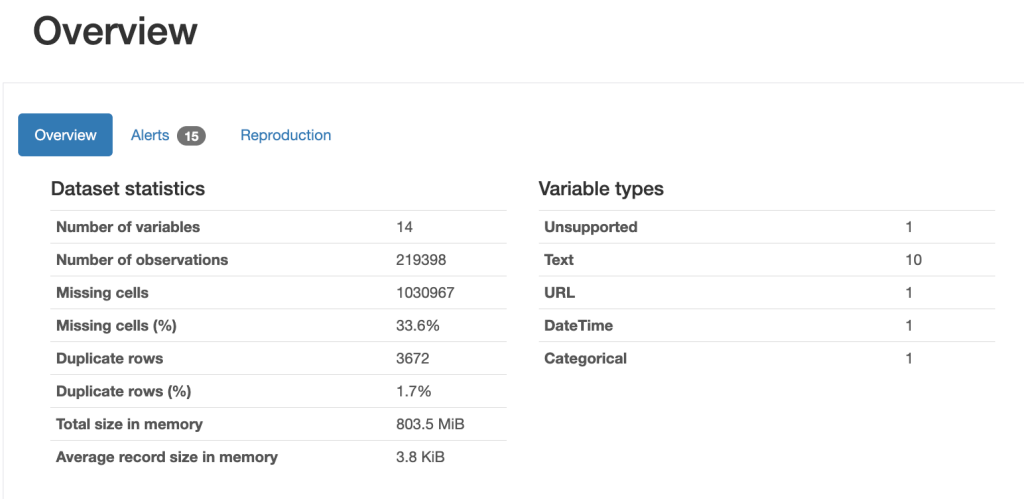


Fig 1.3: y-data profiling report OVERVIEW tab

Fig 1.4: y-data profiling report ALERT tab

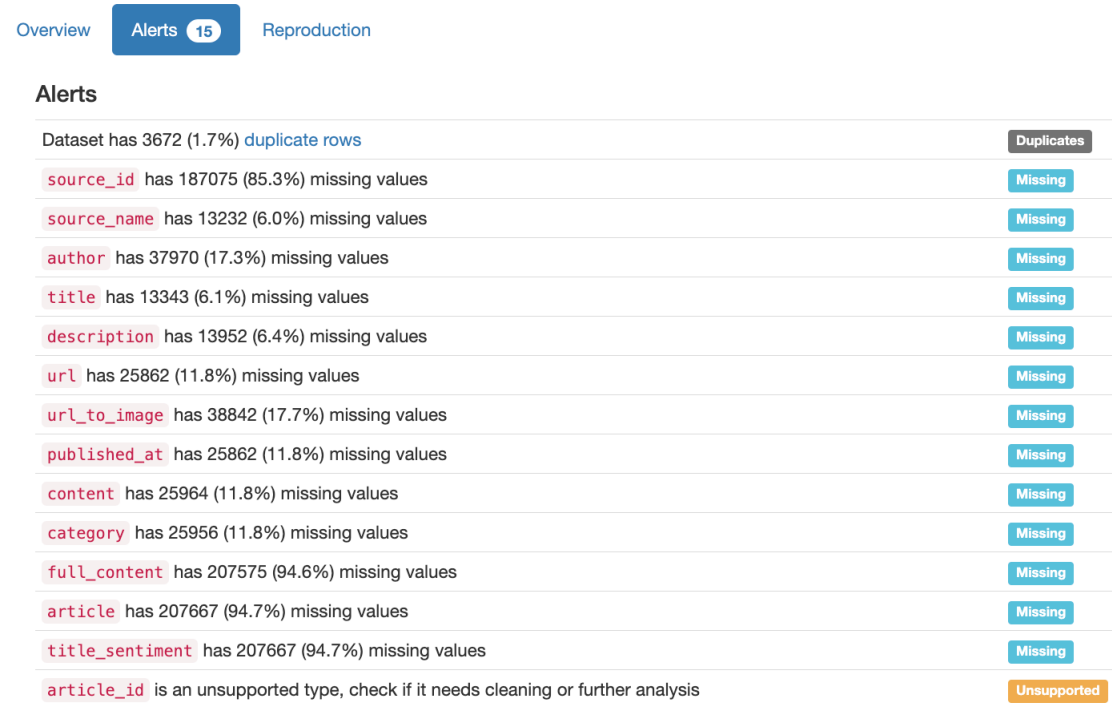


Fig 1.4: y-data profiling report ALERT tab

Fig 1.5: y-data profiling report REPRODUCTION tab

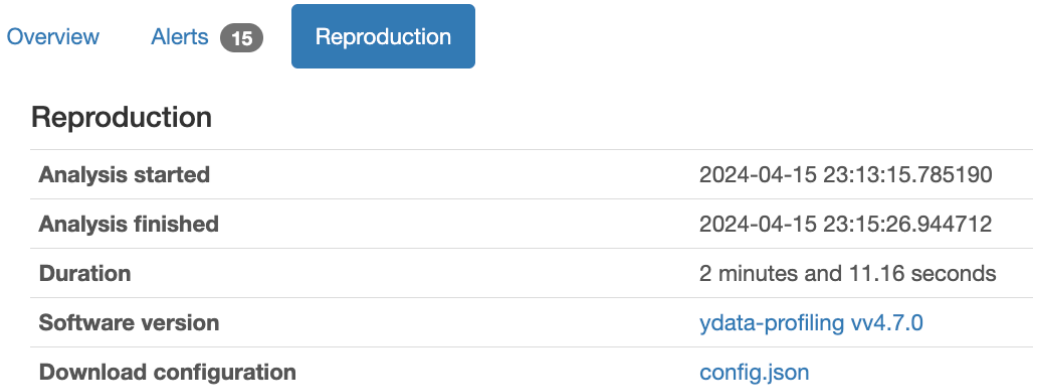
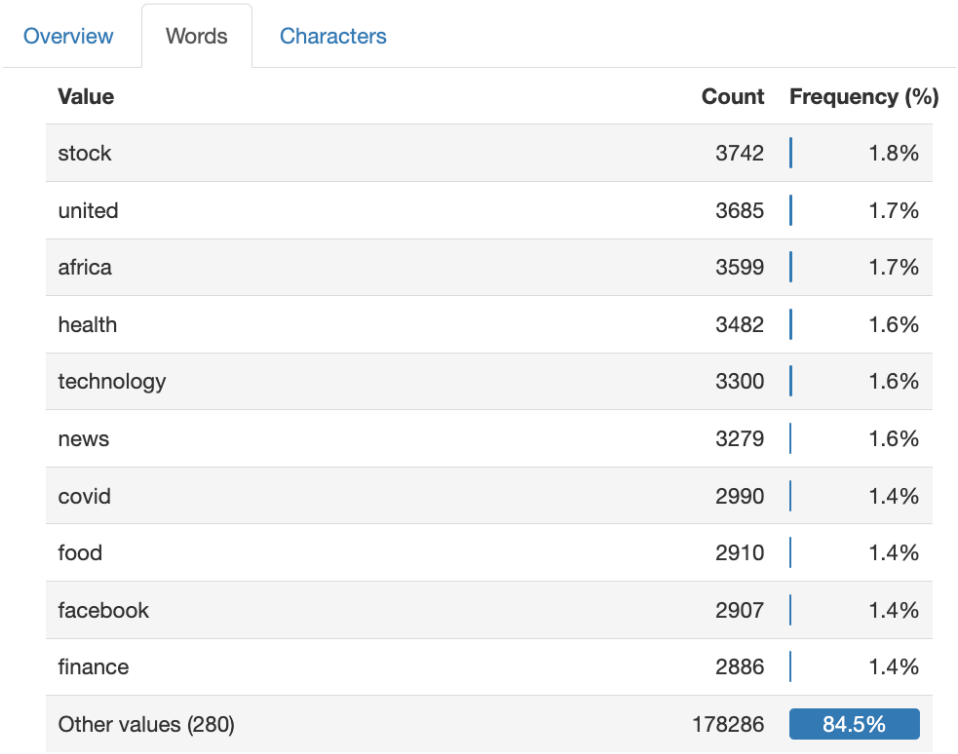


Fig 1.5: y-data profiling report REPRODUCTION tab

Exploring the full Profile Report generated we observe the following:

- This dataset has a fair number of missing values and duplicates.
- Features 'category' and 'title_sentiment' are categorical variables.
- The 'category' feature consists of 259 values. Of these categories for the purpose of our analysis we will focus on the 'Stock' and 'Finance' categories to focus on topics we are interested in.

Fig 1.6: y-data profiling report Variables section 'category' Words tab



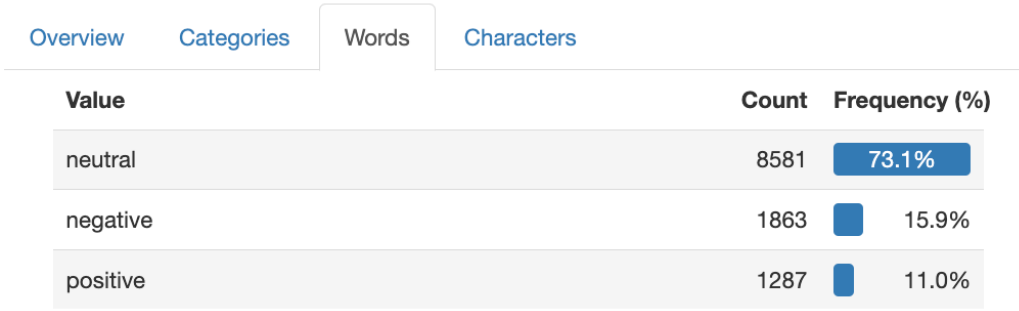
Overview | Words | Characters

Value	Count	Frequency (%)
stock	3742	1.8%
united	3685	1.7%
africa	3599	1.7%
health	3482	1.6%
technology	3300	1.6%
news	3279	1.6%
covid	2990	1.4%
food	2910	1.4%
facebook	2907	1.4%
finance	2886	1.4%
Other values (280)	178286	84.5%

Fig 1.6: y-data profiling report Variables section 'category' Words tab

- The 'title_sentiment' feature consists of neutral, negative, and positive sentiment:

Fig 1.7: y-data profiling report Variables section 'title_sentiment' Words tab



Overview | Categories | Words | Characters

Value	Count	Frequency (%)
neutral	8581	73.1%
negative	1863	15.9%
positive	1287	11.0%

Fig 1.7: y-data profiling report Variables section 'title_sentiment' Words tab

- The correlation heatmap indicates that the most highly correlated variables (>0.6) are: url and source_name, url and title, url and description, url_to_image and url, published_at and source_name, published_at and title, published_at and description, published_at and url, published_at and url_to_image, content and source_name, content and title, content and description, content and url, content and url_to_image, content and published_at, , category and title, category and description, category and url, category and url_to_image, category and

published_at, content and category, title and source_name, and lastly title_sentiment and article:

Fig 1.8: y-data profiling report Correlation Heatmap

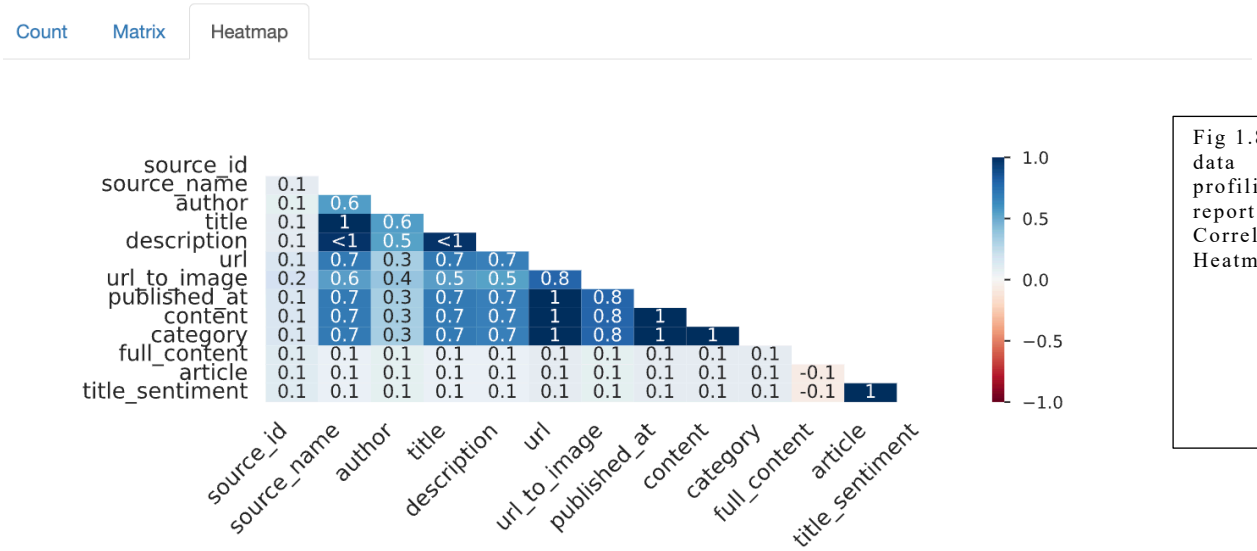


Fig 1.8: y-data profiling report Correlation Heatmap

Data Cleaning and Wrangling

We will filter 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.

4. Approaches/Methodologies Used with Analysis and Findings

When we are dealing with text data or unstructured data our analysis differs from numerical data in several ways, given in the table below:

Table 3: Key differences between Text and Numerical data

Key Aspect	Text Data	Numerical Data
Data Representation	Sequence of words, sentences, or documents	Numeric values arranged in rows and columns
Preprocessing Techniques	Will require tokenization, stemming, lemmatization, stop-word removal)	May require less preprocessing (normalization, scaling, encoding categorical variables)
Feature Engineering	Convert unstructured text into structured features (e.g., TF-IDF, word embeddings)	Feature engineering may involve normalization, scaling, encoding categorical variables
Data Visualization	Word clouds, frequency plots, topic modeling	Histograms, scatter plots, box plots
Purpose of Analysis	Understand distribution and pattern of words	Explore distribution and relationships between variables
Statistical Analysis	Word frequencies, distributions, correlations between terms using techniques like cosine similarity, Pearson correlation on TF-IDF	Descriptive statistics, correlation analysis between numeric variables
NLP Techniques	Sentiment analysis, named entity recognition, part-of-speech tagging, text classification	Regression, classification, clustering
Dimensionality Reduction	PCA, t-SNE on TF-IDF vectors or embeddings	PCA, t-SNE
Domain-Specific Considerations	Dealing with misspellings, slang, abbreviations, language-specific nuances	May have domain-specific considerations but generally simpler than text processing

We can use several approaches to answer our research questions.

Method 1: Simple Text Mining using Statistical Approach

Theoretical Basis for Approach
We use extract Unigrams, Bigrams and N-grams from text. Then we use Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words model to analyze text data. In this approach the goal is to identify and quantify patterns within text, to find important terms and relationships.

This method is good for foundational work when conducting more advanced text analysis.

Practical Findings
(Please note full code can be found in my github repo for this project under notebooks/2_Text_Analysis.ipynb)

The complete results from this approach is in the notebook. This method was a launching pad for the other methods.

Limitations of Approach
While this approach offers valuable insights into text data, it is not without limitations. Some of the constraints include:

- Limited Scope: The approach may overlook nuanced semantic meanings and context, particularly in complex texts.
- Dimensionality Issues: With large datasets, the high dimensionality of feature spaces may pose challenges in computational efficiency and interpretation.
- Dependency on Preprocessing: The effectiveness of the approach heavily relies on preprocessing steps such as tokenization, stemming, and stop-word removal, which may impact the quality of results.

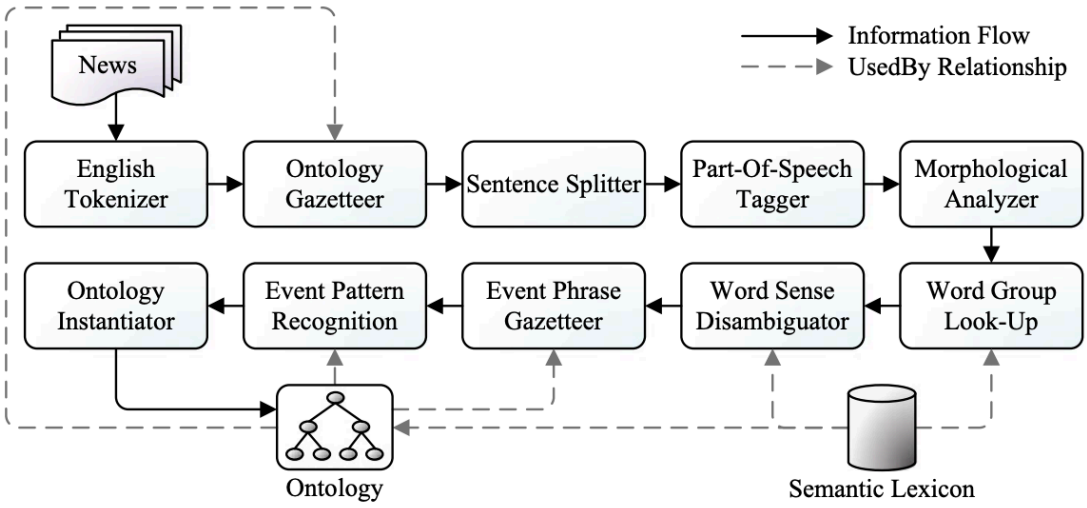
Despite these limitations, the statistical approach to text mining serves as a fundamental methodology for extracting meaningful information from textual data, laying the groundwork for further advanced analyses and applications.

Method 2: Semantics-based Information Extraction

Theoretical Basis for Approach
In Hogenboom et al.'s 2013 paper, *Semantics-based information extraction for detecting economic events*, they propose extracting financial events from news articles using the Semantics-Based Pipeline for Economic Event Detection or SPEED framework (or

pipeline) (Hogenboom et al., 2013). This pipeline follows the process in the diagram below:

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



This method focuses on extracting financial news events from news articles and annotating them with meta-data at a fast pace for real-time use. It is important to capture events that are represented in a machine-understandable way. To identify concepts, lexical representations of concepts retrieved from text are matched with event-related concepts that are available in WordNet (Hogenboom et al., 2013).
To start off the pipeline a domain ontology needs to be defined. Domain ontology consists of concepts related to our domain which is financial markets (Hogenboom et al., 2013).

I defined a simple domain ontology with just the most common finance terms. This can be expanded with full definitions and relationships in the future.
To test that the efficacy of the SPEED pipeline I ran each step on a small sample of text first:

Moderna Beats Sales Forecasts, But Light Guidance Could Rattle Shares The company issued below-consensus sales views for 2023 and 2024. Moderna (MRNA) stock could take a hit Thursday after the Covid vaccine maker handily beat third-quarter sales forecasts, but reported bigger-than-expected losses and issued light guidance for this ye… [+4041 chars]

Summary of Pipeline steps:

1. English Tokenizer: tokenize text documents into components or words (tokens)
2. Ontology Gazetteer: link concepts in texts to concepts found in domain ontology.
3. Sentence Splitter: groups tokens in the text into sentences
4. Part-of-Speech Tagger: each word is tagged with its part-of-speech

5. Morphological Analyzer: tokens are lemmatized (i.e. decomposed to root form based on dictionary meaning)
6. Word Group Look-Up: use WordNet semantic lexicon to identify word groups
7. Word Sense Disambiguator: determine word sense of each word group
8. Event Phrase Gazetteer: link word groups to an ontology.
9. Event Pattern Recognition: combine event with surrounding action words to add meaning and extract event phrases
10. Ontology Instantiator: add knowledge gained to domain ontology

Practical Findings

(Please note full code can be found in my github repo for this project under notebooks/3_Semantic_Approach_and_Information_Retrieval.ipynb)
We get the following result using the above pipeline:

Fig 1.10: result from SPEED pipeline on sample text

```
Tokens: ['Moderna', 'Beats', 'Sales', 'Forecasts', ',', 'But',
Linked Tokens: ['Moderna', 'Beats', 'Sales', 'Forecasts', ',',
Sentences: ['Moderna Beats Sales Forecasts, But Light Guidance
POS Tags: [('Moderna', 'NNP'), ('Beats', 'NNP'), ('Sales', 'NN
Lemmatized Tokens: ['Moderna', 'Beats', 'Sales', 'Forecasts',
Word Groups: {'Moderna': [], 'Beats': ['beat_generation.n.01',
Disambiguated Groups: {'Moderna': None, 'Beats': 'beat_generat
Event Phrases: []
Event Patterns: []
```

Now we see that event phrases and patterns are empty. This is because we need a rich ontology to find event phrases effectively. This could be a potentially interesting area to expand on.

Limitations of Approach

There aren't well-defined performance metrics for this approach. The approach described in the text is comprehensive and aims to tackle event detection and annotation in news articles effectively. However, there are several limitations to consider:

1. **Dependency on Ontology and Semantic Lexicons**: The effectiveness of the pipeline heavily relies on the availability and accuracy of the ontology and semantic lexicons (such as WordNet). If these resources are incomplete or inaccurate, it may lead to errors in event detection and annotation.

2. **Limited Coverage of Events**: The pipeline's performance is limited by the coverage of events included in the ontology. If certain types of events are not represented in the ontology, the pipeline may fail to detect or properly annotate them.

3. **Difficulty Handling Noisy Text**: Natural language text, especially news articles, can contain noise, such as misspellings, jargon, and ambiguous language. While the pipeline attempts to handle noisy linguistic information, it may still encounter challenges in accurately processing such text.

4. **Scalability and Speed**: While the pipeline aims for real-time use, the scalability and speed of the pipeline may become issues when processing large volumes of news articles. As the volume of data increases, the processing time may also increase, potentially affecting real-time performance.

5. **Limited Adaptability to New Domains**: While the pipeline is designed to be generalizable to other domains, adapting it to new domains may require significant effort in creating or modifying the ontology and semantic lexicons to suit the specific domain's terminology and concepts.

6. **Complexity of Event Pattern Recognition**: The event pattern recognition component relies on manually defined patterns (using JAPE rules). Developing and maintaining these rules can be labor-intensive and may require expertise in both the domain and natural language processing.

7. **Evaluation on Limited Dataset**: The evaluation of the pipeline's performance is based on a relatively small dataset of 200 news messages. While efforts have been made to ensure inter-annotator agreement and compare against existing approaches, the generalizability of the results may be limited.

8. **Handling Ambiguity**: The pipeline may struggle with handling ambiguous language or ambiguous word senses, especially in cases where multiple senses of a word are relevant to different contexts.

Addressing these limitations would require further research and development, potentially involving improvements in data resources, algorithms, and evaluation methodologies.

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

Theoretical Basis for Approach

This approach to extracting events from news is to use Topic Modelling techniques such as Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003).

Latent Dirichlet Allocation (LDA) is a widely used generative probabilistic model for analyzing collections of discrete data, such as text corpora (Blei, Ng, & Jordan, 2003). In LDA, each document in a collection is represented as a mixture of topics, and each topic is characterized by a distribution over words. The model assumes that documents are

generated in a three-level hierarchical process, where each word in a document is drawn from one of the topics, and the choice of topic is influenced by the document's underlying topic distribution.

LDA has been applied in various domains, including document modeling, text classification, and collaborative filtering. It offers a flexible framework for understanding the thematic structure of text data and has been instrumental in uncovering hidden patterns and topics within large document collections.
Topic Modelling techniques are unsupervised machine learning algorithms. Text needs to be preprocessed before applying such techniques (Kinariwala & Deshmukh, 2023) as follows:

1. Lowercase conversion: convert all tokens (words) to lowercase
2. Noise removal: remove extra digits or punctuations
3. Stop word removal: remove frequently occurring words like prepositions
4. Stemming: convert text to base form
5. Lemmatization: bring text to root dictionary form
6. Normalization: convert text to standard form

Performance metrics we can use for Topic Modelling are Topic coherence measures or Pointwise Mutual Information (PMI) (Kinariwala & Deshmukh, 2023).

PMI:

Fig 1.11: equation from Kinariwala & Deshmukh, 2023, *Short text topic modelling using local and global word-context semantic correlation*

$$PMI\left(wo_i, wo_j\right) = log\frac{p\left(wo_i, wo_j\right)}{p(wo_i)p\left(wo_j\right)}$$

where

- *wo_i and wo_j are word probabilities that co-occur in the document*
- *p denotes probability*

Fig 1.12: result from ChatGPT, Query: "Topic Coherence Formulas"

1. **UMass Coherence Measure**:
   The UMass coherence measure is based on the pairwise document co-occurrence within the same topic.
   The formula for UMass coherence is:
   $$\text{Coherence} = \frac{\sum_{i<j} \log \frac{D(f(w_i,w_j))+1}{D(f(w_i))}}{\binom{N}{2}}$$
   Where:
   - $D(f(w_i, w_j))$ is the number of documents containing both words w_i and w_j.
   - $D(f(w_i))$ is the number of documents containing word w_i.
   - N is the total number of words in the corpus.
   - The sum is over all distinct word pairs (w_i, w_j) in the topic.
2. **c_v Coherence Measure**:
   The c_v coherence measure computes the coherence based on the pointwise mutual information (PMI) between pairs of words in the same topic.
   The formula for c_v coherence is:
   $$\text{Coherence} = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j<i} \text{PMI}(w_i, w_j)$$
   Where:
   - k is the number of top words chosen from each topic.
   - w_i and w_j are two words from the topic.
   - $\text{PMI}(w_i, w_j)$ is the pointwise mutual information between words w_i and w_j in the corpus. It measures the association strength between the words and is defined as $\log \frac{p(w_i,w_j)}{p(w_i) \cdot p(w_j)}$, where $p(w_i, w_j)$ is the probability of co-occurrence of words w_i and w_j, and $p(w_i)$ and $p(w_j)$ are the probabilities of individual occurrences of words w_i and w_j, respectively.

Practical Findings

(Please note full code can be found in my github repo for this project under notebooks/3_Semantic_Approach_and_Information_Retrieval.ipynb)

For our analysis we got the following score for our performance metrics:

UMass Coherence Score: -0.30226957540121774
C_v Coherence Score: 0.9684235864472729

The coherence scores provide a measure of the interpretability and coherence of the topics generated by the Latent Dirichlet Allocation (LDA) model.

1. **UMass Coherence Score**: The UMass coherence score measures the degree of semantic similarity between words within the same topic based on their co-occurrence in the corpus. A negative UMass coherence score indicates that the

topics are not very coherent or interpretable. In this case, the score of approximately -0.30 suggests that the topics may lack coherence and may not adequately represent distinct themes or concepts.

2. **C_v Coherence Score**: The C_v coherence score, on the other hand, evaluates the coherence of topics by computing the pairwise similarity between the top words selected from each topic. A higher C_v coherence score indicates that the topics are more coherent and interpretable. Here, the score of approximately 0.97 suggests that the topics have a high degree of coherence and represent distinct themes or concepts effectively.

Therefore, while the UMass coherence score indicates that the topics may lack coherence, the high C_v coherence score suggests that the topics are coherent and represent meaningful themes or concepts.

Limitations of Approach

Despite its effectiveness, the approach has limitations. One such limitation is the challenge of tuning model parameters and selecting the optimal number of topics, which can be subjective and may require domain expertise. Additionally, topic models may struggle with capturing semantic coherence across topics, especially in noisy or heterogeneous datasets. Our topics result was diverse and not all of them were relevant so when given larger amounts of topics the resulting topics might become more diverse.

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

Theoretical Basis for Approach

Latent Semantic Analysis (LSA), is a method for extracting meaning from passages of text, based on statistical computations over a collection of documents (Evangelopoulos, 2013).

Latent Semantic Analysis (LSA) is grounded in the Vector Space Model (VSM), a mathematical framework for representing documents as vectors in a space where dictionary terms serve as dimensions (Evangelopoulos, 2013). This model facilitates the representation of a collection of documents as a term-by-document matrix, where the dimensionality of the matrix is determined through term reduction techniques like term filtering and term conflation. By transforming the initial frequency counts of terms in documents, using techniques such as TF-IDF and log-entropy transformations, LSA aims to discount the occurrence of frequent terms and promote less frequent ones.

Through the quantification of documents as vectors in the term space and terms as vectors in the document space, LSA enables the calculation of similarities between terms and documents. The cosine similarity metric is commonly used to quantify the

similarity between two vectors, allowing for the comparison of documents based on their term representations.

However, LSA goes beyond literal term matches by considering terms that are related to query terms, including synonyms and conceptually similar terms, through a statistical analysis of term usage patterns observed throughout the corpus. This approach, introduced in the late 1980s, relies on singular value decomposition (SVD) to decompose the term frequency matrix into term and document eigenvectors, capturing the latent semantic structure of the text data.

Statistical measure such as cosine similarity and Euclidean distance can be used in this approach to assess similarity or dissimilarity between documents.

Practical Findings
(Please note full code can be found in my github repo for this project under notebooks/3_Semantic_Approach_and_Information_Retrieval.ipynb)

After applying this technique, the results were quite abstract so this was difficult to understand.

Limitations of Approach
The Latent Semantic Analysis (LSA) method, while effective in capturing latent semantic relationships in text data, has several limitations:

1. **Dimensionality Reduction**: LSA relies on reducing the dimensionality of the term-document matrix using singular value decomposition (SVD). While this helps in capturing the most important latent semantic dimensions, it may result in loss of information, especially in very high-dimensional spaces.
2. **Bag-of-Words Representation**: LSA treats documents as bags of words and ignores word order and context. This can limit its ability to capture nuances in meaning and context, especially in languages with complex syntax and semantics.
3. **Semantic Ambiguity**: It may not distinguish between different senses of a word or identify synonyms accurately, leading to less precise semantic representations.
4. **Noisy Text**: Sparse matrices can lead to noisy representations and reduced performance in capturing meaningful relationships between terms and documents.
5. **Scalability**: LSA can be computationally expensive for very large datasets, especially during the matrix factorization step. Processing large term-document matrices may require substantial memory and computational resources.
6. **Interpretability**: Hard to Interpret. While LSA provides latent semantic representations of terms and documents, interpreting the meaning of the derived dimensions can be challenging. The semantic dimensions may not always correspond directly to intuitive concepts or categories.

Method 5: Sentiment Analysis of News

Theoretical Basis for Approach

Sentiment analysis of news articles is done to predict impact of events on stock market prices. This involves using several Machine learning algorithms to find sentiment of news articles (Lazrig & Humpherys, 2022).

The algorithms used for this approach are Naïve Bayes, Random Forest, Logistic Regression, AdaBoost, VADER (Valence Aware Dictionary and sEntiment Reasoner).

1. Naive Bayes: calculates the probability of a particular sentiment based on the occurrence of words in the text. It assumes independence among features, making it computationally efficient and often effective for text classification tasks.
2. Random Forest: is an ensemble learning method combines multiple decision trees to enhance predictive accuracy. Random Forest is known for handling noisy data and providing robust results in sentiment analysis tasks.
3. Logistic Regression: is a classification algorithm suitable for binary outcomes. It models the probability of a sentiment class, making it widely applied in sentiment analysis tasks.
4. AdaBoost: is a boosting algorithm that combines weak learners to create a strong learner. AdaBoost iteratively adjusts the weights of misclassified instances, improving the overall accuracy of sentiment predictions.
5. VADER (Valence Aware Dictionary and sEntiment Reasoner): is a lexicon and rule-based sentiment analysis tool designed for social media text, found in the Natural Language Toolkit (NLTK) python library.

The results from these algorithms and functions will help us in evaluating sentiment of news articles, enabling us to answer questions more regarding the potential impact of events on stock market prices.

Practical Findings

(Please note full code can be found in my github repo for this project under notebooks/ 4_Sentiment_Analysis.ipynb)

Fig 1.13: result from Notebook 4_Sentiment_Analysis_and_ML.ipynb

```
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
```

The practical findings from the analysis of sentiment using Naive Bayes, Random Forest, Logistic Regression, and AdaBoost algorithms are as follows:

1. **Naive Bayes**: The accuracy achieved by the Naive Bayes algorithm is 0.86. It performs reasonably well in predicting sentiment, with a precision of 0.81 for class 0 (negative sentiment) and 0.87 for class 1 (positive sentiment). However, it has lower recall for class 0 (0.61) compared to class 1 (0.95), indicating that it may miss some instances of negative sentiment.
2. **Random Forest**: The Random Forest algorithm achieves an accuracy of 0.91, outperforming Naive Bayes. It exhibits high precision for both classes, with 0.94 for class 0 and 0.90 for class 1. The recall values are also balanced, with 0.70 for

class 0 and 0.98 for class 1, indicating that it effectively captures instances of both negative and positive sentiment.

3. **Logistic Regression**: Logistic Regression achieves an accuracy of 0.90, performing slightly worse than Random Forest but better than Naive Bayes. It exhibits balanced precision and recall values for both classes, similar to Random Forest, indicating its effectiveness in capturing sentiment.
4. **AdaBoost**: AdaBoost achieves an accuracy of 0.88, performing slightly worse than Logistic Regression. It exhibits lower precision and recall for class 0 compared to class 1, indicating that it may struggle to capture instances of negative sentiment as effectively as positive sentiment.

Limitations of Approach

While the machine learning algorithms employed in this approach yield relatively high accuracy in predicting sentiment, there are several limitations to consider:

1. **Limited Feature Set**: The sentiment analysis is based solely on the text content of news articles, without considering other contextual factors that may influence sentiment, such as market trends, geopolitical events, or broader economic indicators.
2. **Subjectivity in Labeling**: The sentiment labels assigned to news articles may be subjective and open to interpretation, potentially leading to inconsistencies in the training data and bias in the model predictions.
3. **Assumption of Independence**: The Naive Bayes algorithm assumes independence among features, which may not hold true in the context of text data where words are often correlated. This could limit its effectiveness in capturing complex relationships between words and sentiment.
4. **Imbalanced Classes**: The distribution of sentiment classes (positive vs. negative) may be imbalanced in the dataset, which can impact the performance of the machine learning algorithms and lead to biased predictions.
5. **Overfitting**: Complex models such as Random Forest and AdaBoost may be prone to overfitting, especially if the training data is limited or noisy. This could result in poor generalization performance on unseen data.

Overall, while the machine learning algorithms provide valuable insights into the sentiment of news articles, it's essential to acknowledge these limitations and interpret the results with caution, considering the context and potential biases inherent in the data and model.

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

Theoretical Basis for Approach

The utilization of machine learning techniques to analyze news articles and their connection to stock market prices is grounded in several theoretical concepts within finance and economics.

One fundamental concept is the Efficient Market Hypothesis (EMH), which posits that asset prices fully reflect all available information. If news articles contain relevant and actionable information regarding a company's performance, strategic decisions, or market conditions, it should be swiftly incorporated into stock prices. Machine learning models offer a means to efficiently process large volumes of textual data from news articles, identifying patterns and sentiment that may influence market behavior. Information theory suggests that new information reduces uncertainty. News articles disseminate new information about companies, industries, or economic indicators, prompting investors to reassess their expectations and trading strategies. Machine learning algorithms, through natural language processing (NLP) and sentiment analysis, can parse news articles to extract key information and assess its potential impact on stock prices.

Behavioral finance recognizes that investors may not always act rationally and can be influenced by cognitive biases and emotions. News articles can evoke emotional responses among investors, leading to irrational trading behavior and subsequent stock price movements. By employing sentiment analysis and other NLP techniques, machine learning models can gauge market sentiment and anticipate potential investor reactions to news events.

Event studies analyze the impact of specific events, such as corporate announcements or economic indicators, on stock prices. Machine learning can assist in identifying relevant events from news articles and quantifying their impact on stock prices through abnormal return analysis. This allows researchers and market participants to discern the causal relationship between news events and market dynamics.
In summary, the application of machine learning to analyze news articles and their association with stock market prices draws on various theoretical frameworks within finance and economics. By leveraging machine learning techniques, researchers can extract valuable insights from unstructured textual data, enhancing decision-making processes in financial markets.

In our literature review section, based on Strauss and Smith's (2019) research we saw a basis for using text to assess impact on stock market for Tesla shares and show they react to tweets by its founder or news media coverage.

Practical Findings

(Please note full code can be found in my github repo for this project under notebooks/ 4_Sentiment_Analysis_and_ML.ipynb)

Fig 1.14: result from Notebook 4_Sentiment_Analysis_and_ML.ipynb

```
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
```

The practical findings from the analysis of text features predicting Dow Jones Close price are as follows:

1. **Linear Regression:** The Linear Regression model exhibits exceptional performance, with extremely low Mean Squared Error (MSE) of 0.2448, Root Mean Squared Error (RMSE) of 0.4948, and Mean Absolute Error (MAE) of 0.1768, indicating high accuracy in predicting the DJIA_Close based on the

features extracted from news articles. The R-squared value of 0.9999999974194661 signifies an almost perfect fit of the model to the data, suggesting that the independent variables explain nearly all the variance in the dependent variable.

2. **Ridge Regression:** In contrast to Linear Regression, the Ridge Regression model demonstrates inferior predictive performance, as evidenced by the significantly higher MSE of 25175.6716, RMSE of 158.6684, and MAE of 59.5469. Although the R-squared value remains relatively high at 0.9997346255025521, indicating a strong explanatory power of the model, it is slightly lower compared to Linear Regression. This suggests that Ridge Regression may have slightly less precision in predicting the DJIA_Close based on the news article features, potentially due to regularization constraints imposed by the Ridge Regression algorithm.
3. **Lasso Regression:** The Lasso Regression model yields moderate performance, with a Mean Squared Error (MSE) of 13416.4802, Root Mean Squared Error (RMSE) of 115.8295, and Mean Absolute Error (MAE) of 10.8867. These metrics indicate a reasonable level of accuracy in predicting the DJIA_Close based on the features extracted from news articles. The high R-squared value of 0.9998585780852661 suggests that the model explains most of the variance in the dependent variable, indicating a strong fit to the data. However, the higher RMSE and MAE compared to Linear Regression and Random Forest Regression suggest that Lasso Regression may not perform as well in terms of prediction accuracy.
4. **Random Forest Regression:** The Random Forest Regression model demonstrates excellent predictive performance, with a remarkably low Mean Squared Error (MSE) of 7181.8842, Root Mean Squared Error (RMSE) of 84.7460, and Mean Absolute Error (MAE) of 1.4387. These metrics indicate a high level of accuracy in predicting the DJIA_Close based on the features extracted from news articles. The exceptionally high R-squared value of 0.9999242964027588 suggests that the model explains nearly all the variance in the dependent variable, indicating an outstanding fit to the data. The significantly lower RMSE and MAE compared to other models further underscore the superior predictive power of Random Forest Regression in this context.

Limitations of Approach

On a granular level it can be argued that individual stocks can be affected by the kind of press coverage they get but even in the study we see that the market does not react the way the authors expected immediate. Information dissemination is a factor in how the market reacts. Perception is a very subjective and ergo people's perceptions can change. The information may have been positively when it was first released but with time people's opinions can change.

While the utilization of machine learning techniques to analyze news articles and their connection to stock market prices offers numerous benefits, it also comes with several limitations:

1. **Data Quality and Bias**: The quality and reliability of the news articles used for analysis can vary significantly. Biases in reporting, inaccuracies, or even fake news can distort the findings of the analysis. Additionally, the selection of news sources may introduce bias, as certain sources may have different levels of credibility or agendas.
2. **Linguistic Complexity**: News articles often contain nuanced language, sarcasm, or ambiguity, which can be challenging for machine learning models to accurately interpret. This complexity can lead to misinterpretations or erroneous sentiment analysis results, affecting the reliability of the analysis.
3. **Limited Contextual Understanding**: Machine learning models may lack the ability to fully understand the context surrounding news events. They may struggle to differentiate between positive and negative news within a broader context or accurately assess the significance of specific events relative to others.
4. **Market Efficiency and Randomness**: While the Efficient Market Hypothesis suggests that asset prices reflect all available information, market efficiency is not always perfect. Stock prices can be influenced by factors beyond news events, such as investor sentiment, market sentiment, or random fluctuations. Thus, attributing stock price movements solely to news articles may oversimplify the complex dynamics of financial markets.
5. **Correlation vs. Causation**: While machine learning models can identify correlations between news events and stock price movements, establishing causation is more challenging. News articles may coincide with stock price changes but determining whether the news directly caused the price movement requires careful analysis and consideration of other factors.
6. **Overfitting and Generalization**: Machine learning models trained on historical data may be prone to overfit to specific patterns or noise in the data, leading to poor generalization performance on unseen data. It's essential to validate the models rigorously and ensure they can accurately predict stock price movements beyond the training data.
7. **Regulatory and Ethical Considerations**: Using machine learning to analyze news articles for stock market prediction raises regulatory and ethical concerns. Regulatory bodies may scrutinize the use of algorithms for trading purposes, especially if it involves automated trading or market manipulation. Additionally, ethical considerations regarding data privacy, transparency, and fairness must be addressed to maintain trust and credibility in financial markets.

A comprehensive understanding of finance, economics, and market dynamics, coupled with robust methodology and careful validation, is essential for meaningful insights and informed decision-making in financial markets.

5. Limitations of Research

While the methodologies presented offer valuable insights into the relationship between news events and financial markets, it's essential to recognize several limitations inherent in the research. We've seen how each method had its limitations. However, deeper research with better ontologies and domain-specific knowledge and information can still make it possible to conduct this research and achieve the primary objective of finding associations between news and stock market prices.

6. Conclusion

In conclusion, while the tenets of systematic trading rely on market moving information to devise investment strategies the analysis of news articles to extract this type of information still has numerous challenges.

Text is not as easy to classify semantically and decipher to assess between the dependency and causation of such events on the stock market. Text is ambiguous while numbers are not. Interpretation of text is subjective, and this plays a role in its challenges.

Under Method 6 were able to establish a link with the stock market. Using Method 1 we see on a very high level the patterns of text emerging. In Method 2 we formulate a pipeline that would help us extract events. Method 3 adds to domain knowledge by finding domain related topics. Method 4 was difficult to interpret but this was due to my lack of understanding of this model. Method 5 indicates that text features can predict sentiment of news articles. Further research in this is a possibility since this doesn't feel complete. I made assumptions throughout my project that may have added bias how this research was conducted. For example, I should've used PCA to bring my text features to a lower dimension but instead I used all the TFIDF text features. I used TFIDF instead of Count Vectorizer for my machine learning models because I wanted the significance of topics not just the count of them.

Throughout this project I've tried numerous methods and approaches and encountered roadblocks that prevented me from completing this report. Numbers have an order to them that text data doesn't. This adds a level of complexity in not just applying machine learning techniques but in cleaning and preprocessing that is not as straightforward as it is with numerical data.

My kernels have crashed several times due to the text features overloading my models. This required going back to the drawing board and looking at the problem and text differently. While all the puzzle pieces of my research didn't fit neatly the way one

might expect they helped to draw the big picture of what this project was set out to do which was to explore the link between news articles and the stock market. Events were extracted as an intermediary step to determine whether the news was relevant or contained market moving information. However, after all these trials and tribulations I realize that this was never supposed to be a destination but a journey.

References

1. Saksham, K. (2023). Global News Dataset. https://www.kaggle.com/datasets/everydaycodings/global-news-dataset
2. Github: https://github.com/karnazko27/CIND-820-Big-Data-Analytics-Project
3. Kenton, W. (2022, April 21). Event-Driven Investing Strategies and Examples. Investopedia. https://www.investopedia.com/trading/advanced-trading-strategies-and-instruments/event-driven-investing-strategies-and-examples/
4. Maverick, J.B. (2023, September 30). The Weak, Strong, and Semi-Strong Efficient Market Hypotheses. Investopedia. https://www.investopedia.com/trading/trading-strategies/the-weak-strong-and-semi-strong-efficient-market-hypotheses/
5. Strauss, N., & Smith, C. H. (2019). Buying on rumors: How financial news flows affect the share price of Tesla. *Corporate Communications, 24*(4), 593-607. https://doi.org/10.1108/CC-11-2018-0126
6. Downs, A. (1972). Up and Down with Ecology-the Issue-Attention Cycle. Public Interest, 28(Summer), 38.
7. MacKinlay, A. C. (1997). Event studies in economic finance. *Journal of Economic Literature, 35*(1), 13-39.
8. Hogenboom, A., Hogenboom, F., Frasincar, F., et al. (2013). Semantics-based information extraction for detecting economic events. Multimedia Tools and Applications, 64(1), 27–52. https://doi.org/10.1007/s11042-012-1122-0
9. Kinariwala, S., & Deshmukh, S. (2023). Short text topic modelling using local and global word-context semantic correlation. Multimedia Tools and Applications, 82(48), 26411–26433. https://doi.org/10.1007/s11042-023-14352-x
10. Blei, D. M., Ng, A. Y., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993-1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993
11. Evangelopoulos, N. E. (2013). Latent semantic analysis. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(6), 683-692. https://doi.org/10.1002/wcs.1254
12. Lazrig, I., & Humpherys, S. L. (2022). Using Machine Learning Sentiment Analysis to Evaluate Learning Impact. *Information Systems Education Journal (ISEDJ), 20*(1), 13. ISSN: 1545-679X. Retrieved from https://isedj.org/; https://iscap.info
13. ChatGPT. "UMass and c_v Coherence Measure Formulas." OpenAI, 2024.