Predicting Future Financial Volatility from Financial Reports with SEC 10-K Report Benchmark

Our Project Overview

01 Main goal of our project

Taking corporate annual financial reports as the object of analysis, this study focuses on exploring the predictive ability of textual features on future stock price volatility. Based on empirical tests of machine learning regression models, we systematically evaluate the differences in the predictive effectiveness of two types of textual representations: sparse features (BOW Model) and dense features (BERT, Word2Vec like).

02 Conclusion

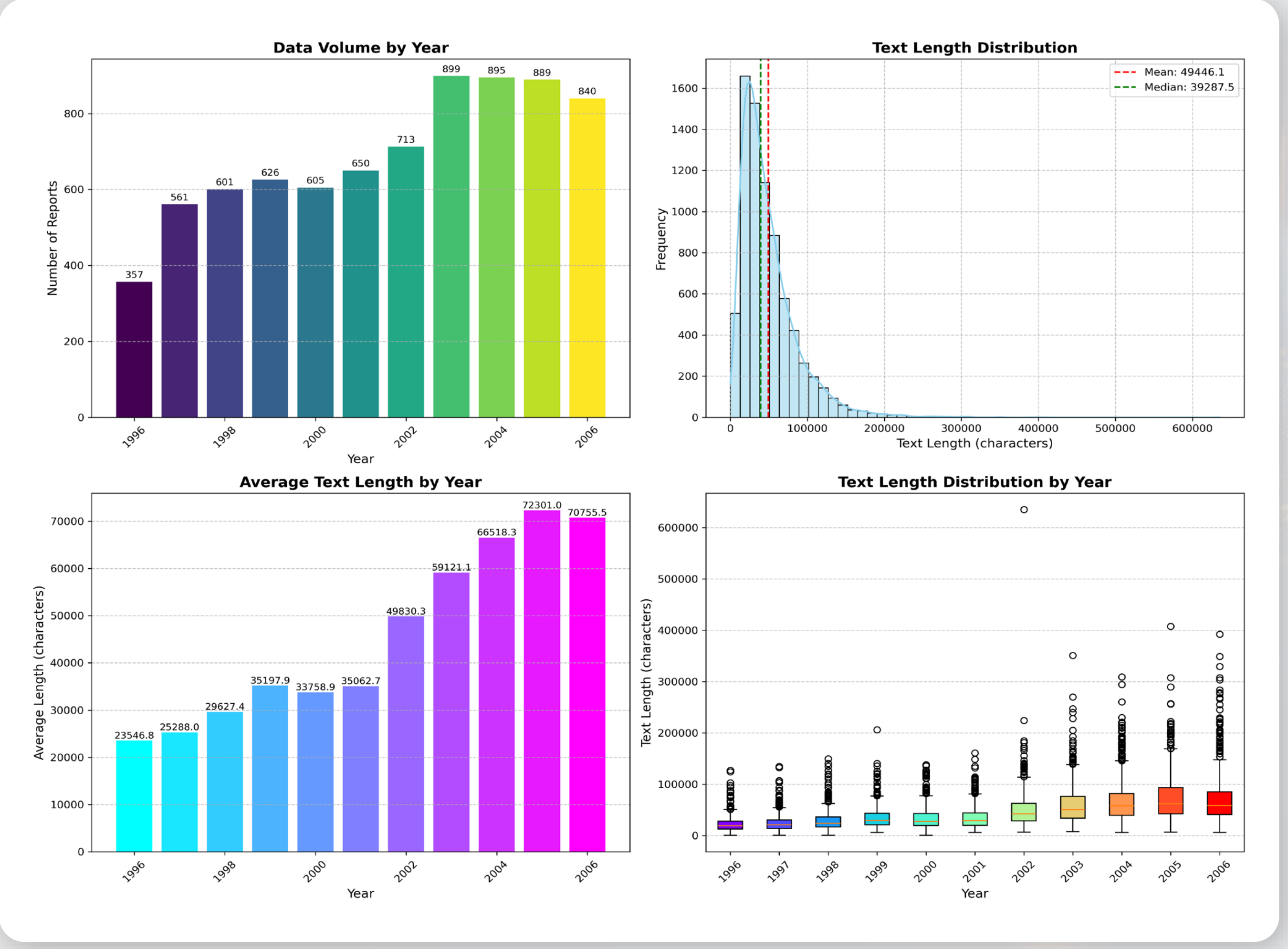
On the dataset studied in this project, the ensemble learning algorithm XGboost has the highest prediction accuracy, the effectiveness of textual information in interpreting volatility exhibits feature type heterogeneity, and the prediction accuracy of volatility based on keyword sparse features is higher than that based on semantically embedding from pre-trained models. This may imply that surface lexical patterns may be more powerful risk indicators than deep semantic representations in financial text analysis scenarios.

* =Ө= +SKGJ M≠ŒJ I

Data Scale

Lexical Diversity

Sentiment Analysis

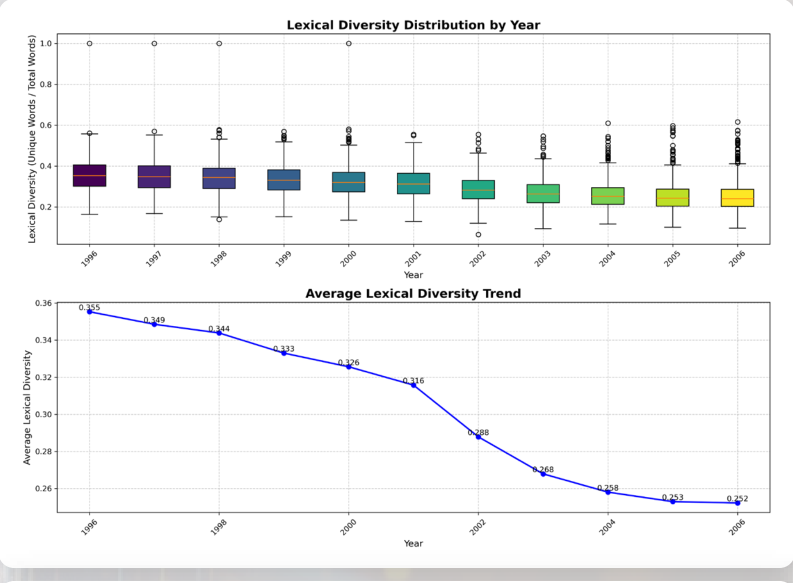


Data Volume by Year

Text Length Distribution

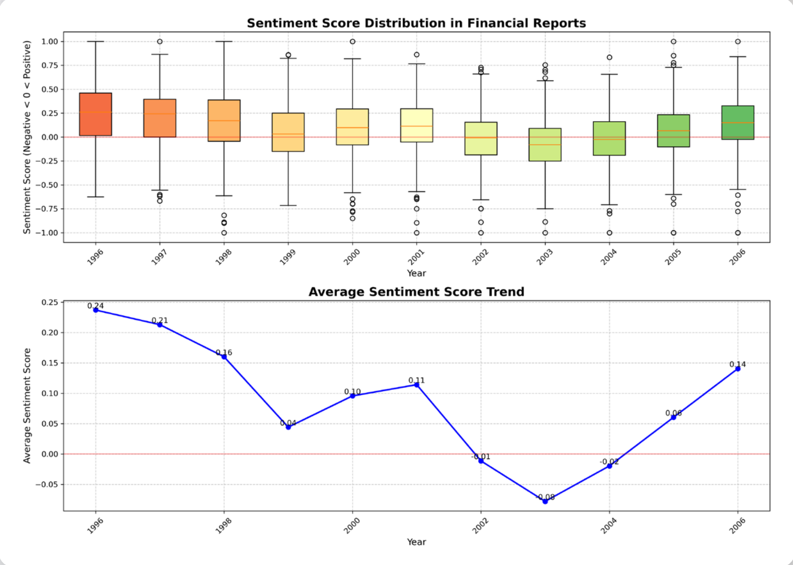
Average Text Length by Year

Text Length Distribution by Year



Lexical Diversity Distribution by Year

Average Lexical Diversity Trend

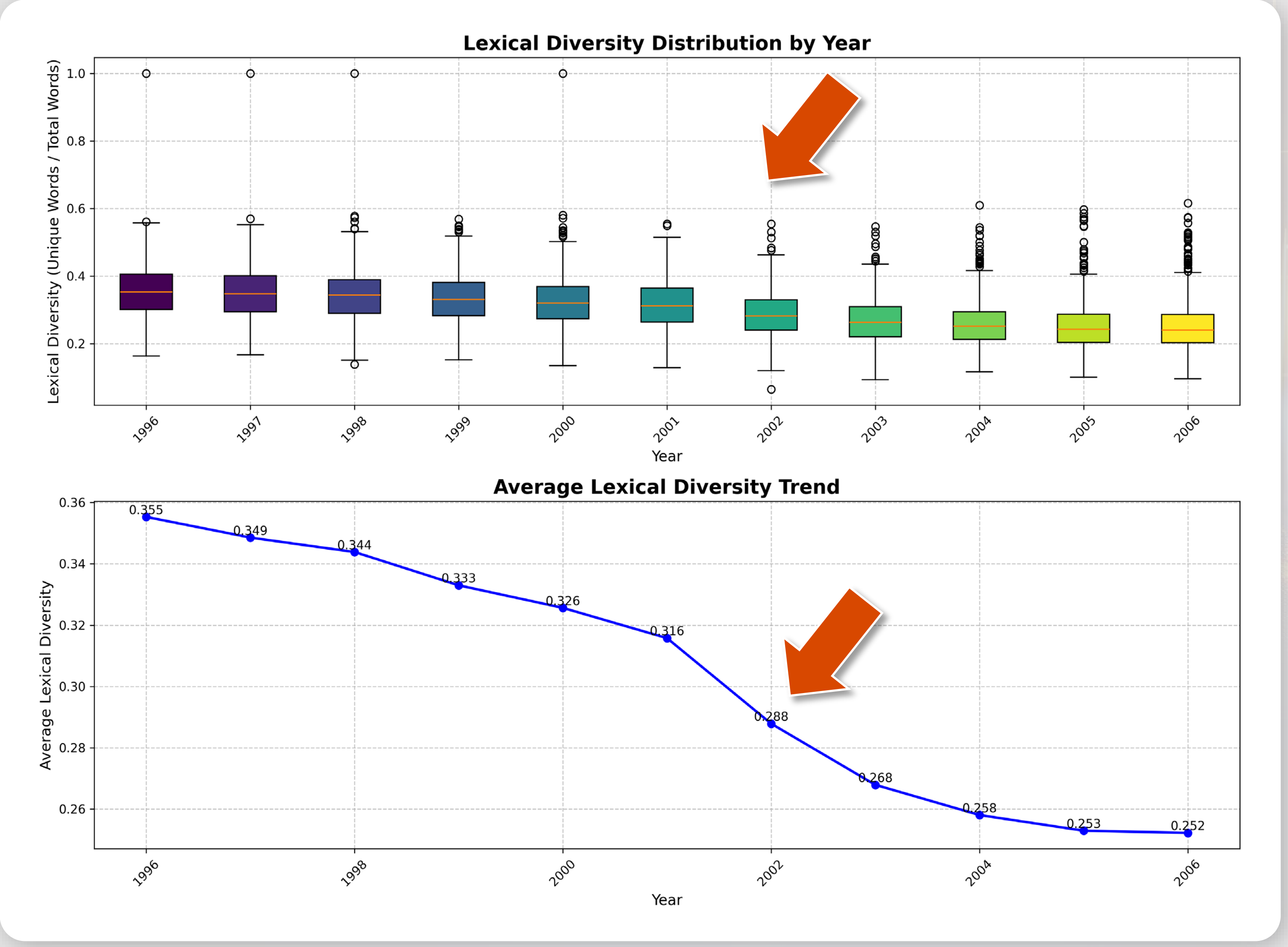


Sentiment Score Distribution in Financial Reports

Average Sentiment Score Trend

Data Exploration

Lexical Diversity & Sentimental Analysis



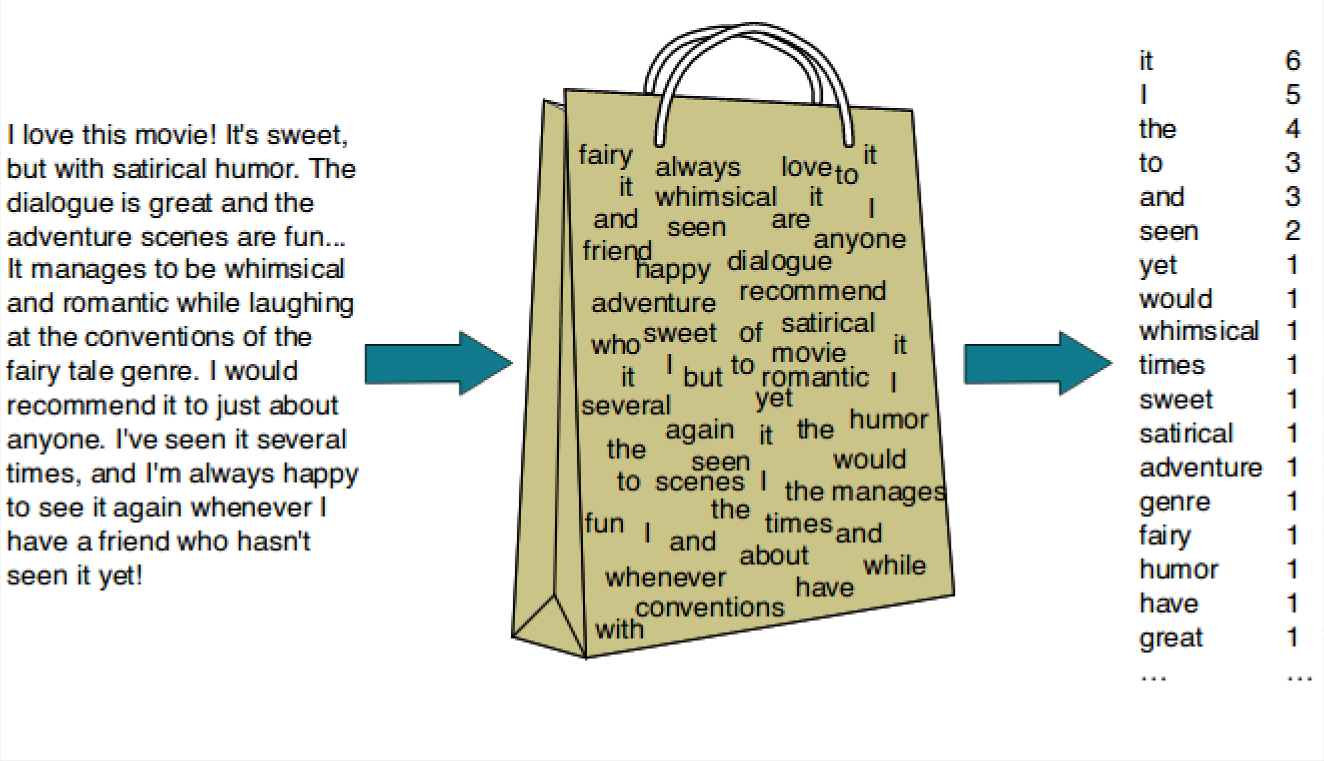
Lexical Diversity Distribution by Year

Average Lexical Diversity Trend

In July **2002**, the United States enacted the *Sarbanes–Oxley Act*, imposing stricter disclosure standards for public companies.

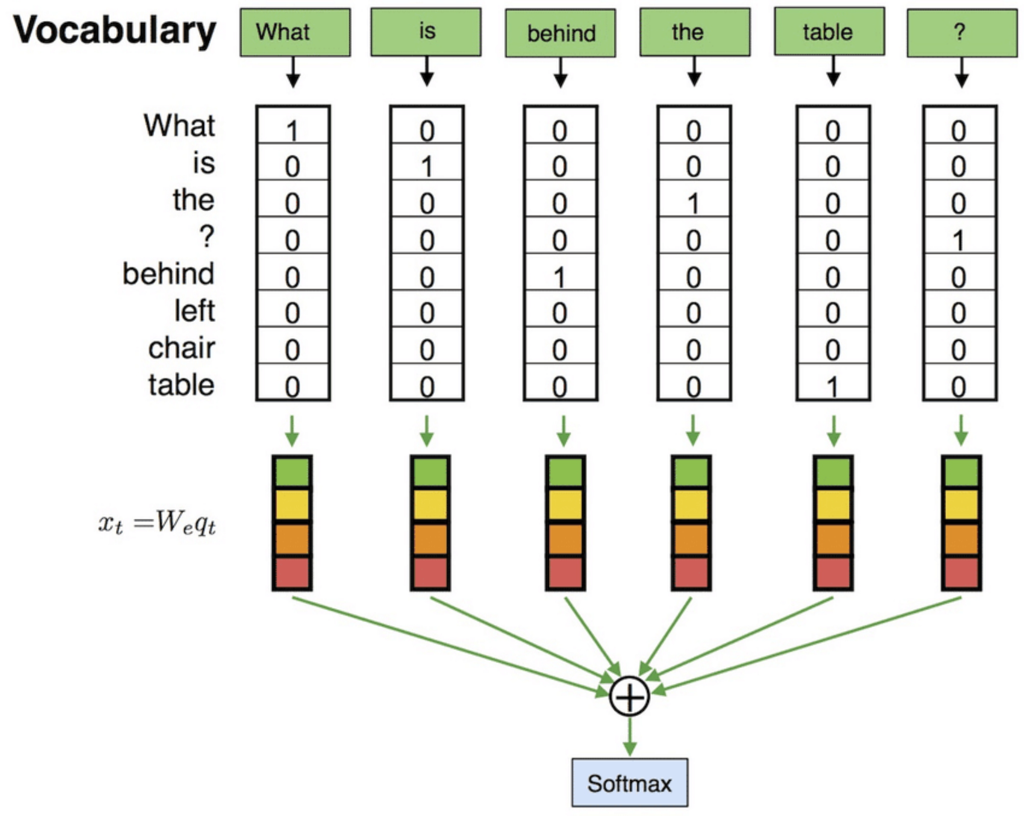
Given the exogenous shock nature of the policy intervention, the project decided to divide the experiment into two time intervals, nearly pre- and post-policy, respectively

Bag of Words (BOW) Model



I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...



Vocabulary	What	is	behind	the	table	?
What	1	0	0	0	0	0
is	0	1	0	0	0	0
the	0	0	0	1	0	0
?	0	0	0	0	0	1
behind	0	0	1	0	0	0
left	0	0	0	0	0	0
chair	0	0	0	0	0	0
table	0	0	0	0	1	0

[NLP: Bag of words and TF-IDF explained! | by Koushik kumar | Medium]

🛍️ Bag of Words meets Random Forest 🌳 | Kaggle

Feature Engineering – Text Analysis

Statistical Features Extraction

Term Frequency (TF)

- Frequency of words in the current document
- In use, we applies a **log transform** to mitigate outlier effects

$$h_j(d) = \log(1 + freq(x_j; d))$$

Bigram - TF

- Extracts **both term frequency and adjacent bigrams**, and compute their probability
- Captures common phrases and neighboring word relationships in text (e.g., “net loss”)

TF - IDF

- Builds on TF by incorporating **Inverse Document Frequency (IDF)**
- Reduce the weight of common words
- Emphasize rare words

Bigram – TF - IDF

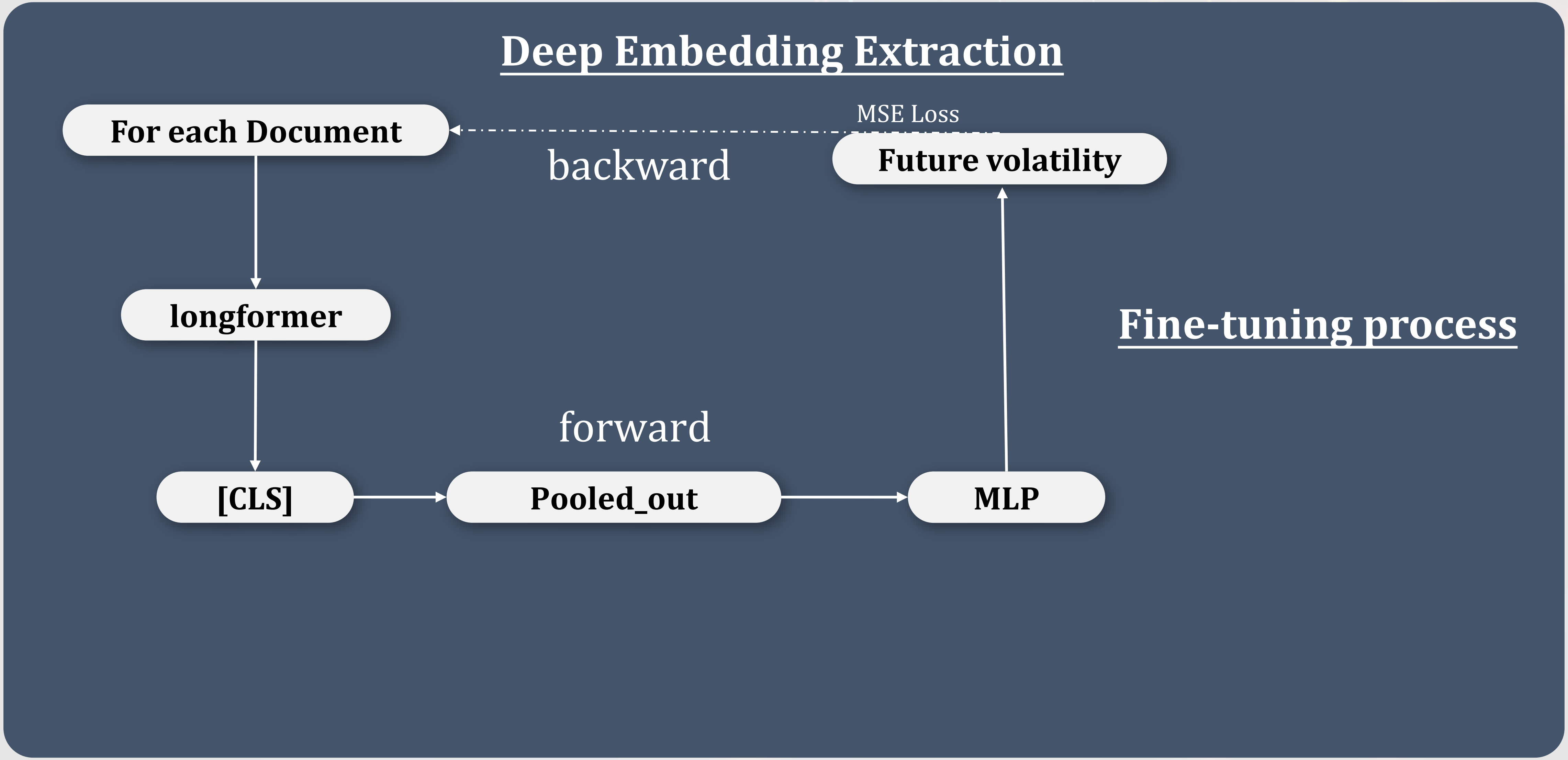
- Builds on Bigram-TF by adding IDF weighting
- Balances phrase frequency with corpus-wide distribution
- Increases discrimination for less frequent phrases

Pre-trained Model Deep Embedding

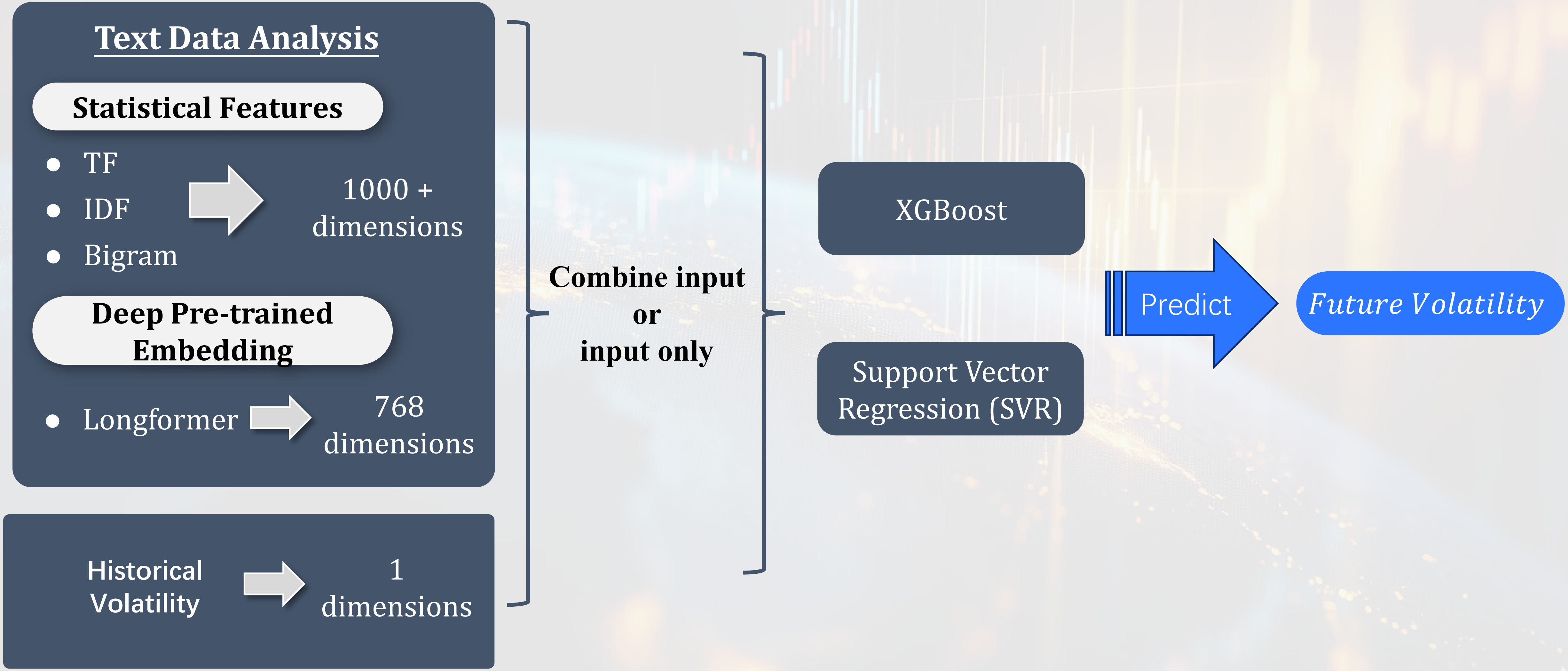
BERT - new Google algorithm update – Greenlogic

KiKaBeN - Longformer: The Long-Document Transformer (2020)

Feature Engineering – Text Analysis



Framework & Workflow



Text Data Analysis

Statistical Features

- TF
- IDF
- Bigram

1000 + dimensions

Deep Pre-trained Embedding

- Longformer

768 dimensions

Historical Volatility

1 dimensions

Combine input or input only

XGBoost

Support Vector Regression (SVR)

Predict

Future Volatility

Training Experiment Design

Training & Test Dataset

- To predict 2006 vol, using dataset from 2001 to 2005
- Similary, using dataset from 1996-2000 to predict 2001 vol

Basic data clearing

- Remove stopwords, Stemming, and convert text to lowcase for traditional stastistical features
- There is no need to do any preprocess to the text which used for pre-trained model

Xgboost Hyperparameters

max_depth=6
min_child_weight=1
min_loss=0
learning rate=0.1
estimators=100
L2 penalty=1

SVR Hyperparameters

kernel = rbf
error penalty = 1.0
tolerance = 0.001

Before the *Sarbanes–Oxley Act*

feature_type	model_type	MSE	R^2
bigram_tfidf_with_logv12	XGBoost	0.18017763	0.5610444
tf_with_logv12	XGBoost	0.18414356	0.55138245
bigram_tf_with_logv12	XGBoost	0.18419326	0.55126137
longformer_bigram_with_logv12	XGBoost	0.18487247	0.54960667
tfidf_with_logv12	XGBoost	0.18668545	0.5451898
bigram_tfidf_with_logv12	SVR	0.19303066	0.52973137
bigram_tfidf	SVR	0.19695076	0.52018107
bigram_tf_with_logv12	SVR	0.19697474	0.52012265
logv_minus_12_only (baseline)	XGBoost	0.19855868	0.51626381
bigram_tf	SVR	0.20185373	0.50823627
bigram_tfidf	XGBoost	0.20215232	0.50750885
tfidf_with_logv12	SVR	0.20340644	0.5044535
tfidf	XGBoost	0.20359978	0.50398249
longformer_bigram	XGBoost	0.20497933	0.50062157
tf	XGBoost	0.20634332	0.49729856
bigram_tf	XGBoost	0.20679932	0.49618765
tfidf	SVR	0.20871195	0.49152802
tf_with_logv12	SVR	0.21012691	0.48808084
tf	SVR	0.21907049	0.46629214
longformer_only	XGBoost	0.32508698	0.20801076
longformer_only	SVR	0.4138766	-0.0083019
longformer_bigram_with_logv12	SVR	0.50983793	-0.2420865
longformer_bigram	SVR	0.509838	-0.2420867
logv_minus_12_only (baseline)	SVR	2.41380051	-4.8805925

Feature	Importance
Historical vol logv12	80.85540771
estat	24.88868141
argentina	10.36391068
tangibl	9.533727646
quantit qualit	8.494257927
review	8.422605515
net loss	7.985421181
energi	7.944610596
seed	7.848720074
equip million	7.541954994
oper effici	7.292068481
corn	7.071839809
rate primarili	7.056113243
suppli chain	6.94379425
feed	6.826294422
offset effect	6.817842007
fx	6.790940285
tax per	6.622413635
apb	6.479614258
texa	6.402114868

After the *Sarbanes–Oxley Act*

feature_type	model_type	MSE	R^2
bigram_tf_with_logv12	XGBoost	0.127489606	0.558582722
longformer_bigram_with_logv12	XGBoost	0.127771108	0.557608055
bigram_tfidf_with_logv12	XGBoost	0.130517287	0.548099743
tf_with_logv12	XGBoost	0.133240954	0.538669375
tfidf_with_logv12	XGBoost	0.135382626	0.531254098
logv_minus_12_only (baseline)	XGBoost	0.143710815	0.502418755
bigram_tf	XGBoost	0.165544409	0.426822585
longformer_bigram	XGBoost	0.166453559	0.423674765
bigram_tfidf_with_logv12	SVR	0.166521462	0.423439658
bigram_tfidf	SVR	0.170356544	0.410161154
bigram_tfidf	XGBoost	0.170797915	0.408632958
bigram_tf_with_logv12	SVR	0.173035605	0.400885229
tf	XGBoost	0.17562322	0.391925926
tfidf	XGBoost	0.177506434	0.385405527
bigram_tf	SVR	0.180300338	0.375731973
tfidf_with_logv12	SVR	0.186845658	0.353069599
tfidf	SVR	0.194527815	0.326471062
tf_with_logv12	SVR	0.19598403	0.321429093
tf	SVR	0.212755473	0.263360011
longformer_only	XGBoost	0.319589357	-0.106539341
longformer_bigram	SVR	0.322647628	-0.117128234
longformer_bigram_with_logv12	SVR	0.322647628	-0.117128235
longformer_only	SVR	0.421106638	-0.4580306
logv_minus_12_only (baseline)	SVR	2.704373234	-8.363563935

Feature	Importance
Historical vol logv12	119.4632797
divert	20.41123962
gener administr	20.03438187
actual futur	15.14949512
length	13.87048531
act amend	13.25977898
initi recognit	12.5438633
cash proce	11.82049561
stock may	10.74275684
fda approv	10.69298553
product shipment	10.28904819
financi account	9.902664185
requir capit	9.301649094
impact inflat	9.010601044
interest entiti	8.736427307
reit	8.259461403
increas billion	8.059524536
regularli review	7.951934814
net loss	7.933226585
coast	7.848445892

)JI?OPNEJI

This project draws the following three core conclusions from a systematic model comparison study:

One, in terms of financial text feature engineering, the hybrid feature that incorporates Word Frequency-Integrated Dual Grammar (TF-IDF + Bigram) and traditional market factors (historical volatility) demonstrates superior predictive capability, which is significantly better than semantically dense features based on fine-tuning of pre-trained models.

Second, the integrated learning approach has significant advantages in this study. Specifically, the Xgboost model significantly outperforms the SVR model in terms of accuracy and becomes the preferred model in this study.

Third, financial text analysis has obvious domain specificity. Surface-level textual statistical features may have an advantage over deep semantic embeddings in terms of interpretability of risk representations. This finding provides a new possibility hypothesis for textual analysis in the financial quantitative domain, which has important research value and practical significance.

Interpretability Challenges

Deep Learning Models

Interpretability challenges with deep learning models like BERT.

Understanding the decision-making process of complex models is difficult.

, POPMA+SKGI M+OEI I

Future exploration includes integration of interpretability frameworks (e.g., SHAP values).

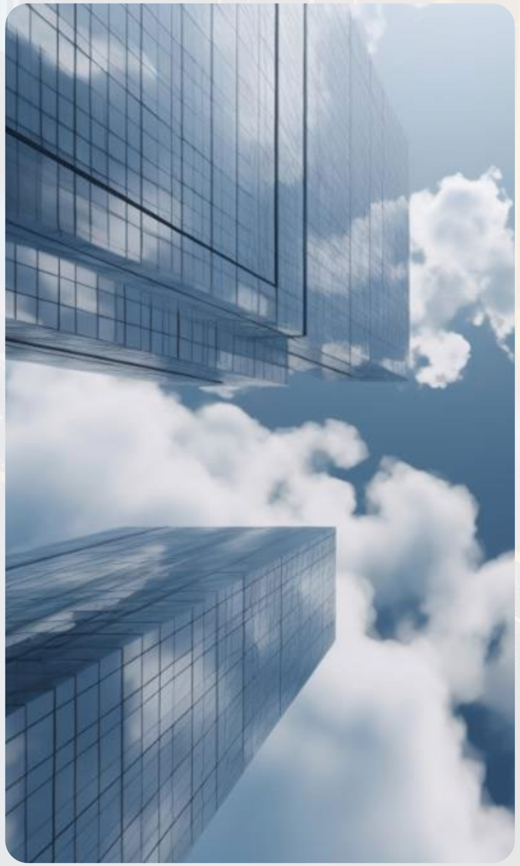
This can help in explaining the model's predictions and improving trust.

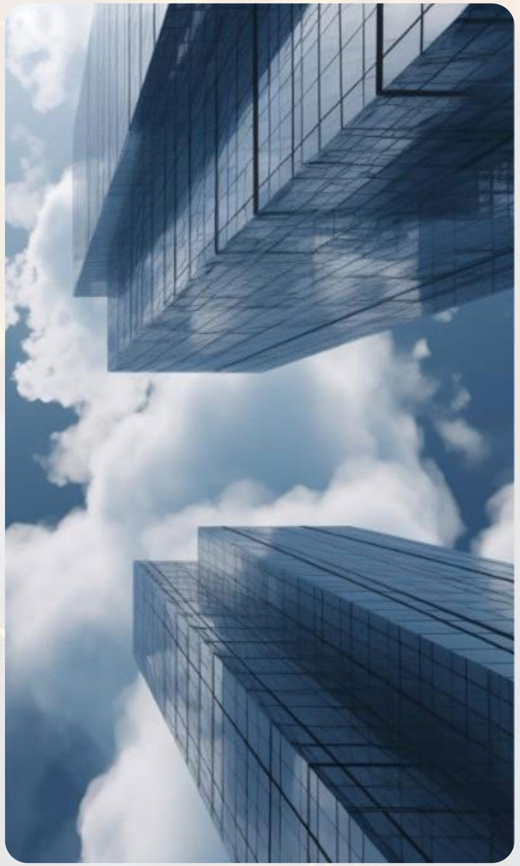
Potential for Expansion

AV[LU[PHSL_WHUZPVU^ P[O SHYNLYHUK UL^ LYKH[HZL[Z(

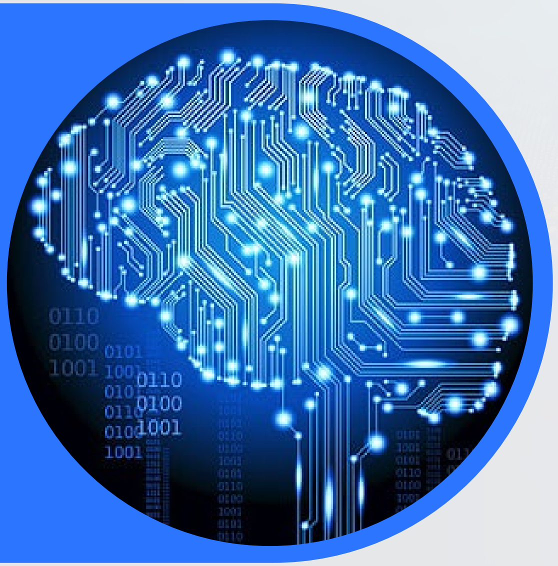
5 VU[PU\ V\ZPT WYV] LT LU[VMT VKLSZ^ P[O T VYL KH[H HUK HK] HUJLK

[LJ OUPX\ LZ(





References



Ali, Amal Al, et al. "A powerful predicting model for financial statement fraud based on optimized XGBoost ensemble learning technique." *Applied Sciences* 13.4 (2023): 2272.

Rawte, Vipula D., Mohammed J. Zaki, and Aparna Gupta. "FETILDA: An Effective Framework For Fin-tuned Embeddings For Long Financial Text Documents." *arXiv e-prints* (2022): arXiv-2206.

Chang, Ariana, Tian-Shyug Lee, and Hsiu-Mei Lee. "Applying sustainable development goals in financial forecasting using machine learning techniques." *Corporate social responsibility and environmental management* 31.3 (2024): 2277-2289.

Oukhouya, Hassan, and Khalid El Himdi. "Comparing machine learning methods—svr, xgboost, lstm, and mlp—for forecasting the moroccan stock market." *Computer Sciences & Mathematics Forum*. Vol. 7. No. 1. MDPI, 2023.

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting Risk from Financial Reports with Regression.

Thank you