

# NLP-Based Analysis of Annual Reports: Asset Volatility Prediction and Portfolio Strategy Application

Insight 

SFI RESEARCH CENTRE FOR DATA ANALYTICS

By Xiao Li, Yang Xu, Linyi Yang, Yue Zhang, Ruihai Dong  
AICS-24 2024-12-10

HOST INSTITUTIONS



PARTNER INSTITUTIONS



FUNDED BY:



# Annual Report (10-K Report)

- Annual reports summarize a company's yearly performance and outlook. (Required by U.S. Securities and Exchange Commission)
- Quantitative Information:
  - Balance sheet, Cash flow,
- Descriptive Information:
  - Risk Factor disclosure, and Management Discussions.



# Research Motivation

## 1 Challenges in Extracting Risk from Long Texts

“Risk Factors” sections in annual reports hold critical insights, but their complex and lengthy structure challenges traditional methods in extracting and quantifying risk information effectively.

## 2 Challenges of Uncertainty in Financial Markets

Traditional methods struggle to capture long-term trends due to delayed market absorption of risk information and the complexities of market volatility.

## 3 The Need for Optimizing Investment Portfolios

Accurate Beta predictions enable optimized investment decisions and improved risk-adjusted returns.

# Past Approaches

## Early Finance Applications

Cash Flow Analysis, Stock movement  
Prediction...

1

2

3

## Advanced NLP Application

Sentiment Analysis, Risk Assessment...

## Leveraging Long Documents?

Train more massive pre-train model?  
Truncate the input?  
Or...?

# Research Content

## 1.Risk prediction:

We use the content of the "Risk Factors" section in the annual report to predict the Beta value (stock volatility) through the model.

## 2.Investment simulation:

The predicted Beta value is used to construct an investment portfolio, and the actual market performance of the portfolio is verified through simulation.

Using Annual Report's risk  
factor section to predict Beta  
(volatility of stock)

# What is Beta?

$$E(R_i) = R_f + \beta_i(E(R_m) - R_f)$$

$$\beta_i = \frac{E(R_i) - R_f}{E(R_m) - R_f} = \frac{Cov(R_i, R_m)}{Var(R_m)}$$

Beta Range	Implication
Beta > 1	Higher volatility than the market
0 < Beta ≤ 1	Lower volatility than the market
-1 < Beta < 0	Slightly inverse relationship with the market
Beta ≤ -1	Strongly inverse relationship with the market with high volatility

Beta is a metric that measures the volatility of a stock or asset relative to the overall market, reflecting its systematic risk, and serves as an important tool for assessing the risk and return of an investment portfolio.

# Methodology: Data Collection and Preprocessing

1

## Data Collection

Extracted "Item 1A: Risk Factors" from S&P 500 annual reports (2010-2020)

2

## Preprocessing

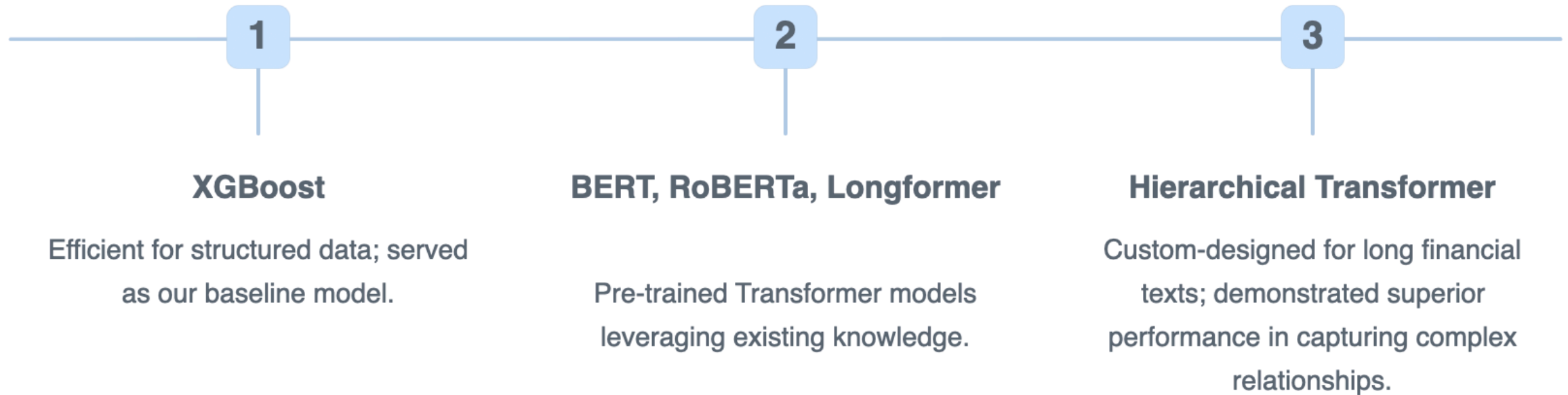
Cleaned text, removed HTML tags and tables

3

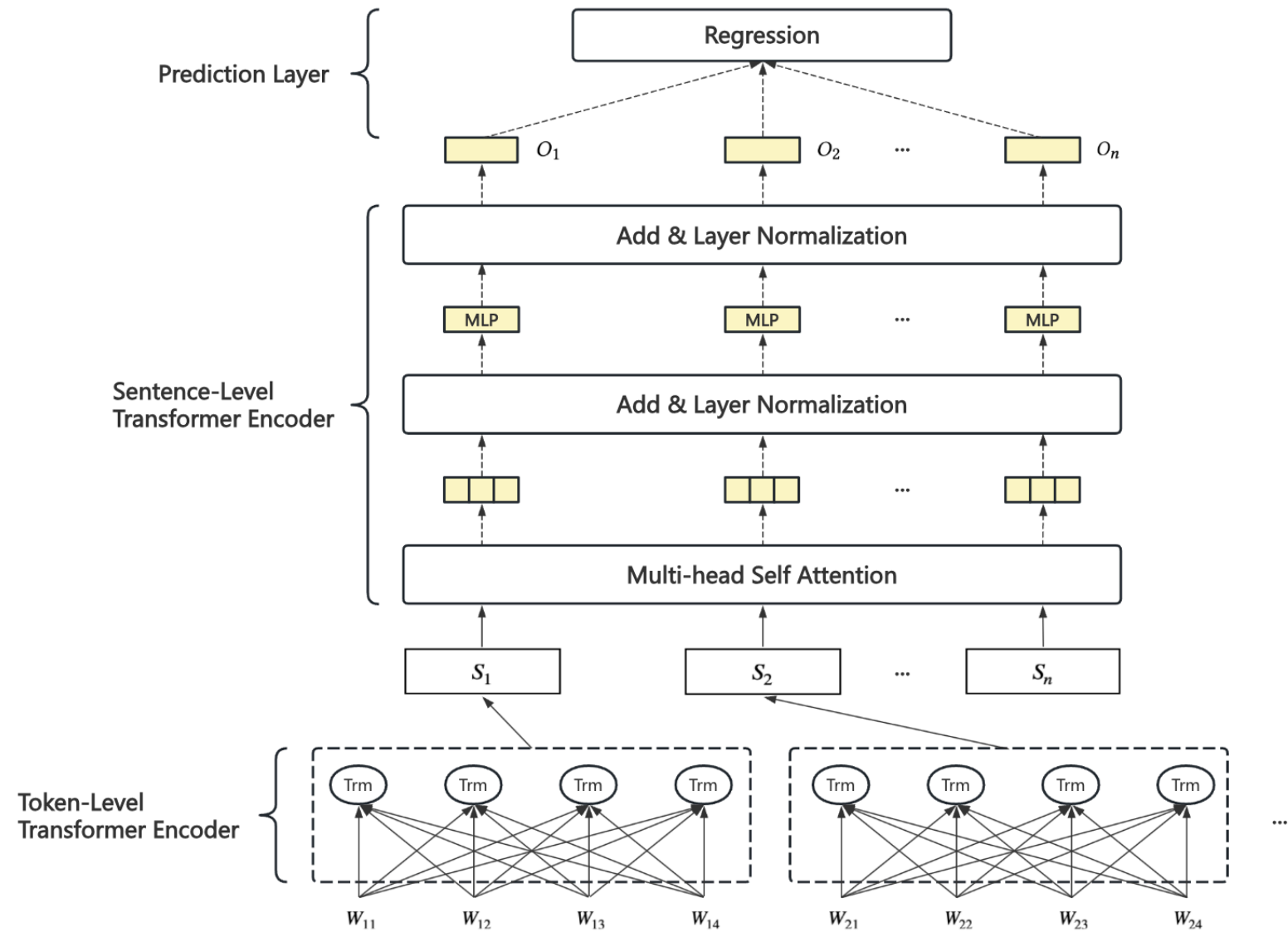
## Data Split

Training set: 2010-2018, Testing set: 2019-2020

# Prediction Models: A Progression of Sophistication



# Hierarchical Transformer Model



# Prediction Result

**Table 1**

Beta Prediction Using 10-K Report “Item 1A: Risk Factor” Section

Models	Mean Square Error (MSE)						
	n=3	n=7	n=15	n=30	n=60	n=90	n=180
XGBoost + TF-IDF	9.53087	1.43068	0.83950	0.30795	0.17176	0.13650	0.09951
BERT (bert-base-uncased)	9.36212	1.40890	0.86511	0.33718	0.18489	0.15794	0.12033
RoBERTa (roberta-base)	9.33685	1.39831	0.82897	0.33769	0.18527	0.15515	0.11957
Longformer (longformer-base-4096)	9.40540	1.43685	0.89168	0.32439	0.17902	0.14855	0.12384
Hierarchical Transformer-based	9.27465	1.41573	0.84309	0.32341	0.17346	0.12015	0.09634

The Hierarchical Transformer model demonstrates significantly lower MSE, particularly for long-term Beta predictions, highlighting its superior ability to capture complex relationships within long financial texts.

# Using Predicted Beta to Construct Portfolio Simulation

# Portfolio Construction Strategy



## **Beta Prediction**

Use model to predict Beta values

## **Company Selection**

Choose 10 highest and 10 lowest Beta stocks

## **Weight Optimization**

Use CML and Monte Carlo simulation

## **Portfolio Formation**

Construct final balanced portfolio

# Beta Prediction Results

# 180

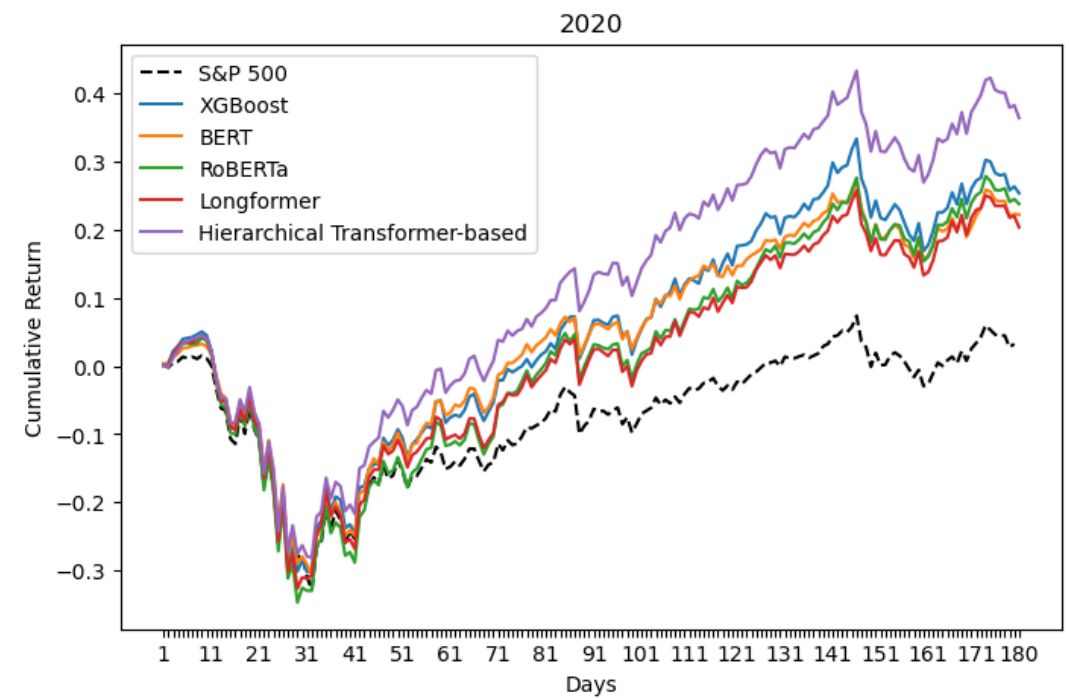
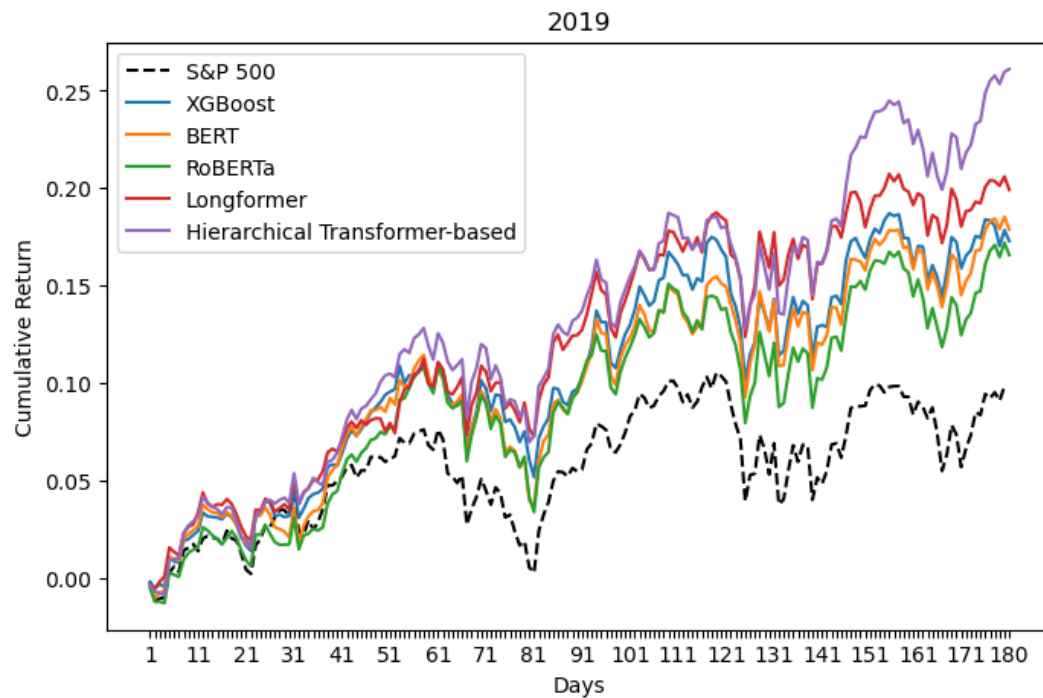
**Best Prediction Horizon**

Days for most accurate Beta predictions across all models

# 21%

**Return Increase**

Average portfolio return increase compared to S&P 500 benchmark



# Key Insights and Future Directions

## **Leveraging LLMs:**

Develop specialized models for annual reports to capture risk information more comprehensively, enhancing prediction accuracy and adaptability.

## **Exploring Cross-Domain Applications:**

Extend the methodology to other text-heavy domains, such as legal documents and medical reports, to test model applicability in diverse scenarios.

## **Incorporating External Factor Analysis**

Integrate external risks and macroeconomic indicators into the prediction framework for a more holistic assessment of company and market volatility.

## **Real-Time Data Integration**

Combine real-time data streams, like news and social media sentiment, with annual report analysis to provide dynamic risk predictions and decision-making tools.

# Thank you!

## Q&A



## **NLP-Based Analysis of Annual Reports:** Asset Volatility Prediction and Portfolio Strategy Application

**Author:** *Xiao Li (UCD)*

Yang Xu, Linyi Yang, Yue Zhang, Ruihai Dong

**Contact:** [xiao.li@ucdconnect.ie](mailto:xiao.li@ucdconnect.ie)

**Event:** AICS 2024

**Date:** Dec 9-10, 2024