CIND 820
Project Abstract

Supervisor: not assigned yet

Karnaz Obaidullah

Student # 501000900

Date: January 22nd 2024

Event Classification using Reuters News Data

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.* Furthermore, events that classified in similar categories can be retrieved using *Recommender systems.* 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).

Other areas where event classification could be useful can be in social media marketing and product marketing. In both arenas, public opinion about relevant events can determine the

demand for consumer products and are thus important to consider before undertaking high budget marketing campaigns that may prove to be disastrous. One such example is the Pepsi Ad with Kendall Jenner (Jack of Marketing, 2023).

Chosen Dataset

For this project, I have chosen the Reuters Text Categorization dataset (Lewis, 1997) from the UC Irvine Machine Learning Repository. This is a collection of documents that appeared on Reuters newswire in 1987. The dataset is comprised of 5 features and 21,578 instances of data. 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

In achieving our objective of Event Classification of the Reuters News Dataset, we encounter the need to answer some questions before proceeding.

1. 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*?
2. 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?
3. How does classifying events help us in formulating an event-driven investment strategy? Does it satisfy the objective of taking advantage of temporary market mispricing?

References

1. Lewis, D. (1997). Reuters-21578 Text Categorization Collection. UCI Machine Learning Repository. https://doi.org/10.24432/C52G6M
2. 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/
3. 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/
4. Jack of Marketing. (2023, May 19). Why was Pepsi's Kendall Jenner ad a marketing disaster? The Marketing Jack. https://www.themarketingjack.com/blogs/why-pepsi-kendall-jenner-ad-marketing-disaster