TEAM MEMBER

FPT Education
FPT UNIVERSITY

PREDICTION OF

BANK SHARE PRICES

USING DATA MINING:

A case study of listed banks in Ho Chi Minh stock exchange

TEAM MEMBER

SUPERVISOR:

Nguyen Phu Ha

LEADER

Pham Thi Ngoc Ha
HS130118

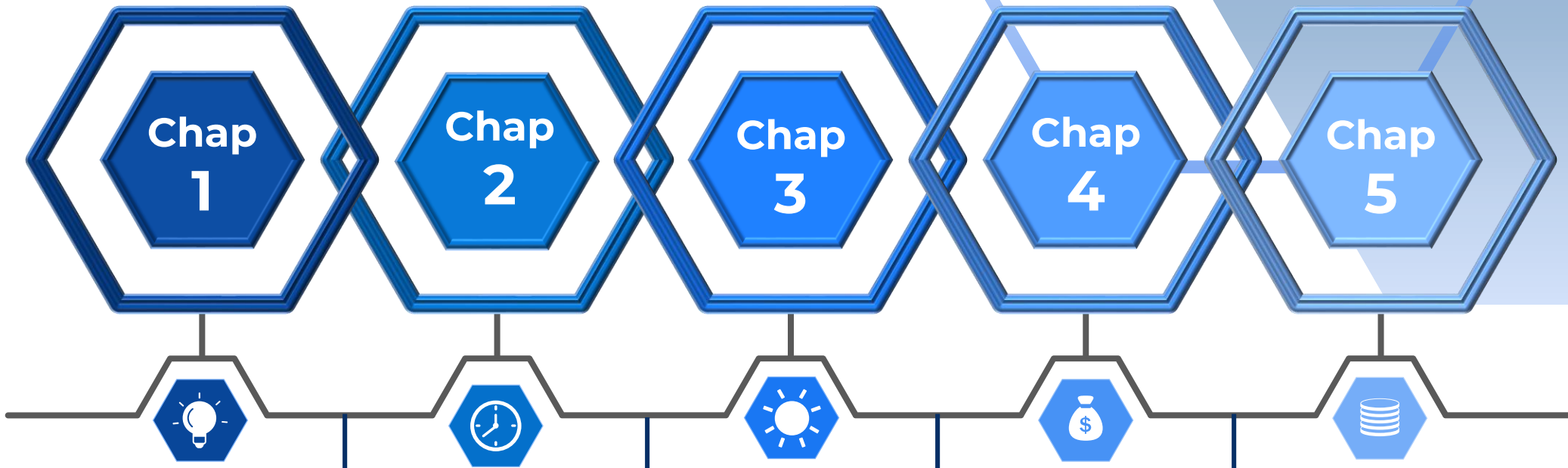
MEMBER

Nguyen Van Tuan
HS130363

MEMBER

Nguyen Thi Nga
HS130097

TABLE CONTENT



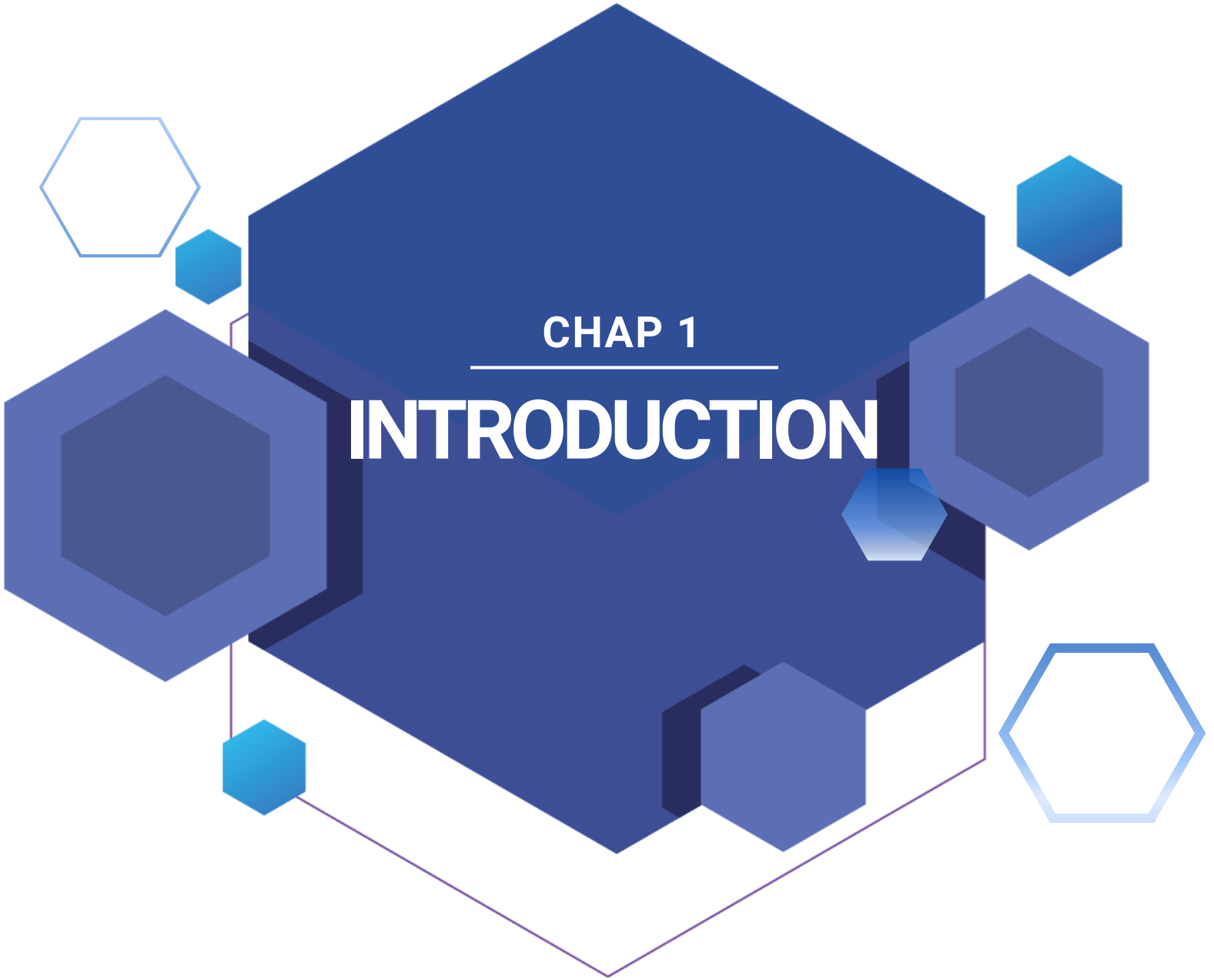
Brief overview of the research topic, topic's research objective, research question, research subjects, research scopes, thesis structures.

Theoretical and empirical review in available study on the research topic, application of data mining methods in stock price predicting to support this thesis

Discussion of methods of analysis, data collection used in this thesis and building detail of predicting stock price programs.

trend, seasonality variation of 14 listed banks on HOSE and predictions of bank share prices using data mining.

Findings and reccomendations to regulators and investors


CHAP 1
INTRODUCTION

TOPIC BACKGROUND

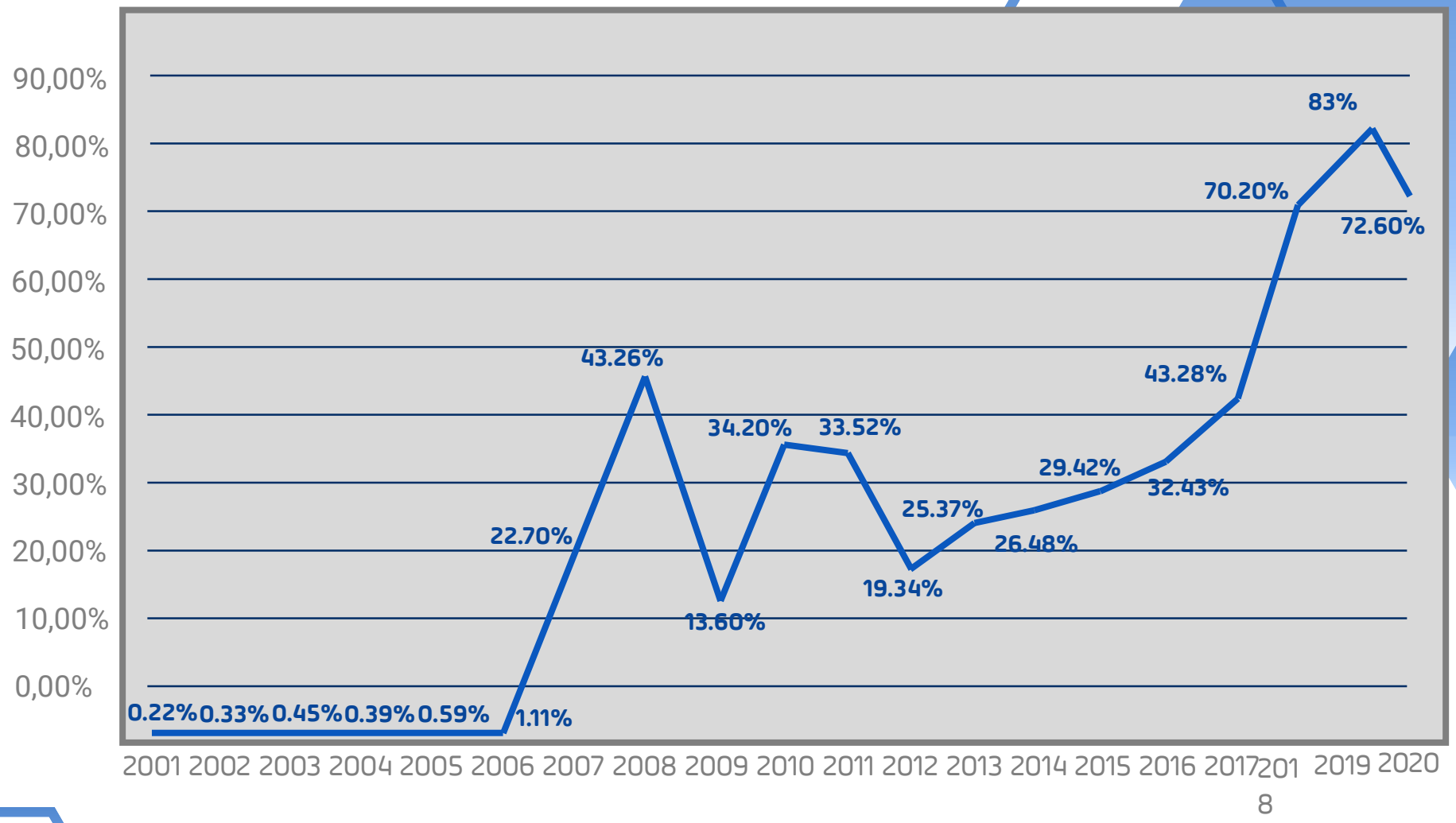


Figure 1.1: 20-year development journey of Vietnamese securities.(2001-2020)

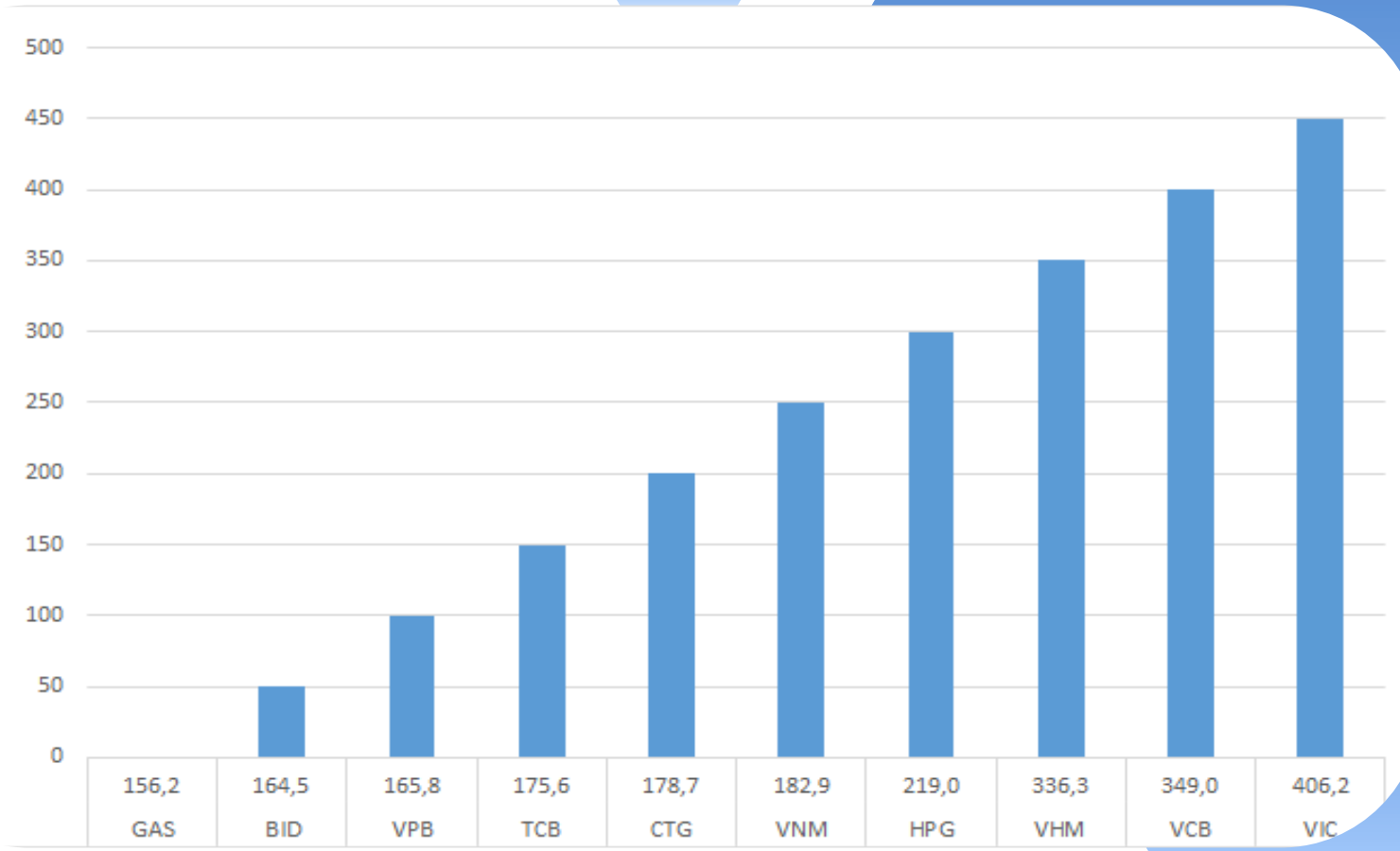


Figure: 1.2: Top 10 stocks by market capitalization on HOSE by the end of May 19, 2021

Why is data miningconsidered to be powerful in predicting stock?

REASEARCH OBJECTIVES

1. Systematize the theoretical basis of share prices and behavior of bank share prices

2. Apply research methods using data mining techniques to build a test program to forecast bank stock prices

3. Conclude findings from deploying the method and share experience about how to improve the accuracy in forecasting bank share prices.

4. Propose recommendations for investors and regulators based on the research results.

RESEARCH QUESTION

1. How does data-mining affect the valuation models of bank stocks in Vietnam's stock market?
2. Which methods of data mining is the most powerful in predicting the behavior of bank share prices and how does it affect investors' decisions in Vietnam's stock market?
3. What are the recommendations for individual and institutional investors?

RESEARCH SUBJECTS

- Stock price of banks listed on HOSE
- Closing market prices and price volatility during the period between January 2014 and August 2021
- Application of data mining in predicting for shares prices of 03 representatives, 03 representatives of bank shares

RESEARCH SCOPE

Scope of contents:

A case study of 3 bank shares BID, TCB and TPB.

About scope of time:

January 2014 - August 2021

Scope of space:

14 banks listed in HOSE;

.In the application of data mining in predicting bank share prices: 03 shares of listed banks,

CHAP 2

LITERATURE REVIEW

Previous literature and studies relevant to the research subjects

Tô Nguyễn Nhật Quang (2007), "Ứng dụng mô hình GAAR và ANFIS dự báo thị trường chứng khoán"

Đặng Hồng Phú (2008), "Ứng dụng microsoft time series xây dựng hệ thống dự báo thị trường chứng khoán Việt Nam".

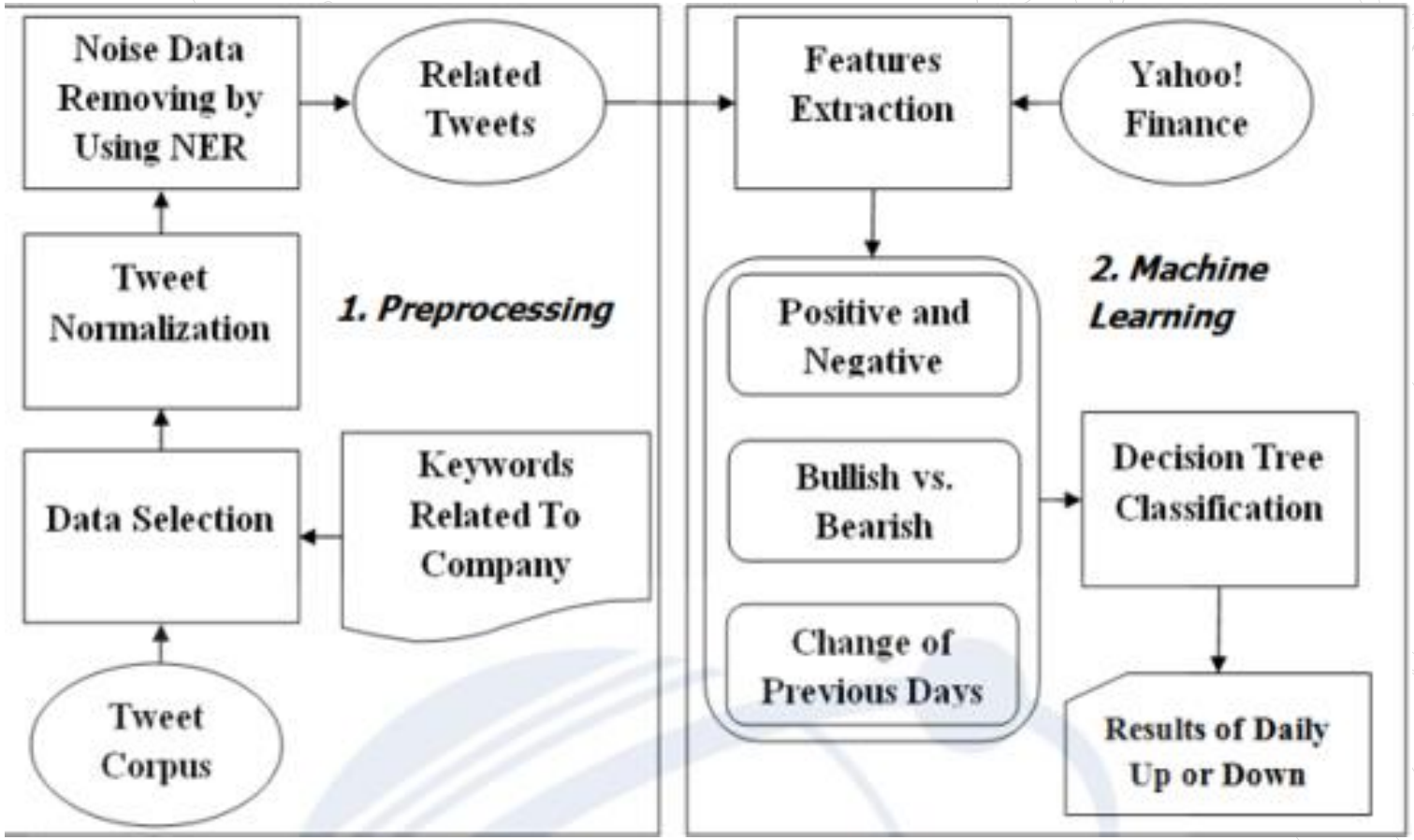
Nguyễn Thị Việt Hà(2020)" Tổng quan về khai phá dữ liệu và phương pháp khai phá luật kết hợp trong cơ sở dữ liệu ".

Trịnh Thanh Ngọc (2013), "Ứng dụng nền tảng Twitter trong việc dự báo thị trường chứng khoán"

Việt Nam

ThS. Lê Thị Thu Giang, ThS. Vũ Thị Huyền Trang (2021)"Ứng dụng mạng nơ-ron nhân tạo (ANNs) trong dự báo giá đóng cửa các mã cổ phiếu niêm yết trên sàn chứng khoán "

Trịnh Thanh Ngọc (2013), "Ứng dụng nền tảng Twitter trong việc dự báo thị trường chứng khoán"



ABROAD

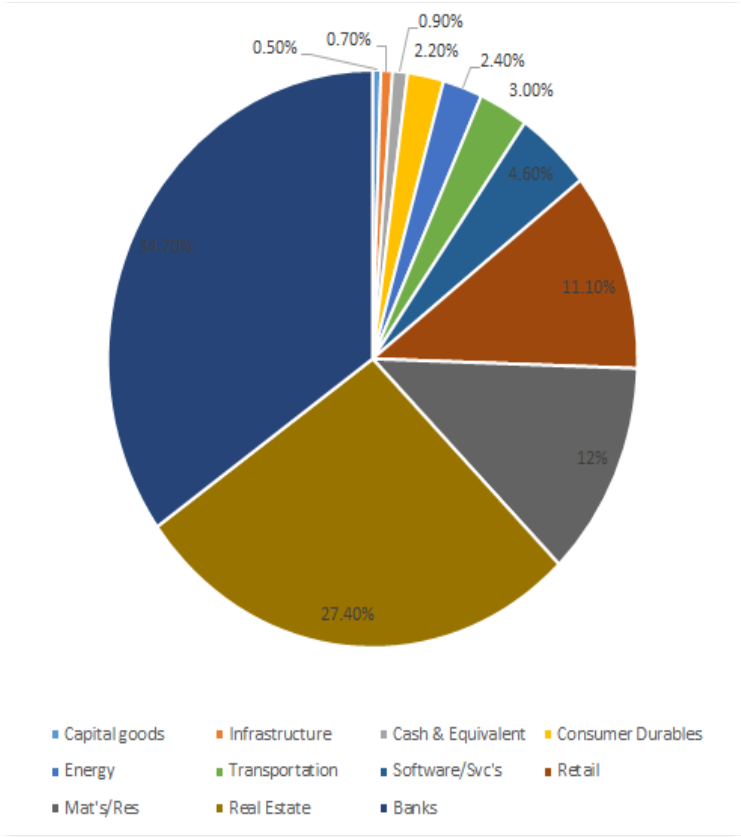
1.Kim-Georg Aase (2011), “Text Mining of News Articles for Stock Price Predictions”,
Master Thesis. Norwegian University of Science and Technology, Department of Computer and Information Science.

Zhichao Han (2012), “Data and text mining of financial markets using news and social media”,
Master thesis, The University Of Manchester.

1.Simon Bacher (Oct 2012), “Mining Unstructured Financial News to Forecast Intraday Stock Price Movements”,
Master Thesis, University Mannheim.

Brett DRury(2014), “A Text Mining System for Evaluating the Stock Market's response To News”,
Doctoral Program in Computer Science of the Universities of Minho, Aveiro and Porto.

Overview of bank share



The "KING" stock

Measurement of bank share price volatility

Absolute price volatility $\Delta_P = P_{i,t} - P_{i,t-1}$

Relative price volatility $\%\Delta_P = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t}}$

Price volatility ratio $\Delta_P = \frac{P_{i,t}}{P_{i,t-1}}$

FACTORS AFFECTING BANK STOCK PRICES

Macroeconomic factors

- GDP: Mehr-un-nisa and Mohammad Nishat (2012)
- Inflation: Mohammed Omran and John Pointon (2001).
- Interest rate: Mahmudul and Salah Uddin (2009)
- Central bank money supply: Chen Shaoping (2008)
- Exchange rate: Noel Dilrukshan Richards and John Simpson
- Production Index: Author George Filis (2009)

Microeconomic factor

- Bank operating hours: Liargovas and Skandalis (2008)
- Bank size: Amarjit and colleagues (2010)
- Profitability ratio: ROA, ROE: Abu Hashi, K. (2003)
- Price to Earning Ratio: Khan and Amanullah (2012)
- Non-performing loan ratio, NPL:he definition of bad debt of Vietnam in the Law on Credit Institutions 2010
- Earning per shares (EPS): Uddin study & Rajib Hossain (2013)

Psychology factor

Overconfidence: Qureshi et al. (2012) and Bashir et al. (2013)

A-Excessive :
Gervais et al. (2002)

Herd mentality:
Biekhchandani and Sharma. 2001

psychology of risk:
Olsen (2007, 2008)

Excessive pessimism

Data mining in predicting bank stock prices

Efficient market hypothesis

The efficient market hypothesis (or efficient market theory) is a hypothesis of financial theory that asserts that market.

Three versions

WEAK

stock prices fully reflect information published in the past

SEMI-STRONG

stock prices have been fully influenced by both previous disclosures and those have just been disclosed

STRONG

prices represent all past and present publicly available information, as well as internal information

Efficient market hypothesis

The efficient market hypothesis (or efficient market theory) is a hypothesis of financial theory that asserts that market.

EFFICIENT MARKETS DO NOT ALWAYS APPLY WHEN ??

- Small Firm Effect: Small firms have unusually high returns over a long time.
- Reversal: stocks are showing low returns today tend to offer high returns in the future, and vice versa. again.
- Market overreaction:Stock prices often overreact to the latest information and valuation errors are adjusted slowly.
- The January effect:
  Over a long time, stock prices tend to rise abnormally high from December to January

Machine learning

Machine gaining knowledge is an AI software (AI) that offers structures the ability to robotically examine and enhance informationat the same time as now no longer being expressly programmed.

SUPERVISED LEARNING

Input & Output
Data
Classification
regression

Predictions &
Predictive modes

UNSUPERVISED LEARNING

Input
Data
Clustrering
Association

Patterns / Structure
Discovery

Text mining

The changing unstructured text into a structured format process to find patterns with new meanings and insights is known as text mining.

DATA

- STRUCTURED
- UNSTRUCTURED
- SEMI-STRUCTURED

1960s
First introduced

1990s
Became more popular

FOREX EXCHANGE RATE FORECAST

- Historical trends (C. Goodhart, 1989)
- News reports (GPC funf, 2002)
- Historimacro articles (MDD evans, 2008)
- Even Twitter (TTVu, 2012)

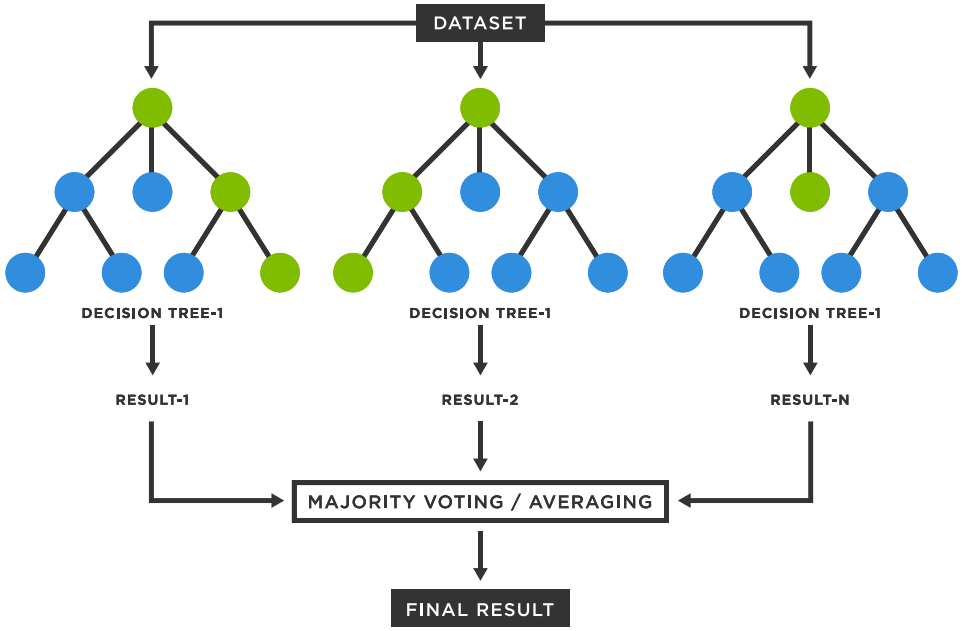
Decision tree

In the field of machine learning, a decision tree is a type of predictive model, that is, a mapping from observations about a thing/phenomenon to conclusions about a target value of the object/phenomenon

SHOULD I ACCEPT A NEW JOB OFFER ?

Random Forest

Random Forest
is a supervised machine learning algorithm.
This is an algorithm built on Decision Trees
for predictive modeling and behavioral analysis.

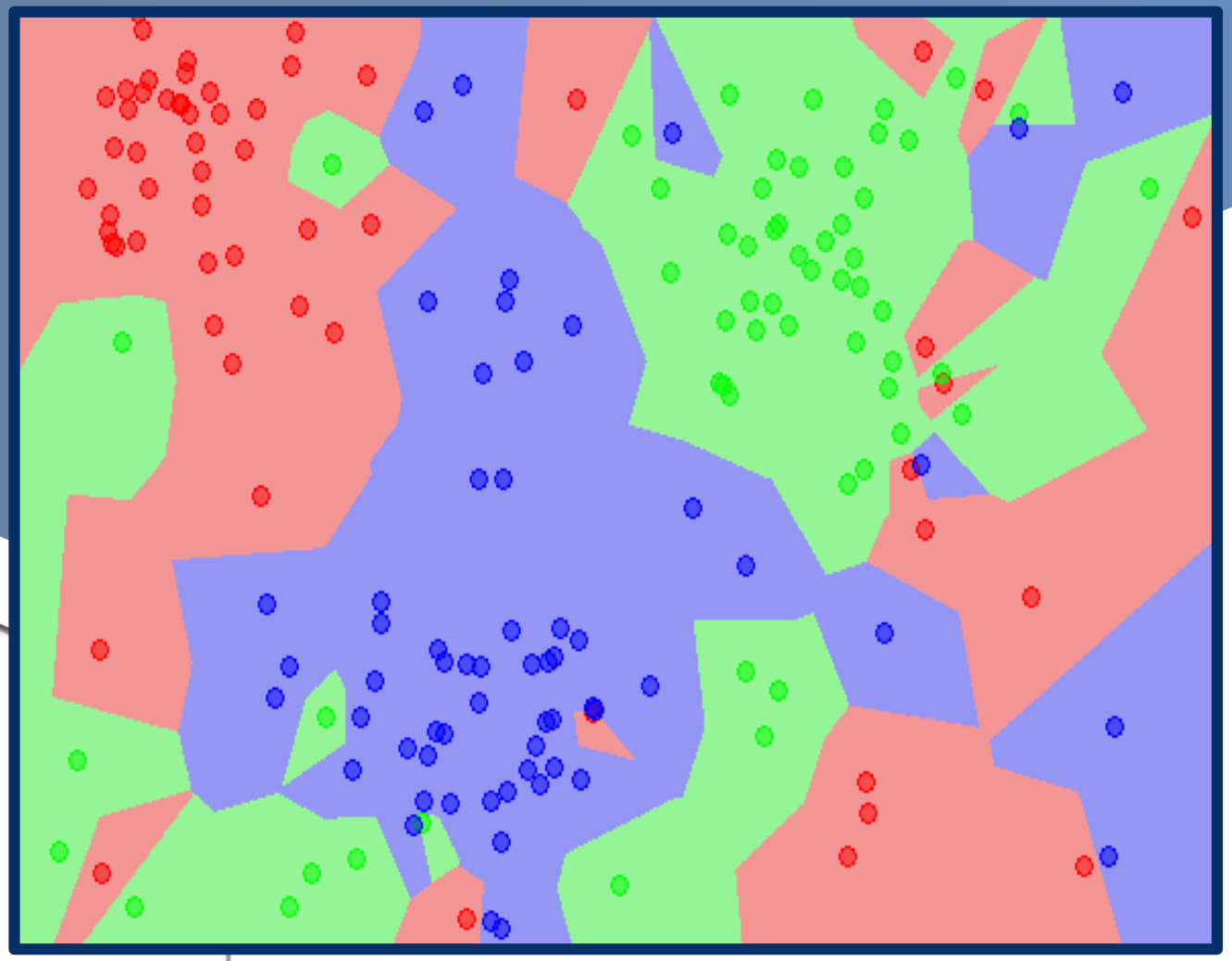


Advantage of random forest

- Advantage of **giving estimates** of variable importance
- Provide an out-of-the-box method for **working with missing data**
- **Handle big data** with many variables up to thousands
- Handles variables **quickly**, suitable for complex tasks

K-nearest neighbor

K-nearest neighbor is
one of the handiest supervised basic algorithms
in Machine Learning.

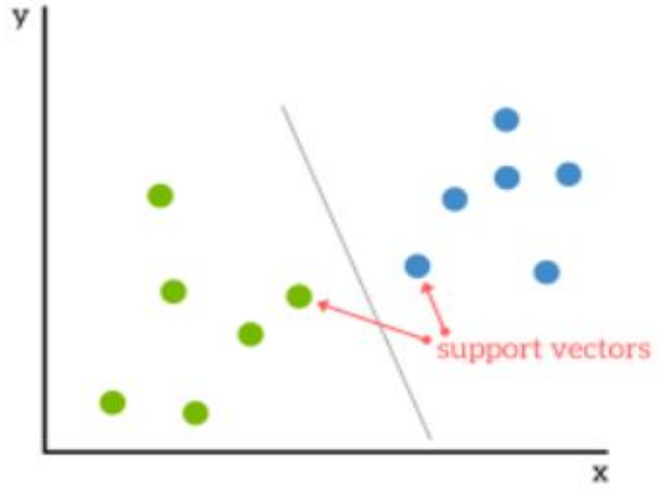


K-nearest neighbor may be implemented for

- Supervised recognize-problems
- Classification
- Regression

Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learningthat can be used for classification and regression.



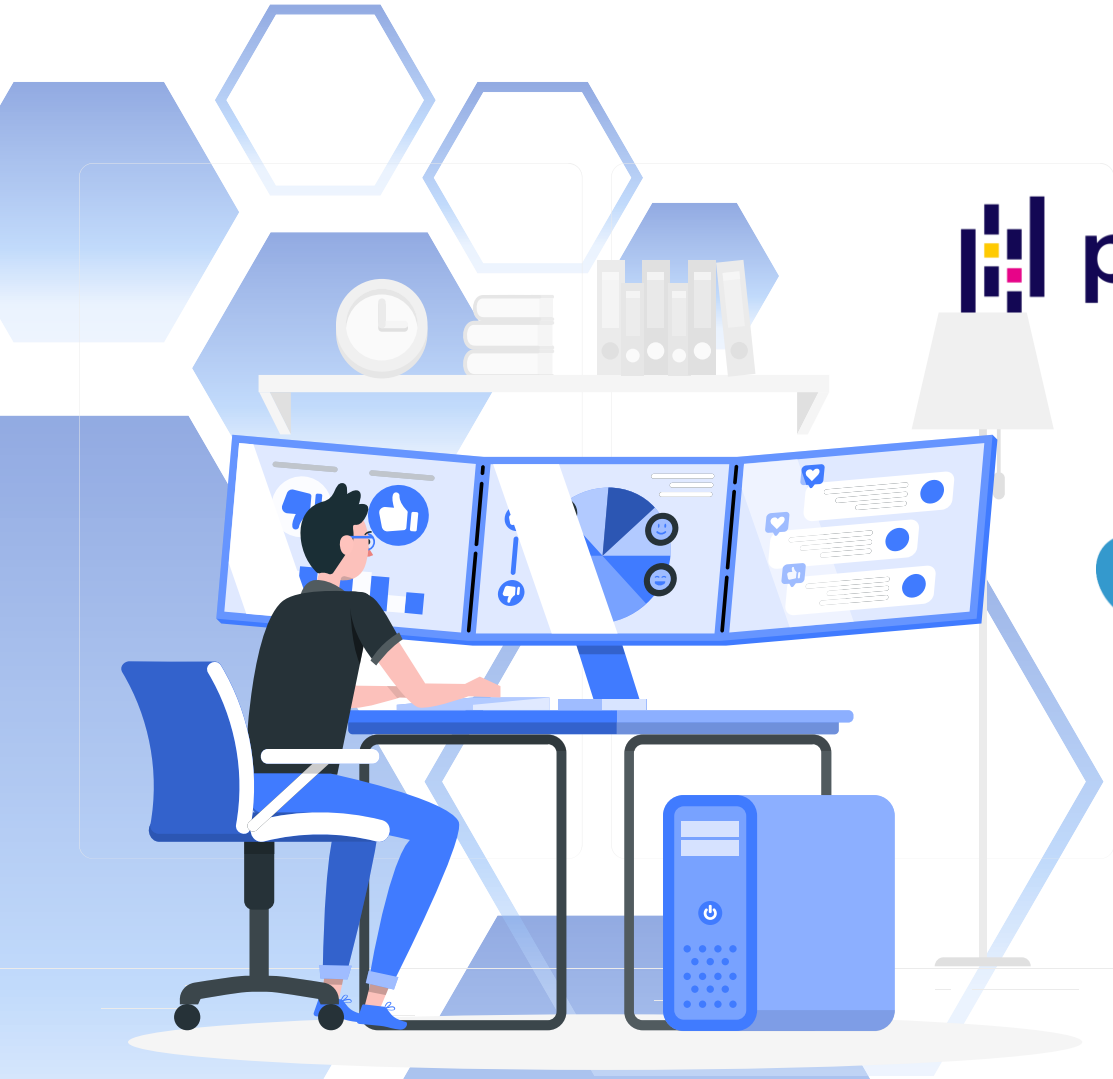
Test Error Minimisation

Text classification

Unknown topic

- known text classes
- known text classes

Support tools using on research

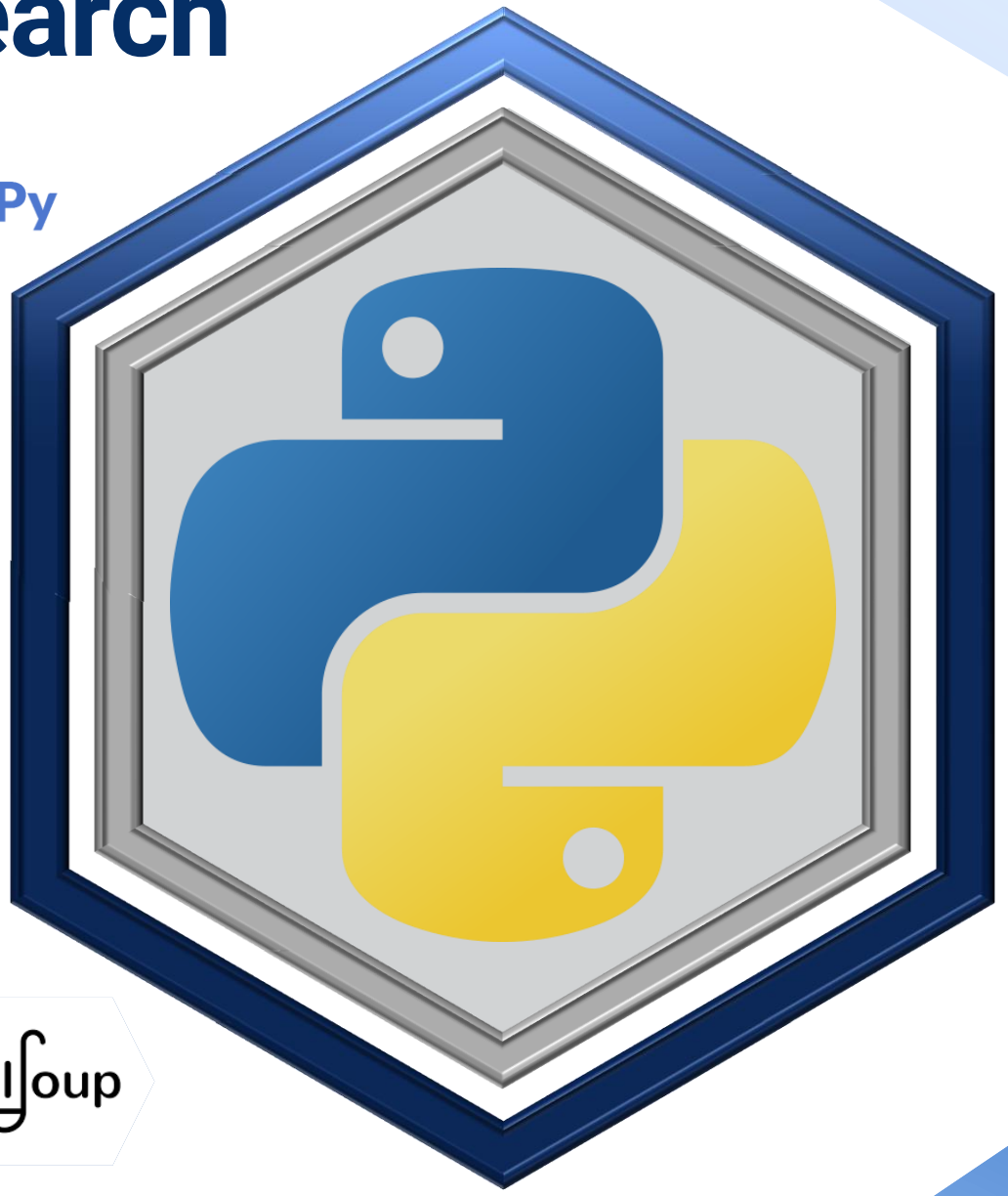


NumPy

pandas

scikit learn

Beautiful Soup

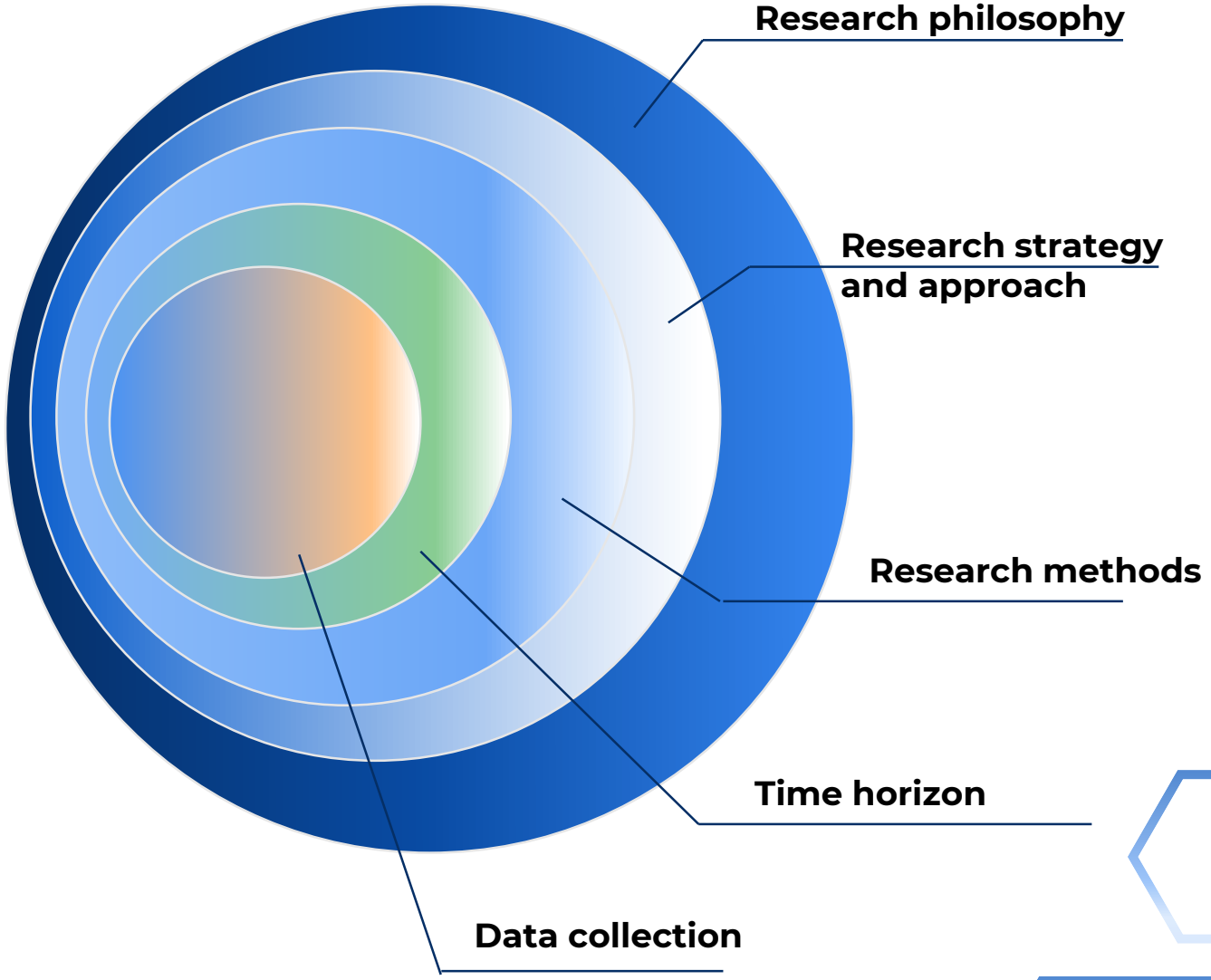


CONCLUSION OF CHAPTER 2


CHAP 3
METHOLOGY

Research onion

1. **Research philosophy**
   Positivist
2. **Research strategy and approach**
   Case study
3. **Research methods**
   Quantitative and qualitative
4. **Time horizon**
   Cross-sectional
5. **Data collection**
   Secondary data



Research design

Data source

News relevant to bank shares

History of bank share prices

Data analysis process

Remove unnecessary characters

separate words

Elimination of stop word

Combine new

Labeling news lean on history stock prices

Representation of articles in vector space

Running model

Training data

Machine model: SVM, K-neihgbor, Random forest, Decision tree

Test data

Value efficiency

Data sources

Bank news

Stock market news

January 2014 - August 2021

23.879 articles

	Time	Total of news
CafeF cafef.vn	23/1/2014 - 2/8/2021	10989
VnExpress Tin nhanh Việt Nam	23/1/2014 - 2/8/2021	2887
Báo Thanh Niên	3/4/2018 - 2/8/2021	2100
Stockbiz	2/1/2017 - 2/8/2021	7903

Web scraping

News

websites pages,
Unstructured Data

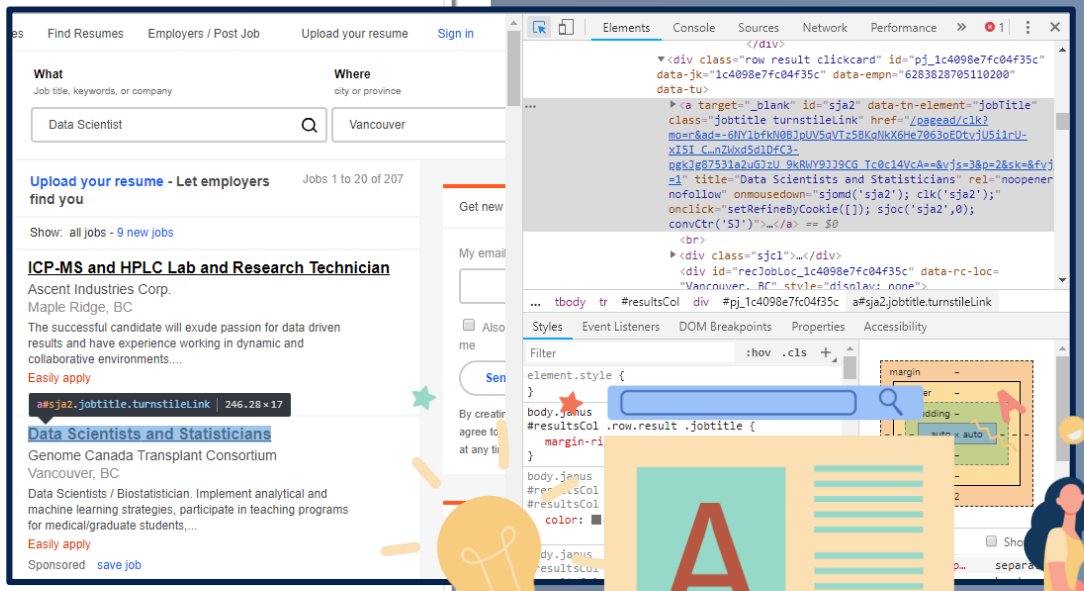
web scraping/
Data Extraction

Folder year

Folder months

Folder day

File News in one day



History stock prices

Cophieu68.vn

BID
TCB
TPB

Data analysis process

Step 0

Thêm người nhà Sếp ngân hàng bị phạt vì mua bán chứng khoán không công bố thông tin

01/06/2021người_nhà_sếp ngân_hàng phạt mua_bán chứng_khoán không công_bố thông_tin

khoán không công bố thông tin

khoán không công bố thông tin

chứng_khoán không công_bố thông_tin

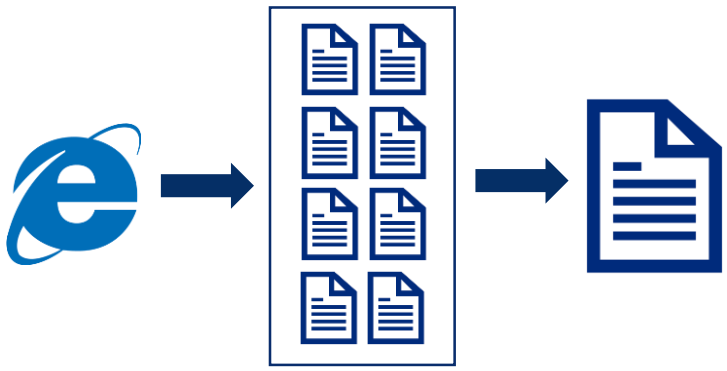
Step 04

Eliminate stop words (a ha, amen, ...)

thêm người_nhà_sếp ngân_hàng bị phạt vì mua_bán chứng_khoán không công_bố thông_tin

Step 5: Combine news

STEP 5.1



Sigle news of one day

Total news of one day

Cafef.vn

STEP 5.2

Cafef.vn

thanhnien.vn

Vnexpress.net

Stockbiz.vn

Total new of one day (4 website)

Labeling news to prepare data for the training phase

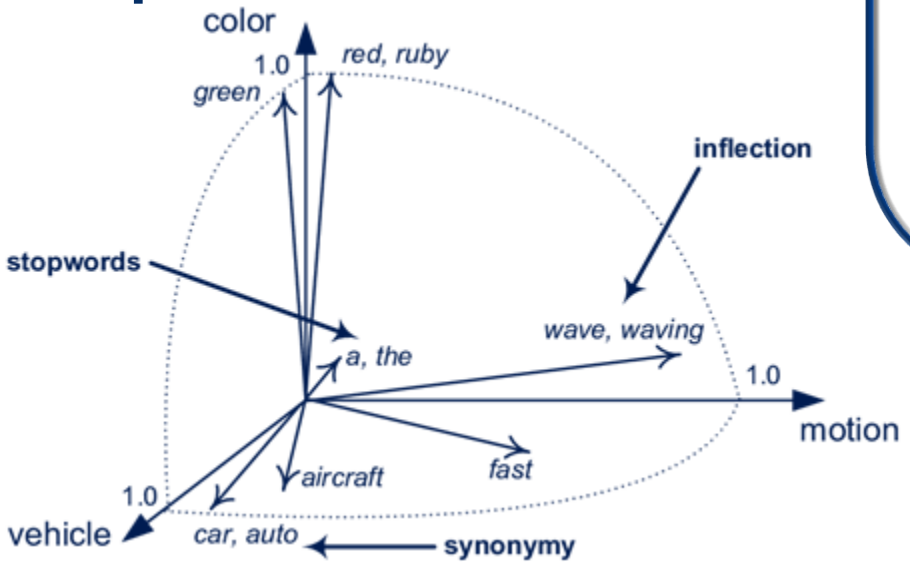
Good 1

Bad -1

Medium 0

01/06/2021người_nhà_sếp ngân_hàng phạt mua_bán chứng_khoán không công_bố thông_tin1

Representation of articles in vector space



Run model stage

Day	Text	Change
1	a	1
2	b	-1
3	c	0
...	...	...
N	n	→ ? (1,-1,0)

MACHINE LEARNING
4 ALGORITHM

K-neighbor
SVM
Random forest
Decision tree

Input train

Day	Text
1	a
2	b
3	c
...	...
100	g

Output train

Change
1
-1
0
...
1

Model

Input

Day	Text
N	n

Output

Change
1

MACHINE LEARNING MODEL

- K-neighbor
- SVM
- Random forest
- Decision tree

$$Accuracy = \frac{TU + TD + TN}{TU + TD + TN + FU + FD + FM}$$

TU: Is the correct forecast number for the uptrend

TD: Is the correct forecast number for the downtrend

TN: Is the correct forecast number for the trend not to increase, not to decrease

FU: Is the number of false predictions for an uptrend

FD: Is the number of false predictions for a downtrend

FN: Is the number of false predictions for a trend that is not increasing, not decreasing

STEP 1: 70%: 30%

Day	Text	Change
1	a	1
2	b	-1
3	c	0
...	...	...
100	g	1

70% tranining

X train: day, text (A, B, C,...., N): **Input**

Y train: change (1,0,-1): **Output**

30% testing

X test: day, text (A, B, C...., N): **Input**

Y test : change (1,0,-1)

STEP 2,3,4: Building and testing model

Step 2: Built model
Step 3: Testing model
Step 4: Caculated accurency of model

70% tranining
X train
Y train

Step 2

SVM, K-neighbor
Decision tree, Random forest

→ Model

30% testing
X test

Step 3

Model of SVM, K-neighbor
Decision tree, Random forest

→ Result

↓ Compare with Y test

Step 4

Accuracy of Model (SVM, K-neighbor, Decision tree, Random Forest)

Compare with Y test

	Test result	Y test	
Text A	1	1	✓
Text B	1	-1	✗
Text C	-1	-1	✓
Text D	0	1	✗

✓ **Correct** ✗ **Incorrect**

Correct forecast number for the uptrend (TU)

Correct forecast number for the downtrend (TD)

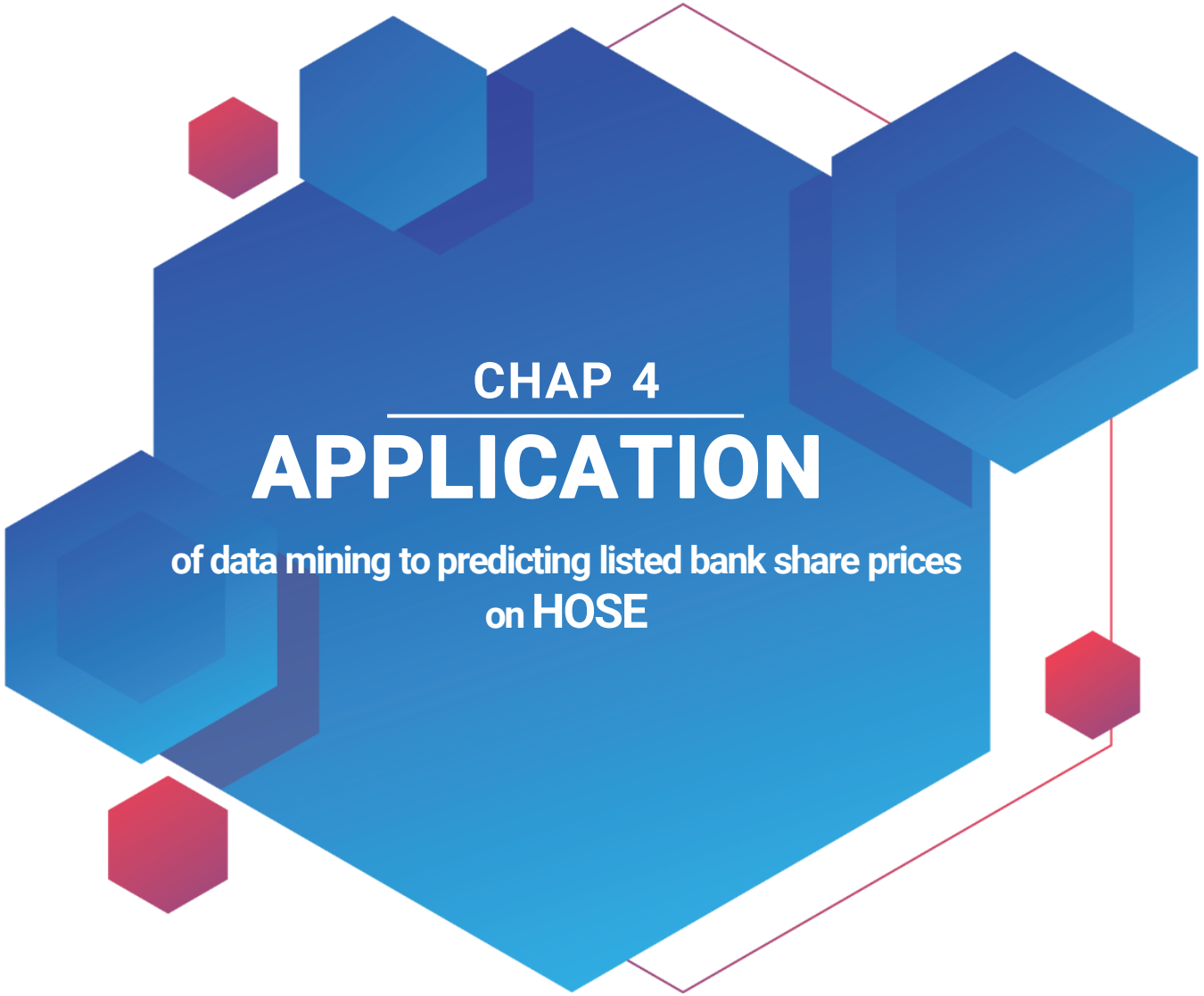
Correct forecast number for no change (TN)

The number of false prediction for the uptrend (FU)

The number of false prediction for the downtrend (FD)

The number of false prediction for no change (FN)

$$Accuracy = \frac{TU + TD + TN}{TU + TD + TN + FU + FD + FM}$$



CHAP 4

APPLICATION

of data mining to predicting listed bank share prices on HOSE

Overview of listed banks on HOSE

- Listing and issuance situation,
- Descriptive statistics,
- Moving average of 03 consecutive quarters
- Trends and seasonality

2014-2021.

Stock ID	First trading date	First trading day	First listed volume	Listed shares	Shares outstanding
ACB	12/09/2020	130.200	110.004.656	2.161.558.460	2.701.948.075
BID	24/01/2014	18.800	2.811.202.644	4.022.018.040	4.022.018.040
CTG	16/07/2009	40.100	121.211.780	3.723.404.556	3.723.404.556
EIB	27/10/2009	29.000	876.226.900	1.235.522.904	1.229.432.904
HDB	05/01/2018	39.600	980.999.979	1.608.848.818	1.593.767.296
LPB	09/11/2020	14.200	646.000.000	1.074.638.915	1.074.638.915
MBB	01/11/2011	13.800	730.000.000	2.798.756.872	2.798.756.872
MSB	23/12/2020	17.000	1.175.000.000	1.175.000.000	1.157.000.000
STB	12/07/2006	78.000	189.947.299	1.885.215.716	1.803.653.429
TCB	04/06/2018	102.400	1.165.530.720	3.504.906.230	3.504.906.230
TPB	19/04/2018	32.450	555.000.000	1.071.671.722	1.071.671.722
VCB	30/06/2009	60.000	112.285.426	3.708.877.448	3.708.877.448
VIB	10/11/2020	18.500	564.442.500	1.109.387.852	1.553.142.993
VPB	17/08/2017	39.000	1.405.908.635	2.456.748.366	2.456.748.366

Table: 4.1 Overview of listed banks in HOSE

Descriptive statistics of listed bank share prices

Bank	Stock ID	Maximum	Minimum	Mean	Standard Deviation
Asia Commercial Bank	ACB	46,7	15,20	24,74	7,96
JSC Bank For Investment And Development Of Vietnam	BID	47,90	12,70	28,06	11,55
Joint Stock Commercial Bank for Industry and Trade	CTG	52,7	13,80	22,28	8,61
Vietnam Commercial Joint Stock Export Import Bank	EIB	30,3	11,65	16,46	4,29
Ho Chi Minh City Development Joint Stock Commercial Bank	HDB	45,9	17,80	30,14	7,04
LienViet Post Joint Stock Commercial Bank	LPB	29,95	5,60	12,73	6,25
Military Commercial Joint Stock Bank	MBB	43,35	13,10	19,88	7,01
Vietnam Maritime Commercial Join Stock Bank	MSB	30,2	18,80	23,62	5,90
Sai Gon Thuong Tin Commercial Joint Stock Bank	STB	30,6	7,30	14,24	4,83
Vietnam Technological and Commercial Joint Stock Bank	TCB	91,7	15,00	32,41	20,34
Tien Phong Commercial Joint Stock Bank	TPB	36,75	17,15	24,14	4,98
Bank for Foreign Trade of Vietnam	VCB	116,4	26,10	57,75	23,62
Vietnam International Commercial Joint Stock Bank	VIB	49,45	12,80	25,48	10,90
Vietnam Prosperity Joint Stock Commercial Bank	VPB	67,7	16,95	31,60	15,84

Table 4.2 Descriptive statistics of listed bank share price(Q2.2014-Q2.2020)

highest-closing-price banks

Highest share price (VCB)

Moving average of 03 consecutive quarters of listed bank stock prices during 2014-2021

Data

- **Large group** (BID, CTG, VCB)
- **Medium group** (TCB, MBB, ACB, STB, HDB)
- **Small group** (VIB, TPB, MSB, LPB, EIB)

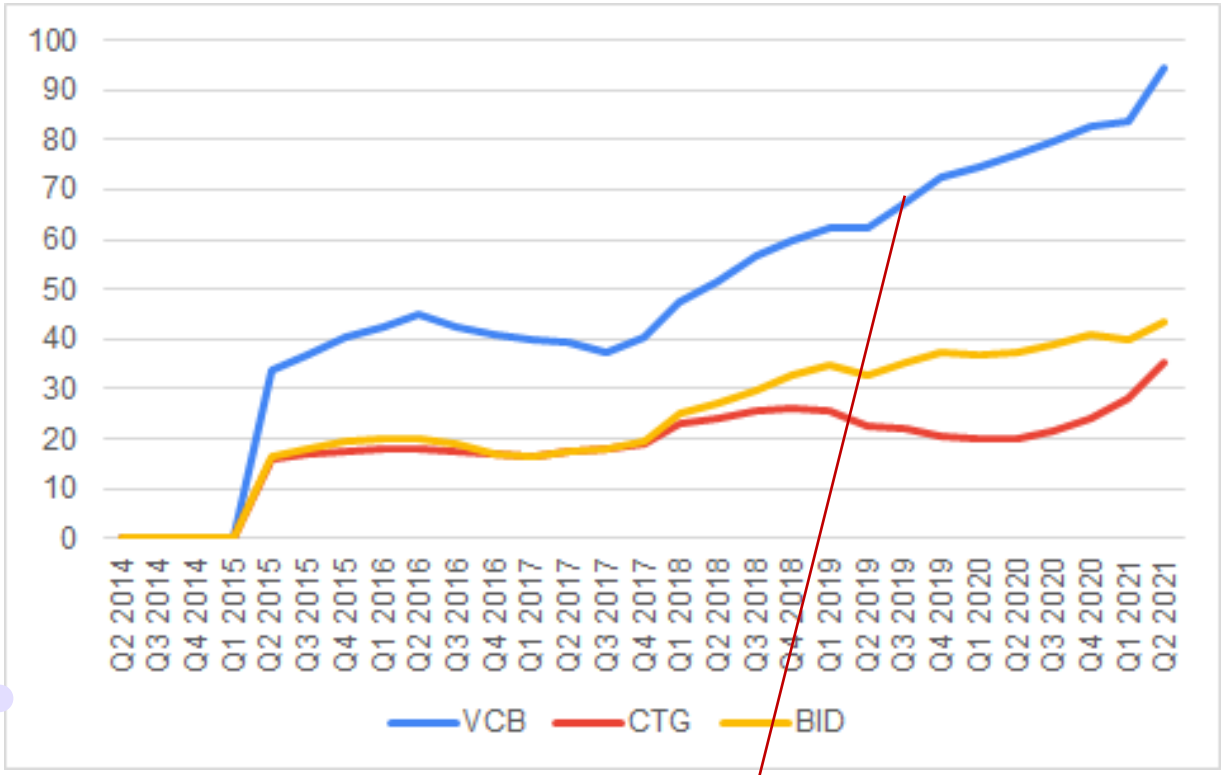


Figure 4.1 Moving average share prices of the three big banks listed on HoSE

VCB share fluctuates strongly and have highest price level

The dramatic increase in volume was ACB and peaked in the middle of 2018 with a closing price of 36,700 VND.

MBB had the highest closing price in the middle of 2018 at 26,000 VND.

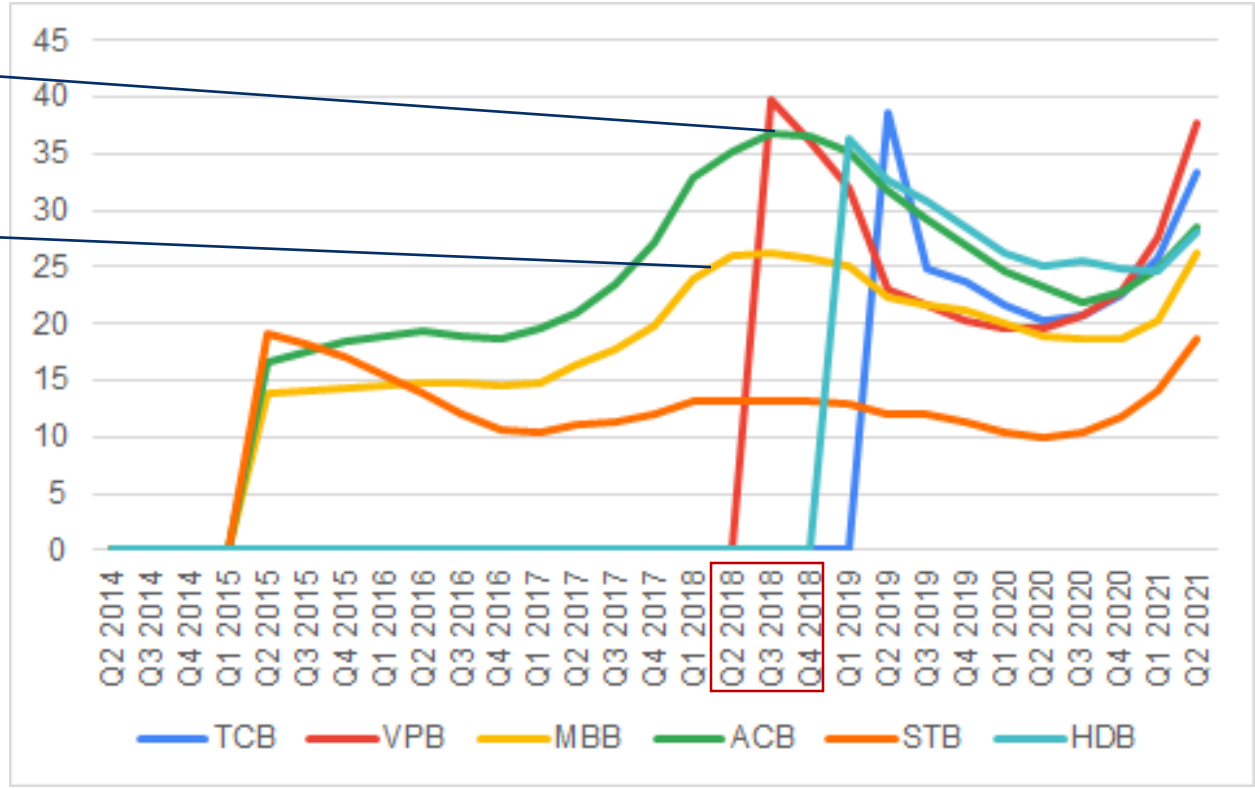


Figure 4.2 Moving average share prices of medium banks listed on HoSE

The series has achieved a non-uniform motion state.

TPB is the smallest bank but has the highest moving average in mid-2018 with a closing price of 36,000 VND.

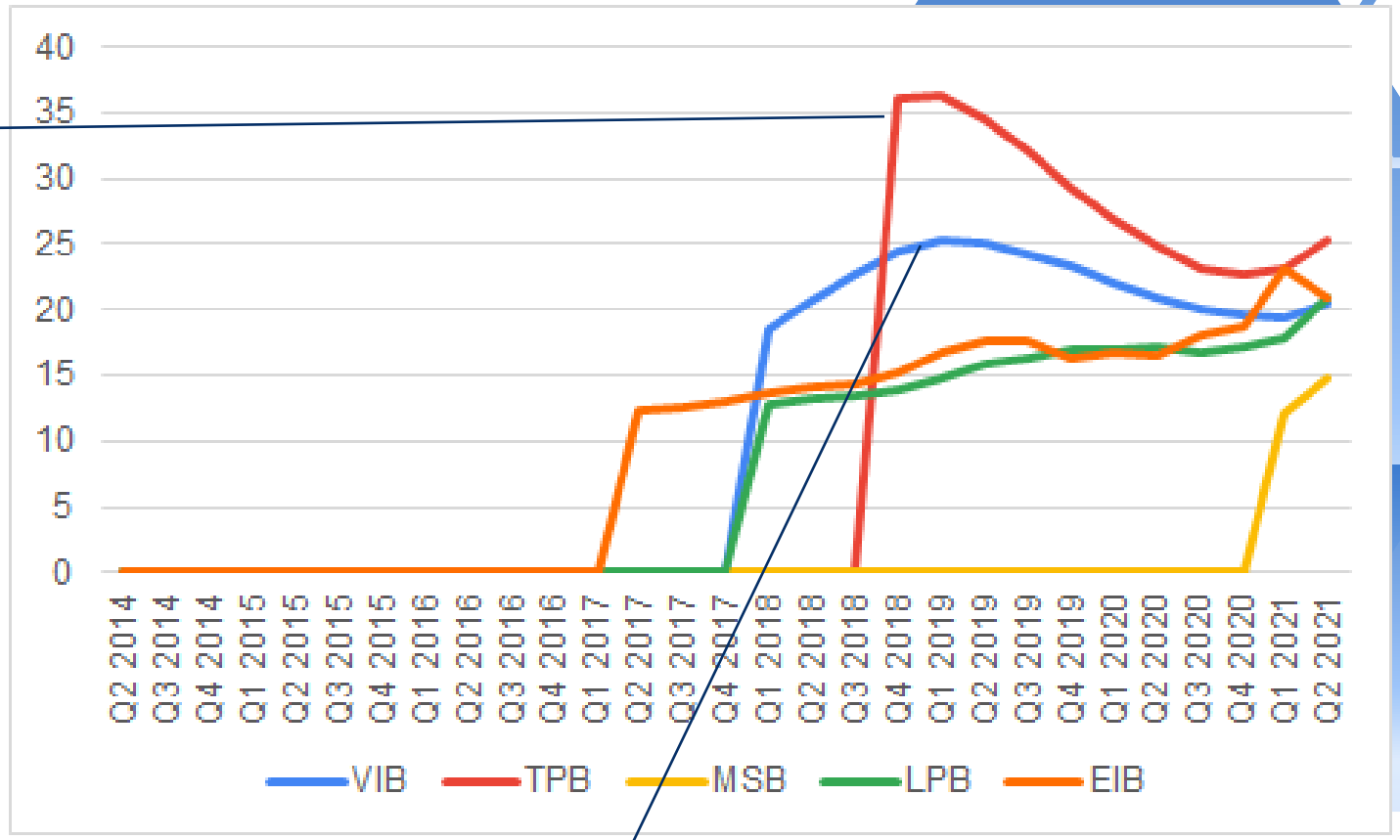


Figure 4.3 Moving average prices of small banks listed on HoSE

The 2nd highest moving average out of 5 small banks, belongs to VIB with the highest price in early 2019

Trends and seasonality of listed bank shares on HOSE

2014 - 2021

Over 4 quarters since being established, the HOSE market has achieved impressive results in trading volume as well as number of investors.

The transaction volume of VCB tended to decrease in the last 3 quarters of the year. Especially fell sharply in the last quarter when it dropped to nearly 100 million VND.

Stock ID	Q1	Q2	Q3	Q4
VCB	-0.69777	1.28467	-1.87247	1.28557
CTG	0.14368	-0.57894	0.88534	-0.45007
BID	0.36994	0.90119	0.24226	-1.51339
TCB	-1.99092	-2.43973	-4.30759	8.73824
VPB	-1.59583	0.00833	5.40833	-3.82083
MBB	0.10104	-0.51414	0.11622	0.29687
ACB	-0.94048	-0.5619	1.19881	0.30357
STB	0.14054	-0.37429	-0.07429	0.30804
HDB	0.62403	-2.57656	1.65558	0.29695
VIB	0.34539	-2.46949	2.67872	-0.55461
TPB	0.98437	-0.61741	-2.15134	1.78437
MSB		0.94554	0.30268	-0.28839
LPB	0.04911	0.80625	-1.26696	0.41161
EIB	-0.46823	-0.09635	-0.53385	1.09844

Table 4.4. Shows the trends and seasonality of the 14 banks listed on the HOSE

APPLICATION OF TEXT MINING IN PREDICTING THE PRICES OF BID, TCB AND TPB

Some function of the module using in research

Stt	Name of module	Funtion
1	**Download news**	Use Beautiful soup library to take content from html tag of web news and history data price from cophieu68.com
2	**Data-processing**	Use Word_tokenizer of Pyvi library. Pyvi is a popular library on Vietnamese language processing
3	**Process machine model**	Scikit-learn is probably the most useful library for machine learning in Python. The Sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

SPLIT DATA

- 70% training model
- 30% testing model

```python
def training(bank, file):
    df = pd.read_csv(f'4. training-data/{bank}/{file}-{bank}.txt')
    df = df[df['Change'] != 2]

    feature_extraction = TfidfVectorizer()
    X = feature_extraction.fit_transform(df["text"].values)
    y = df["Change"].values

    # chia dữ liệu 70% training 30% test
    X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=0.3, random_state=1)
```

```python
clf_SVC = SVC(probability=True, kernel='rbf', C=0.01)
clf_SVC.fit(X_train, y_train)
predictions_SVC = clf_SVC.predict(X_test)
print(f"{bank},{file} SVM: ", accuracy_score(y_test, predictions_SVC))
```

Example of fit data on SVM

BEST MODEL WITH BID, TCB, TPB SHARE PRICES

Best model wirh BID

Best model wirh TCB

Best model wirh TPB

	Model	Accuracy
BID	SVM	45.5%
	K-neghbor	36.69%
	Dicision Tree	43.71%
	Random forest	44.42%
TCB	SVM	38.49%
	K-neghbor	44.77%
	Dicision Tree	47.79%
	Random forest	48.12%
TPB	SVM	42.32%
	K-neghbor	38.31%
	Dicision Tree	40.32%
	Random forest	41.12%

Table 4.7. Test result with BID, TCB, TPB share prices

BEST WEBSITE WITH BID, TCB, TPB SHARE PRICES

	Time	The total of data source	Accurate
StockBiz	2/1/2017 - 2/8/2021	7903	52,55%
Vnexpress	23/1//2014 - 2/8/2021	2887	43,93%
Cafef	23/1//2014 - 2/8/2021	10989	44,77%
Thanhnien	3/4/2018 - 2/8/2021	2100	44,94%

Stockbiz is the best website

Table 4.8. Accurate with each web use model SVM to BID

Vnexpress is the best website

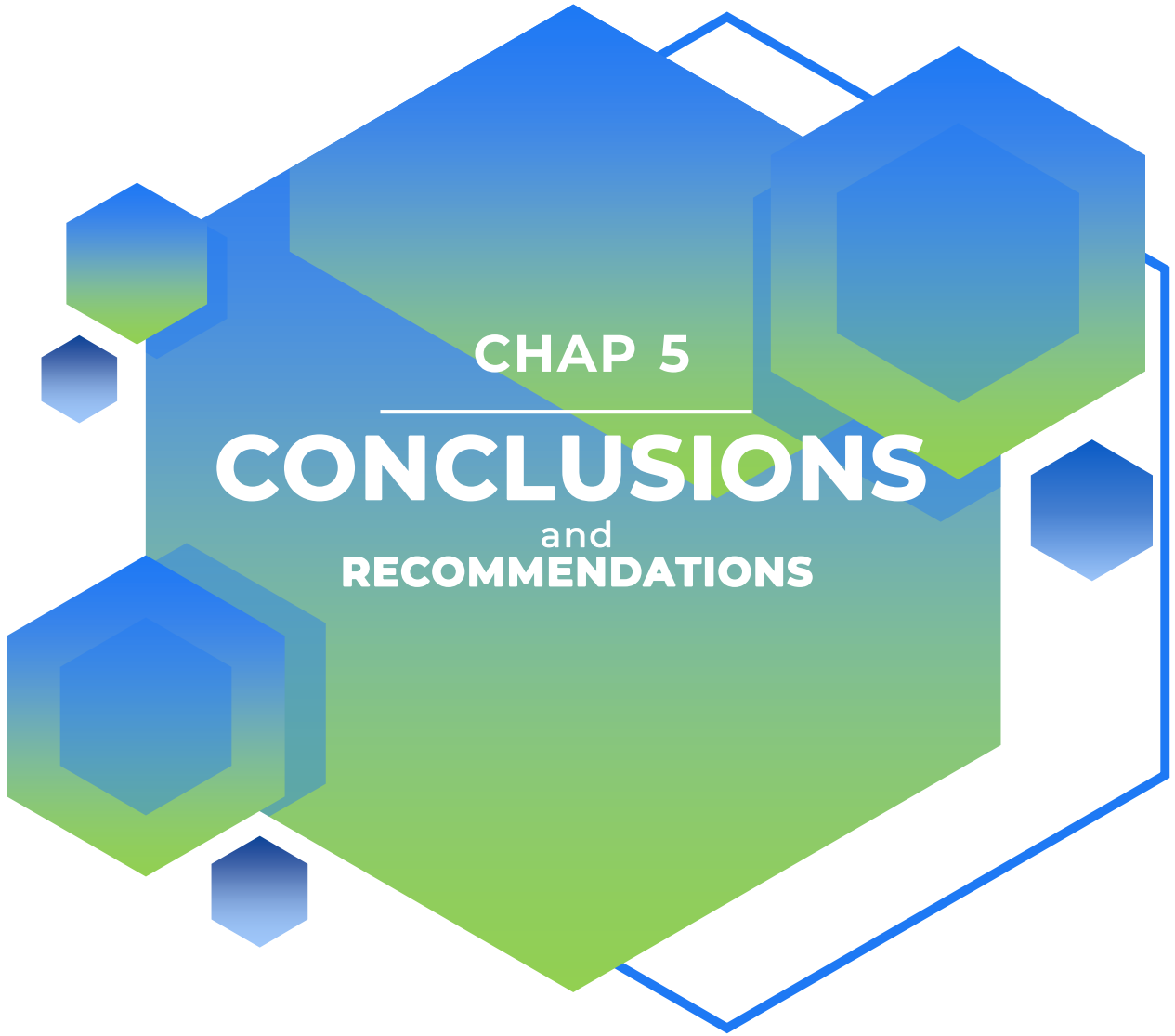
	Time	The total of data source	Accurate
StockBiz	2/1/2017 - 2/8/2021	7903	44,17%
Vnexpress	23/1//2014 - 2/8/2021	2887	52,94%
Cafef	23/1//2014 - 2/8/2021	10989	43,39%
Thanhnien	3/4/2018 - 2/8/2021	2100	48,31%

Table 4.10. Accurate with each web use model SVM to TPB

	Time	The total of data source	Accurate
StockBiz	2/1/2017 - 2/8/2021	7903	49,78%
Vnexpress	23/1//2014 - 2/8/2021	2887	43,63%
Cafef	23/1//2014 - 2/8/2021	10989	52,36%
Thanhnien	3/4/2018 - 2/8/2021	2100	52,60%

Table 4.9. Accurate with each web use model Random forest to TCB

Thanhnien is the best website


CHAP 5
CONCLUSIONS
and
RECOMMENDATIONS

Q1: How does data mining affect the valuation models of bank shares in Vietnam's stock market?

Q2: How does data mining affect investors' decisions in Vietnam's stock market?

Q3: What are the recommendations for managers and investors

SUMMARY OF FINDINGS

CafeF cafef.vn | Báo Thanh Niên | Stockbiz | VnExpress

23.879 articles

January 2014 - August 2021

BID TCB TPB

DATA

Cleaning the data

Labelling the article based on the trending of stock prices

Divided into training data (70%) and testing data (30%)

Representation of articles in vector space

Bulit model and testing model

Pros and cons of using data mining in the bank stock prediction

PROS

52.94%

New join stock market invester

Institutional investor

Bank manager

CONS

Vietnamese is more complicated than English to processing

Not many Vietnamese data processing methods

The fomat of some websites is difficult to scraping

As in the study, we can see that on each type of bank (large, medium, and small), the performance of each model is different, the performance for the web is also different

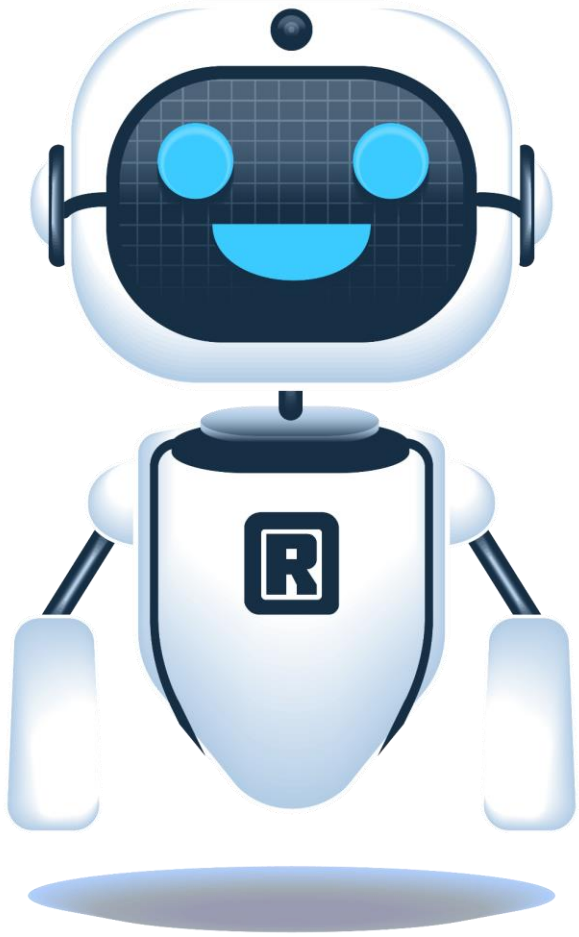
Disadvantage of the program

38.31% - 52.94%

Using the stock news website has not provided sufficiently a comprehensive view of vietnamese stock news.

Number of news used as inputs for the program maybe not large enough

Not convenient to fit data everyday,
the accuracy of model can be decrease day by day



At 0 am

Scrap all news on the day before

Delete the oldest data

Build a new model for the current day

Predict share prices for the current day

RECOMENDATION

Pay more attention to the impact of market information on bank share prices
News that occurs in 24 hours is especially important

Cultivate new methods such as data mining to choose the most suitable method for each market stage.

Individual investors

Select reliable sources that is the most appropriate for the selected bank stock

selecting the methods and evaluating approach is of importance

Bank manager

Applying data text mining in forecasting the market situation

Institutional investor

need to invest in new technology and forecast methods.

Adding data mining as an ecosystem of methods

Suggestions for further research

Expanding scope of research to 14 banks stocks listed on HOSE, and step by step increase the research scope into all the banks trading and listing on stock exchanges.

Expand the sources of news to the top 10 popular news websites about finance

plus opinions of top financial experts and consulting companies in the market.

develop the Vietnamese language processing program and integrate them into Python library

Add the model identifying adjectives, adverbs, and verbs to the program.