1. Motivation & Problem Statement

Public companies in the United States are required to file 10-Q quarterly reports with the SEC, which provide detailed insights into their financial health, operations, and risk factors. These filings contain vast amount of information about the direction of the company, but are highly unstructured, making it difficult for non industry professionals and even professionals to pick companies that will succeed.

This project aims to solve the problem of automatically forecasting stock price movement based solely on the textual content of 10-Q filings using NLP techniques as well as determine the factors that effects a company's success and failure..

Why this matters in NLP:

It demonstrates the real-world application of text analysis and understanding on financial documents. The 10-Q filings are also written in formal, domain-specific language, providing a challenging NLP use case. It contributes to the growing interest in quantitative finance, which uses machine learning and nlp in decision making.

Real-world applications:

Automated investment decision-making, alert systems for major corporate developments, and a retail investor tools for summarizing filings

2. Research Questions & Goals

Research Questions:

Can NLP models extract meaningful signals from 10-Qs to predict next-quarter stock price movements?
Which sections (e.g., MD&A, risk factors) carry the strongest predictive power?
Does summarization improve forecasting performance?

Goals:

- Develop an end-to-end pipeline that parses, preprocesses, and models 10-Q filings.
- Train a model that outputs the projected percentage change in stock price over the next quarter.

3. Dataset

10-Q Filings:

- **Source**: SEC EDGAR API (free, no API key required)

- **Collection Plan**: Scrape all 10-Q filings from the last 10 years using the EDGAR index files and Company Submissions API.
- **Estimated Size**: 100,000+ filings

Stock Price Data:

- **Source**: Yahoo Finance API (via libraries like yfinance)
- **Plan**: For each 10-Q filing, retrieve the stock's closing price on the filing date and 3 months later to compute the **quarterly return** (target variable).
- **Format**: JSON/CSV time series data matched to each CIK and filing date

---

4. Methodology

- **NLP Model**: Fine-tune a transformer model such as FinBERT or bert-base-uncased
- **Preprocessing**:
    - Extract and clean specific sections like MD&A
    - Tokenize and truncate/pad documents
    - Label each input with the % price change over the next quarter
- **Tech Stack**:
    - Python
    - Hugging Face Transformers
    - PyTorch or TensorFlow
    - SEC EDGAR and Yahoo Finance APIs
    - spaCy for preprocessing

5. Evaluation Plan

- **Metrics**:
    - For regression: MSE, MAE, R^2

    - For classification (up/down/neutral): Accuracy, Precision, Recall, F1

- **Baselines**:
    - Naive baseline (e.g., always predict 0% change)
    - Bag-of-Words + linear regression
    - Sentiment score from FinBERT

6. Expected Challenges

- **Long documents**: 10-Qs can exceed token limits of many transformer models. Will mitigate by focusing on key sections or using sliding windows.
- **Noise in stock price**: Price movements are driven by many external factors (macro news, earnings calls). May need smoothing or averaging.
- **Label mismatch**: Time alignment between filing dates and stock movement can be tricky—requires precise matching.

7. Timeline & Milestones

Step	Milestone
1	Finalize data sources, set up EDGAR & Yahoo Finance API scripts
2	Download and preprocess 10-Qs + stock data
3	Build baseline models (BoW, sentiment-only)
4	Fine-tune transformer model
5	Evaluate, tune, and analyze results
6	Finalize report and visualizations

8. References

- SEC EDGAR API Documentation
- FinBERT: Financial Sentiment Analysis with BERT
- Hugging Face Transformers: https://huggingface.co/
- Yahoo Finance API via Python: https://pypi.org/project/yfinance/