Project Proposal

1. Project proposal and dataset

As part of McGill's AI society's Accelerated Introduction to Machine Learning Bootcamp, the proposed final project is a financial news sentiment analysis model. This project idea was inspired by Stefan Jansen's book "Machine Learning for Algorithmic Trading: Predictive models to extract signals from market and alternative data for systematic trading strategies with python, 2nd Edition," and will follow much of the methodologies introduced.

The dataset utilized is provided by Hugging Face: https://huggingface.co/datasets/financial_phrasebank, and consists of 4840 sentences from financial news categorized by sentiment. In the case of financial information, the use of statistical techniques for sentiment analysis has been low due to the lack of quality labeled datasets. For this reason, the financial_phrasebank will serve as a basis for model training.

2. Methodology

a) Data Preprocessing:

Dataset composition: 2879 neutral, 1363 positive, 604 negative. As a consequence of the imbalanced data, the underrepresented classes (positive and negative) will be "repeated." Furthermore, the training texts will be "cleaned" to only include words.

b) Machine Learning Model:

The project aims to classify financial news sentiment into three categories: negative, neutral, or positive. A Naive Bayes classifier will be implemented for this purpose, as the algorithm is very popular for text classification:

- Low computational cost
- performance comparable to neural networks for text classification
- simplicity of application

c) Evaluation Metric

- Confusion matrix
- accuracy

3. Application

For users to evaluate an article's sentiment, two options will be present in a web app: the article's URL link or text insertion. If the article's URL is provided, a web scrapper will gather the news article and feed it to the classification model. Once the user clicks "analyze," the article's sentiment will be displayed, including the automatic detection of the company (named entity recognition).