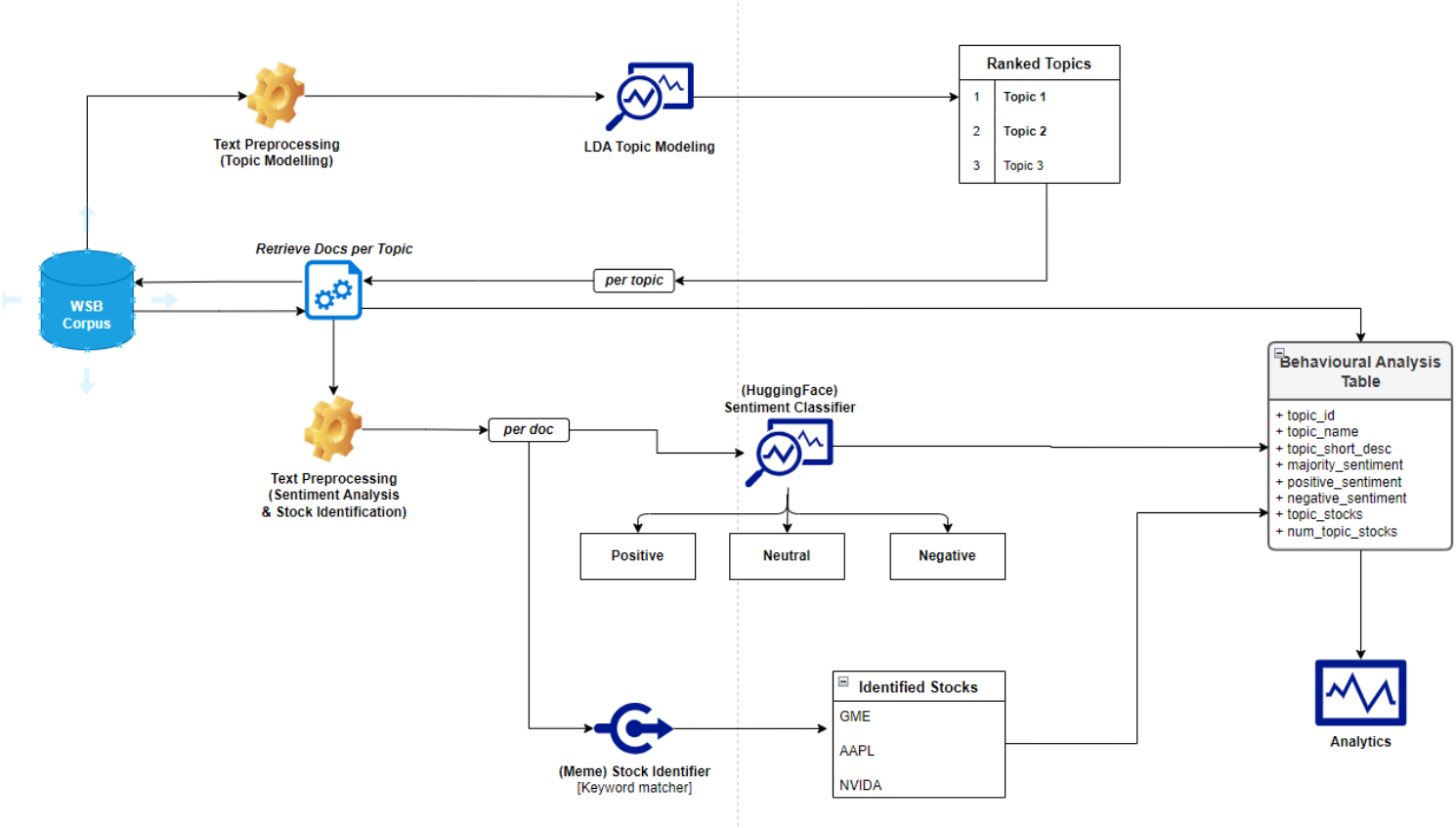
HEC Montreal
Msc. Data Science and Business Analytics

ML1 Project

Study Plan
MATH 60611A
Machine Learning for Large-Scale Data Analysis and Decision Making



Hair Albeiro Parra Barrera, Jonathan Alejandro Gonzales Perez, Cédric Lam
March 8, 2024

1 Study Plan

1. **Question to Answer:** Can we analyze the behavior and psychology of investors on r/wallstreetbets?. To do this, the goal is to conduct an in-depth analysis of investor psychology in a uniquely speculative and informal setting, including flagging potentially new meme stocks. From this, we will be able to draw insights into the behavior of retail investors in hype-driven markets, and a model that flags potential meme stocks based on subreddit activity patterns.

2. **Dataset:** Kaggle Reddit Wallstreetbets:
https://www.kaggle.com/datasets/gpreda/reddit-wallstreetsbets-posts

3. **Tentative Machine Learning Methods:**

   - **Latent Dirichlet Allocation (LDA) for Topic Modelling:** LDA will be used to identify latent topics within the subreddit posts, potentially revealing the dominant themes and discussions among investors.
   - **Sentiment Analysis with Pretrained Language Models (LLMs) via the Huggingface Interface:** This approach involves using pretrained language models to analyze the sentiment of posts, which can provide insights into investor sentiment towards specific stocks or market trends.
   - **Statistical Learning Methods and Data Analytics:** Once topics and sentiments are identified, various statistical learning methods and data analytics techniques will be applied to extract meaningful patterns and insights from the data.

4. **Questions & Discussion:**

   - What steps can we take to ensure the scalability and efficiency of our data processing and analysis pipeline, especially when dealing with large volumes of Reddit posts, to maintain the timeliness of our insights into investor behavior?
   - How can we enhance the robustness of our machine learning model, particularly in terms of feature selection and model validation, to ensure its effectiveness in flagging potential meme stocks accurately?