Data Analysis Project 2.

Textual Analysis using Natural Language Processing

1. Project Overview

The objective of this project is to develop a natural language processing model that can predict firm outcomes based on textual analysis. The model should be used to explore the explanatory power of new data on firm future performance, helping investors make informed decisions. Your analysis can involve but is not limited to machine learning, sentiment analysis, large language models. Contrast to Project 1, this project will give you enough flexibility to explore. Thus, all works involving textual analysis using NLP models will be acceptable.

Example idea:

Use text data in annual report (10-K) to predict firm stock return, stock volatility, ROA, leverage, cash, M&As, innovations, etc.

2. Grading and Submission:

The grading of the project will be divided into two parts: Group-level and Individual-level works.

At the group-level:

- 10% credits will be granted based on the completion and degree of perfection.
- Submit the program (code) used for basic data collection and preprocessing.
- Submit a short summary of data collection and preprocessing.
- Submit any other thing that you decided to share at the group-level.

At the individual-level:

- 10% credits will be granted based on the completion and machine learning performance (i.e., completion rate*F1 score).
- Submit the program (code) used for individual analysis (anything that your group did not share).
- Submit a short summary of (1) additional data sources that are not shared by the group (2) descriptions of firm outcome selection (3) descriptions of model selection (4) report your model performance (5) economic interpretation on important features and its effects on firm outcomes.

3. Project Goals

- **Develop Big Data Use Case in Finance:** Create predictive models to explore firm level performance.
- **Feature Engineering:** Learn how to decode and encode textual data into numeric features, using them in predictive models.
- **Backtesting and Validation:** Implement rigorous backtesting to evaluate model performance over historical data and validate its accuracy on out-of-sample data.
- **Deployment:** Deploy the model as an interactive tool or API for real-time predictions.

4. Functional Requirements

4.1 Data Collection

- **Textual Data Source:**
  - Download the cleaned files for 10-K/10-Q (textual data for US public firms' annual financial report) from https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/
  - You might have limited capacity of handling the full sample. You can select a subsample (e.g., 1-5 years only) from this dataset.
  - If you have alternative sources of textual data, feel free to use them instead.

- Other data you might be interested in for firm-level outcomes:
  - Historical stock price data from reliable financial data providers such as Yahoo Finance, CRSP, or CSMAR other financial databases.
  - Firm-level financial information: Compustat or CSMAR, Bloomberg.
  - M&As and firm innovation data are upon request from the instructor.
  - If you are interested in any other database, feel free to contact me.
- **Data Frequency:** Yearly.

4.2 Model Development

- **Model Selection:** Very flexible depends on your focus. Here are some examples:
  - For text embedding: word-to-vec, sentence transformer
  - For text classification: **Support Vector Machines (SVM)**
  - For analysis: NLTK (sentiment analysis), LDA (detect important topics)

6. Resources Required

- **Tools:**
  - Python, TensorFlow/PyTorch, Scikit-learn, Pandas, SQL databases, Cloud services (AWS/Azure/GCP).
  - Most open resource references might use python code. I would suggest you use Python. Feel free to search on StackFlow.com for the basic set up.
  - Google colab is a highly suggested platform for python programming and data storage. https://colab.research.google.com/
- **References**:
  - Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, *66*(1), 35-65.

    - Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of financial economics*, *110*(3), 712-729.

- **Weblink tutorials:**
    - https://www.tensorflow.org/hub/tutorials/tf2_text_classification
    - https://www.datacamp.com/tutorial/text-analytics-beginners-nltk
    - You can easily find many other helpful resources online. Feel free to use them as your guidance.

7. Success Criteria

- **Accuracy:** The model achieves an acceptable level of prediction accuracy (e.g., R^2, F1 score).
- **Usability:** Your model should be user-friendly and provide actionable insights for investors.
- **Performance:** The model is capable of providing predictions within the specified time frame and handling the expected data volume.