|  |
| --- |
| VIETNAM NATIONAL UNIVERSITY, HANOI  **INTERNATIONAL SCHOOL**  **GRADUATION PROJECT**  **PROJECT NAME: SENTIMENT ANALYSIS AND ITS CORRELATION WITH STOCK PRICE MOVEMENTS IN VIETNAM’S FINANCIAL MARKETS**  **Student’s name**  **NGUYEN THI THUY LINH**  ­­­­­­  *Hanoi - 2025* |
| VIETNAM NATIONAL UNIVERSITY, HANOI  **INTERNATIONAL SCHOOL**  **GRADUATION PROJECT**  **PROJECT NAME: SENTIMENT ANALYSIS AND ITS CORRELATION WITH STOCK PRICE MOVEMENTS IN VIETNAM’S FINANCIAL MARKETS**  SUPERVISOR: PhD. TRUONG CONG DOAN  STUDENT: NGUYEN THI THUY LINH  STUDENT ID: 21070628  COHORT: QH-2021-Q INS4011  MAJOR: BUSINESS DATA ANALYTICS  *Hanoi - 2025* |

**ACKNOWLEDGEMENT**

Throughout the creation of this project and the related study, " Sentiment Analysis And Its Correlation With Stock Price Movements In Vietnam’s Financial Markets " ,I am truly appreciative of Mr. Truong Cong Doan's extraordinary knowledge, direction, and unwavering support. My ability to overcome obstacles and improve the caliber and inventiveness of my work has been greatly aided by his wise counsel, unwavering support, and considerate criticism.

I am incredibly appreciative of my family's tolerance, love, and continuous support at this difficult time. Their confidence in my skills has consistently given me strength, particularly during times of uncertainty and fatigue.

Lastly, I would want to thank the larger academic community for laying the groundwork for this study through their groundbreaking papers and work in the domains of robotics, deep reinforcement learning, and control systems. Their input has greatly influenced and directed the course of my study.I want to sincerely thank everyone who has helped me (directly or indirectly) to make this project a reality. This success is a result of your faith in me.

**STATEMENT OF ORIGINALITY**

Under the direction and supervision of Mr. Truong Cong Doan, I thus certify that this thesis is the product of my own independent effort. The discoveries and information provided here are unique and the result of my own work. In the references section, all analyses, assessments, and interpretations derived from outside sources have been appropriately referenced and recognized.

In compliance with academic standards, all materials—including reviews, comments, and data—that were obtained from other authors or organizations have been properly cited and acknowledged.

I take full responsibility for any plagiarism or academic dishonesty found in this thesis. The Vietnam National University International School in Hanoi disclaims all responsibility for any possible copyright violations resulting from this work.

|  |  |
| --- | --- |
|  | *Ha Noi, 6/20/2025*  Author  Linh  Nguyen Thi Thuy Linh |

**ABSTRACT**

The use of sentiment research in financial modeling has been increasingly popular in recent years, particularly in emerging markets where investor behavior is extremely influenced by news and public opinion. This study examines, with an emphasis on the banking industry, the relationship between sentiment gleaned from financial news and changes in stock prices within Vietnam's financial markets. A sizable corpus of news stories on banking institutions written in Vietnamese are used to generate sentiment indices (positive, neutral, and negative) using sophisticated Natural Language Processing (NLP) techniques.

The study examines the impact of these sentiment indicators on trading volume, stock price volatility, and overall market behavior by combining them with historical stock market data. The study aims to find predictive correlations between news sentiment and transient market swings using statistical techniques and machine learning models. In addition to assessing the potential of sentiment-driven signals for forecasting and investing decision-making, the objective is to comprehend how sentiment affects market dynamics.

The results show that sentiment, especially bad news, has a quantifiable impact on price patterns and trading behavior in the Vietnamese stock market. These findings emphasize how crucial it is to incorporate qualitative data into financial analysis and show how sentiment analysis may be a useful tool for analysts, investors, and policymakers. In addition to offering useful implications for creating sentiment-aware investing strategies in Vietnam's dynamic and changing financial environment, this study adds to the expanding corpus of research on behavioral finance in emerging economies.

**TABLE OF CONTENTS**

[LIST OF TABLES 1](#_Toc201298223)

[LIST OF FIGURES 2](#_Toc201298224)

[LIST OF ABBREVIATIONS 4](#_Toc201298225)

[CHAPTER 1. INTRODUCTION 5](#_Toc201298226)

[1.1. Background 5](#_Toc201298227)

[1.2. Problem Statement 5](#_Toc201298228)

[1.3. Research Objective 7](#_Toc201298229)

[1.4. Scope and Limitations 7](#_Toc201298230)

[CHAPTER 2. LITERATURE REVIEW 9](#_Toc201298231)

[2.1. Introduction to the project 9](#_Toc201298232)

[2.2. Introduction to PhoBERT 10](#_Toc201298233)

[CHAPTER 3. METHODOLOGIES 12](#_Toc201298234)

[3.1. Data Collection 12](#_Toc201298235)

[3.1.1. Crawl News 12](#_Toc201298236)

[3.1.2. Crawl stock 15](#_Toc201298237)

[3.2. Data processing 16](#_Toc201298238)

[3.2.1. Training data 17](#_Toc201298239)

[3.2.2. Data testing 18](#_Toc201298240)

[3.3. Model 18](#_Toc201298241)

[3.3.1. Support Vector Machine 18](#_Toc201298242)

[3.3.2. Convolutional Neural Network 18](#_Toc201298243)

[3.3.3. Long Short-Term Memory 18](#_Toc201298244)

[3.3.4. PhoBERT 19](#_Toc201298245)

[3.3.5. Logistic Regression 19](#_Toc201298246)

[3.3.6. Deep Neural Network 19](#_Toc201298247)

[3.4. Metric evaluation 20](#_Toc201298248)

[3.4.1. Classification Problem 20](#_Toc201298249)

[3.4.2. Classification Metrics 21](#_Toc201298250)

[3.5. Power BI 22](#_Toc201298251)

[3.6. Streamlit 22](#_Toc201298252)

[CHAPTER 4. EXPERIMENT 24](#_Toc201298253)

[4.1. Dataset 24](#_Toc201298254)

[4.1.1. News.csv 24](#_Toc201298255)

[4.1.2. Full News.csv 26](#_Toc201298256)

[4.1.3. Stocks.csv 26](#_Toc201298257)

[4.2. Relationship designing 31](#_Toc201298258)

[4.3. Data visualization 33](#_Toc201298259)

[4.3.1. Structure of report in Power BI 33](#_Toc201298260)

[4.3.2. Functions of sections in a sheet 34](#_Toc201298261)

[4.4. Data processing 37](#_Toc201298262)

[4.4.1. Data Cleaning 37](#_Toc201298263)

[4.4.2. Text Preprocessing 37](#_Toc201298264)

[4.4.3. Sentiment Label Encoding 38](#_Toc201298265)

[4.4.4. Data Splitting 39](#_Toc201298266)

[4.4.5. Feature Extraction 40](#_Toc201298267)

[4.4.6. Text Representation with BERT 41](#_Toc201298268)

[4.4.7. Smote 41](#_Toc201298269)

[4.5. Model 42](#_Toc201298270)

[4.5.1. Support Vector Machine 42](#_Toc201298271)

[4.5.2. Convolutional Neural Network 43](#_Toc201298272)

[4.5.3. Long Short-Term Memory 43](#_Toc201298273)

[4.5.4. PhoBERT 44](#_Toc201298274)

[4.5.5. Logistic Regression 44](#_Toc201298275)

[4.5.6. Deep Neural Network 45](#_Toc201298276)

[4.6. Fill data 46](#_Toc201298277)

[CHAPTER 5. EXPERIMENTAL RESULTS 49](#_Toc201298278)

[5.1. Demo 49](#_Toc201298279)

[5.1.1. Home 49](#_Toc201298280)

[5.1.2. Dashboard 50](#_Toc201298281)

[5.2.2. Models Overview 72](#_Toc201298282)

[CHAPTER 6. CONCLUSION 75](#_Toc201298283)

[REFERENCES 77](#_Toc201298284)

# LIST OF TABLES

[Table 3.1: Classification Metrics in a Confusion Matrix Visualization 21](#_Toc201297749)

[Table 3.2: Classification Metrics: Definitions and Formulas 21](#_Toc201297750)

[Table 4.1: Attributes in file News 25](#_Toc201297751)

[Table 4.2: Attributes in Stocks.csv 27](#_Toc201297752)

[Table 4.3: Banks and financial institutions operating in Vietnam 28](#_Toc201297753)

[Table 4.4: Function of each component in the sheet 34](#_Toc201297754)

# LIST OF FIGURES

[Figure 3.1: Research Framework 12](#_Toc201297961)

[Figure 3.2: Overview of Data Collection Using Selenium 14](#_Toc201297962)

[Figure 3.3: Configuration of DNN 19](#_Toc201297963)

[Figure 4.1: 15 head rows samples of News.csv 24](#_Toc201297964)

[Figure 4.2: 15 head rows samples of Stocks.csv 26](#_Toc201297965)

[Figure 4.3: Data Schema for Data Visualization 32](#_Toc201297966)

[Figure 4.4: Data Cleaning 37](#_Toc201297967)

[Figure 4.5: Text Preprocessing 38](#_Toc201297968)

[Figure 4.6: Label Encoding 39](#_Toc201297969)

[Figure 4.7: Data Splitting 40](#_Toc201297970)

[Figure 4.8: Feature Extraction 40](#_Toc201297971)

[Figure 4.9: Tokenizer BERT 41](#_Toc201297972)

[Figure 4.10: Text-to-Vector conversion 41](#_Toc201297973)

[Figure 4.11: Saving label encoder 46](#_Toc201297974)

[Figure 4.12: The function for filling the data 47](#_Toc201297975)

[Figure 4.13: Filling data 48](#_Toc201297976)

[Figure 5.1: Sentiment insight demo 49](#_Toc201297977)

[Figure 5.2: Overview dashboard in streamlit 50](#_Toc201297978)

[Figure 5.3: Overview News 51](#_Toc201297979)

[Figure 5.4: Top 5 banks with the most news through all time 52](#_Toc201297980)

[Figure 5.5: Top 5 bank have the most positive/ negative/ neutral news through all time 53](#_Toc201297981)

[Figure 5.6: Top 5 bank have the most positive news through all time 55](#_Toc201297982)

[Figure 5.7: Top 5 bank have the most negative news through all time 56](#_Toc201297983)

[Figure 5.8: Top 5 bank have the most neutral news through all time 57](#_Toc201297984)

[Figure 5.9: Analysis News and Sentiment Frequency over time period 59](#_Toc201297985)

[Figure 5.10: Analysis Sentiment Distribution in News 60](#_Toc201297986)

[Figure 5.11: Annual News Volume Analysis 61](#_Toc201297987)

[Figure 5.12: Overview stocks 62](#_Toc201297988)

[Figure 5.13: Change price frequency by banks 62](#_Toc201297989)

[Figure 5.14: Quarterly Transaction Volume Analysis of Banks in 2024 64](#_Toc201297990)

[Figure 5.15: Analysis of Opening Prices Across Time 65](#_Toc201297991)

[Figure 5.16: Stock Price Summary Table 66](#_Toc201297992)

[Figure 5.17: Overview relationship between News & Stocks 68](#_Toc201297993)

[Figure 5.18: Correlation Analysis Between Transaction Volume and Sentiment Score through all time 69](#_Toc201297994)

[Figure 5.19: Daily Transaction Volume and News Frequency 70](#_Toc201297995)

[Figure 5.20: Daily Sentiment Distribution and Closing Price Analysis 71](#_Toc201297996)

[Figure 5.21: Model performance metric 73](#_Toc201297997)

[Figure 5.22: Get in touch with me of message 74](#_Toc201297998)

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| **SVM** | Support Vector Machine |
| **CNN** | Convolutional Neural Network |
| **DNN** | Deep Neuron Network |
| **BERT** | Birdirectional Encoder Representations from Transformer |
| **TF-IDF** | Term Frequency – Inverse Document Frequency |
| **NLP** | Natural Language Processing |
| **SMOTE** | Synthetic Minority Over-sampling Technique |
| **LSTM** | Long Short Term Memory |

# INTRODUCTION

## Background

Due in large part to economic reforms, greater foreign investment, and the quick development of financial technologies, Vietnam's financial markets have grown significantly in recent years. Specifically, the stock market has grown to be an important aspect of the nation's economy, drawing in both institutional and individual investors. Market sentiment has become one of the most important factors affecting stock price fluctuations and investor behavior in this dynamic climate.

Sentiment analysis, a branch of natural language processing (NLP), has become a popular method for analyzing qualitative information from social media, financial news, and publications in order to determine the mood of the market. News that is either positive or bad can have a big impact on investors' opinions, which can change market patterns and cause price volatility. In emerging markets like Vietnam, there is still a dearth of research on the relationship between public sentiment and stock performance, despite the fact that many studies in Western markets have demonstrated a high correlation. Vietnam's economy heavily relies on the banking industry, and market activity is often influenced by news about banks. Investors and governments can both benefit greatly from understanding how news sentiment affects bank stock performance. In order to evaluate the tone of financial news and investigate its relationship to stock price fluctuations in Vietnam's banking industry, this study employs sentiment analysis techniques. The study intends to facilitate the development of data-driven, sentiment-informed investment strategies that are suited to the Vietnamese financial environment by filling this vacuum in the literature.

## Problem Statement

In the financial markets, making quick and accurate decisions frequently depends on interpreting qualitative information, especially news and public opinion, in addition to numerical data. The behavioral and psychological aspects of market activity are frequently overlooked by classic stock price forecasting models, which have mostly depended on quantitative data like historical prices, trading volumes, and financial ratios.

Public mood and the media have a particularly strong influence on Vietnam's financial sector, which is still relatively new and expanding quickly. Price volatility rises as a result of investors' frequent reactions to news headlines and market rumors, particularly in delicate industries like banking. Nevertheless, there is still a dearth of systematic research that measures and examines this relationship in the context of the Vietnamese financial scene, despite the obvious influence of news emotion on investor behavior.

Furthermore, despite the fact that sentiment analysis employing Natural Language Processing (NLP) has demonstrated encouraging outcomes in developed markets, its use in Vietnam is still constrained by the country's linguistic complexity, data accessibility, and lack of domain-specific sentiment models in Vietnamese. As a result, there is a disconnect between sentiment-based forecasting's theoretical promise and its actual application in regional market settings.

By examining the degree to which sentiment gleaned from financial news affects stock price fluctuations, trading volume, and market behavior in Vietnam—with a particular emphasis on the banking sector—this study fills this knowledge vacuum. The objective is to determine whether sentiment can accurately forecast short-term stock performance and investigate the ways in which this data may be used to create better trading strategies.

This study intends to advance academic knowledge as well as the creation of sentiment-driven tools for analysts, investors, and financial institutions functioning in Vietnam's developing market environment by determining and measuring the relationship between news sentiment and stock market behavior.

## Research Objective

This study's main goal is to investigate and evaluate the relationship between sentiment in financial news and changes in stock prices in Vietnam's banking industry. In order to do this, the study concentrates on the following particular objectives:

**- Sentiment Extraction and Quantification**: Using Natural Language Processing (NLP) methods, financial news stories about Vietnamese banks are systematically categorized into three sentiment categories: neutral, negative, and positive. The goal of this classification is to convert qualitative news information into data that can be analyzed and organized.  
**- Evaluation of the Impact of Sentiment on Market Dynamics**: Examine the relationships between different sentiment patterns and important stock market metrics like price swings, trading volume, and overall market volatility. This goal aims to determine how news sentiment affects investor behavior and market movements in the near and perhaps long term.

**- Finding Predictive Patterns**: Look into the possibility that sentiment data from the past might be used to accurately predict how stocks will perform in the future. As a supplemental tool for stock market forecasting, the study attempts to ascertain the predictive validity of sentiment by comparing sentiment trends with subsequent stock price changes.

## Scope and Limitations

***Scope***

This study investigates the relationship between financial news sentiment and stock price movements, focusing on the banking sector in Vietnam. It analyzes Vietnamese-language news articles related to publicly listed banks on CafeF, using Natural Language Processing (NLP) to classify sentiment into positive, neutral, and negative categories. The study integrates this sentiment data with historical stock prices, trading volumes, and volatility to explore potential predictive patterns.

***Limitations***

The study has a number of shortcomings:  
- **Language barriers**: The intricacy of financial terminology may be beyond the scope of NLP techniques for Vietnamese.  
- **Data limitations**: Timestamps or other information may be missing from some news sources. **- Limited industry focus**: It's possible that the findings don't apply to industries other than banking. **- Scope in the short term**: The study focuses on immediate impacts rather than long-term patterns.  
- **Few origins of sentiment**: Social media and forums are not included in the analysis; only official news stories are.

# LITERATURE REVIEW

## Introduction to the project

Financial research has focused on the connection between stock market performance and sentiment surrounding financial news. The impact of market mood generated from financial news on stock price fluctuations and trading volumes has been the subject of numerous studies.

Using data from the Financial Times spanning six years, Alanyali et al. (2013) examined the relationship between stock market activity and financial news coverage. The study showed that stock trading volume tended to rise in parallel with the frequency of news mentions, suggesting a direct relationship between media exposure and investor activity.

Similarly, Alamsyah et al. (2019) explored the impact of headline news sentiment on stock returns in Indonesia. By applying sentiment analysis techniques such as Naïve Bayes and Support Vector Machine (SVM), they found a significant correlation between news sentiment and stock price fluctuations. Their findings emphasize the importance of sentiment analysis in predicting stock market behavior and enhancing investment decision-making​.

In the context of Vietnam, Tung et al. (2021) applied PhoBERT—a pre-trained Vietnamese language model—for the classification of stock-related news articles. Their findings showed that PhoBERT surpassed other models in performance, reaching an accuracy of up to 93% in sentiment classification. This underscores its effectiveness and potential for analyzing financial news in emerging markets such as Vietnam.

Further extending the literature, Fazlija and Harder (2022) applied BERT-based models to analyze financial news sentiment and predict the price direction of the S&P 500 index. Their research confirmed that sentiment extracted from news content significantly improves the accuracy of stock price direction prediction, especially when combined with historical market data​.

Collectively, these studies provide a strong foundation for understanding the dynamic interplay between market sentiment and stock performance. However, there is a noticeable gap in similar research within Vietnam's financial markets. This study aims to bridge that gap by applying sentiment analysis techniques to Vietnamese banking sector news to explore its influence on stock price movements.

## Introduction to PhoBERT

PhoBERT is a language model specially built for the Vietnamese language, based on the same powerful Transformer architecture as BERT (Bidirectional Encoder Representations from Transformers). It was developed by the VinAI Research team to tackle the unique challenges of processing Vietnamese text — such as complicated word segmentation, tone markers, and compound words — which are very different from languages like English.

Unlike English, where spaces denote word boundaries, Vietnamese often places spaces between syllables rather than complete words. This presents significant challenges for conventional tokenization techniques. To handle Vietnamese- specific language challenges, PhoBERT employs a custom word segmentation method and is trained on a vast corpus tailored for Vietnamese, making it more accurate than generic multilingual models, which includes data from sources such as Wikipedia, online news platforms, and social media. This large-scale training allows PhoBERT to capture both semantic and syntactic subtleties of Vietnamese text more effectively than generic multilingual BERT models.

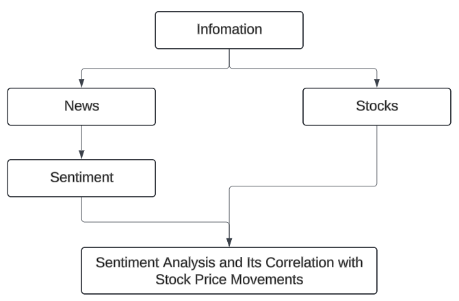
The model is released in two configurations—PhoBERT-base and PhoBERT-large—each offering different levels of model depth and computational capacity. PhoBERT has set new performance benchmarks for several Vietnamese NLP tasks, such as part-of-speech tagging, named entity recognition, and especially sentiment analysis.

In this study, PhoBERT is used to extract and analyze the emotional tone in Vietnamese financial news articles. Because it can understand the context of Vietnamese text so well, it provides more accurate results than older machine learning models or rule-based approaches. This helps us better understand how the mood of news coverage might influence stock price movements in Vietnam’s banking sector.

# METHODOLOGIES

The availability of timely and accurate information is crucial for investors and shareholders when making informed investment decisions. In Vietnam's dynamic financial markets, financial news plays a significant role in shaping market sentiment and influencing stock price movements. Positive news often drives stock prices upward, while negative news can trigger declines. Understanding this relationship is essential for predicting market trends and making strategic investment decisions.

This study focuses on analyzing the correlation between sentiment in financial news and stock price movements within Vietnam’s banking sector. By leveraging advanced natural language processing (NLP) techniques and financial data analysis, the study aims to uncover how market sentiment impacts stock performance. The research framework is designed as follows:



**Figure 3.1: Research Framework**

## Data Collection

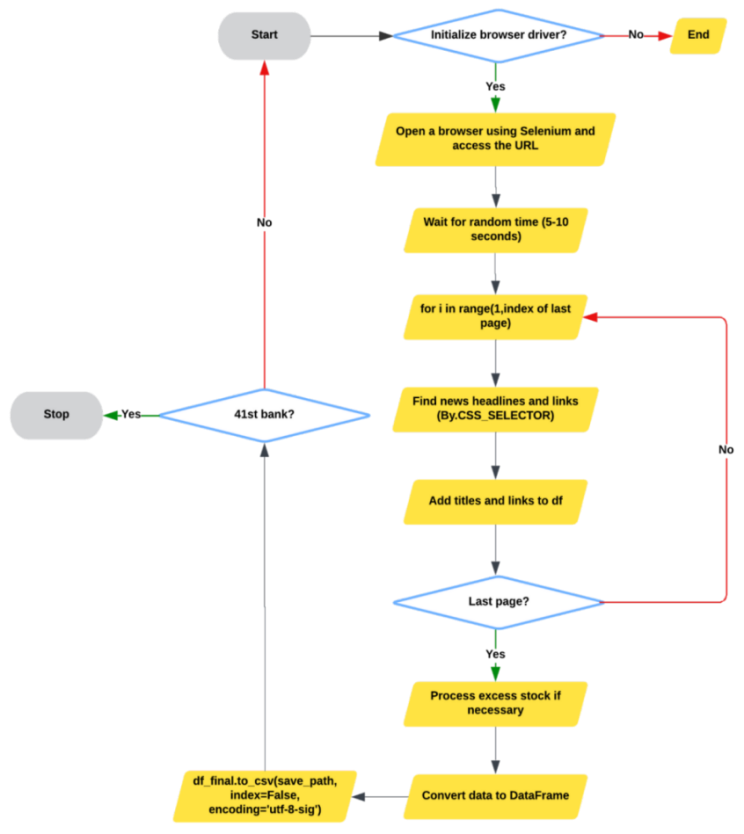
### Crawl News

At the beginning of my research on “sentiment analysis and how it relates to stock price movements in Vietnam’s financial markets”, I realized I needed two main types of data: news articles and stock market data. After exploring a few options, I decided to get everything from **Cafef.vn**. This site is popular in Vietnam for financial news and market updates, especially in the banking sector. It gave me access to both kinds of data I needed in one place, which made the whole process more convenient.

For the news part, we collected things like the article titles, short summaries, and the dates they were published—these are the kinds of details that help when analyzing sentiment. For stock data, I focused on the basics: opening and closing prices, price changes, and trading volumes.

To actually collect all this data, I used two main methods: **web scraping** and **APIs**. Using both helped me save time and effort. I chose **Python** for the job because it’s easy to use and has lots of helpful libraries. For scraping the news, I used **Selenium**, and for pulling stock data and timestamps from the pages, I used **BeautifulSoup**.

Web scraping turned out to be really helpful. It let me pull a lot of data directly from websites, especially when APIs weren’t available or didn’t give me enough info. Basically, it works by sending requests to a web page, reading its structure (HTML), and picking out the parts I want—then I saved everything into files like CSV or JSON. With Python and its tools like Selenium and BeautifulSoup, we were able to gather real-time data, track patterns, and build a clean dataset for the next steps of our project.



**Figure 3.2: Overview of Data Collection Using Selenium**

I used Python, Selenium, and other essential libraries like pandas, numpy, time, random, and csv to collect the data I needed for our project. I concentrated on using pertinent keywords for each business to look for news about banks and financial institutions that were listed on the Cafef.vn platform. Once the browser was launched and directed to the appropriate page, I allowed a short delay—usually between 5 to 10 seconds—to ensure the page had enough time to fully load, as the process was sensitive to internet speed and website responsiveness.

Despite minor variations in the content of each bank, Cafef's website maintained a uniform general structure, allowing me to use a single crawling strategy. Using CSS selectors, I collected the data after running a loop across every page that was accessible for every institution. Since there were frequently less items on the last pages than on the ones before, we included some additional logic to deal with certain situations and guard against data loss. Following collection, the data was arranged using pandas and stored in a CSV file with UTF-8-SIG encoding to ensure that Vietnamese characters were accurately maintained. To speed things up, the crawling was run in parallel for multiple banks and companies, which helped me save time without compromising coverage.

#### Besides news content, I also needed other information like the publication time. This was retrieved using API methods to complete the dataset. In total, I collected news data for 53 Vietnamese banks and financial firms, resulting in more than 20,000 records. The time span of the data varied by institution, but most ranged from early 2000 to the end of April 2025. But only possessing the unprocessed text was insufficient; sentiment analysis was required. In order to do this, I manually classified 8,000 news pieces into three groups: positive, neutral, and negative. I were unable to hand label the complete dataset due to time and resource limitations. In order to address this, I used the labeled data to train a machine learning model, which I then applied to the remaining records. By integrating automation and expert input, this method balanced scalability and quality, enabling us to easily produce a completely labeled dataset.

### Crawl stock

This section will explain how to obtain stock price information for Vietnamese banks from the CafeF website using Python and several tools, such as pandas, BeautifulSoup, and requests.

This data crawler's goal is to collect stock price information for multiple banks listed on CafeF, including trading date, opening price, closing price, highest price, lowest price, price change, and trading volume.

The Implementation Process:

- Create a List of Banks: First, we create a dictionary that links bank names to the appropriate stock symbols.

- Fetch JSON Data: We use the get\_json function to send HTTP requests to the CafeF API, which provides historical stock price data. The URL is created using the stock symbol and page index.

- Response processing: The response was successful. If successful, we will extract the stock price data from the JSON response.

- Data storage: Create a list of relevant data, convert to Pandas DataFrame for easy processing, and then export to CSV file.

- Data export: The final data is saved as CSV file. After the data collection process is complete, I collected from cafeF a comprehensive and well-structured dataset of stock prices of various banks over a certain period of time, including: trading date, starting price, closing price, highest price, lowest price in the transaction, price change, and trading volume.

Analysis will be much easier with data in this structured way. I can analyze various investment techniques, identify trends, and study market behavior. With this amount of insights, analysts and investors can better understand the market size, make the most optimal decisions and optimize their portfolios.

## 3.2. Data processing

In this part, data preprocessing was showed up financial news and stock data for sentiment analysis and predictive modeling. The key preprocessing steps consisted of:

Data Cleaning:

- Removed special characters, punctuation, and numbers.

- Converted all text to lowercase for consistency.

Text Processing:

- Applied tokenization to split text into individual words.

- Removed Vietnamese stop words to eliminate non-informative terms.

- Performed lemmatization to reduce words to their root forms.

Sentiment Labeling:

- Manually labeled the first 6,000 records as Positive, Neutral, or Negative.

- Encoded sentiment labels numerically for model training.

Data Splitting:

- Divided the manually labeled data into training (80%) and testing (20%) sets.

- Reserved the remaining 12,000 records for automatic sentiment prediction using the trained model.

Feature Extraction: Used TF-IDF Vectorization and advanced word embedding models (e.g., PhoBERT) to convert text into numerical features for machine learning.

This preprocessing ensured the data was clean, structured, and ready for accurate sentiment classification and stock price correlation analysis.

### Training data

The dataset is divided into two parts:

- Training set (80%): Used to train the sentiment classification model.

- Testing set (20%): Used to evaluate the performance of the model.

Initially, the first 8,000 records were manually labeled with either Positive, Neutral, or Negative sentiment. Different models were tested and PhoBERT was selected for its superior accuracy. This model was then used to automatically label the remaining 12,000 records.

After the labeling process was completed, the entire dataset of 20,000 records (a combination of manually labeled data and model-predicted data) was used to retrain the sentiment analysis model. This comprehensive dataset allowed the model to learn from a larger and more diverse dataset, improving its accuracy and robustness in sentiment classification.

### Data testing

The model was rigorously tested on the complete dataset to evaluate its performance. Key metrics such as accuracy, precision, recall, and F1-score were used to measure the model’s ability to correctly classify sentiment as Positive, Neutral, or Negative.

This final step that the model was well-calibrated and capable of accurately analyzing sentiment across the entire financial news dataset.

## 3.3. Model

### 3.3.1. Support Vector Machine

SVM is a classification algorithm that attemps to find the best boundary that separates different classes, particularly effective in text-based problems due to its ability to handle high- dimensional data. SVM works well when the data has clear boundaries between classes, making it suitable for handling complex sentiment classification tasks.

### 3.3.2. Convolutional Neural Network

CNN is a deep learning model traditionally used for image recognition but is also effective for text classification. CNN captures local patterns in data through convolutional filters, allowing it to detect key features in text. In sentiment analysis, CNN can recognize important phrases and word patterns that contribute to sentiment classification.

### 3.3.3. Long Short-Term Memory

Long Short-Term Memory networks (LSTMs) are a special kind of Recurrent Neural Network built to capture patterns and dependencies over extended sequences. They’re particularly powerful in natural language processing since they can retain contextual cues across long stretches of text. This ability to remember context makes LSTM architectures especially useful for sentiment analysis, where grasping the nuance of words and phrases is vital.

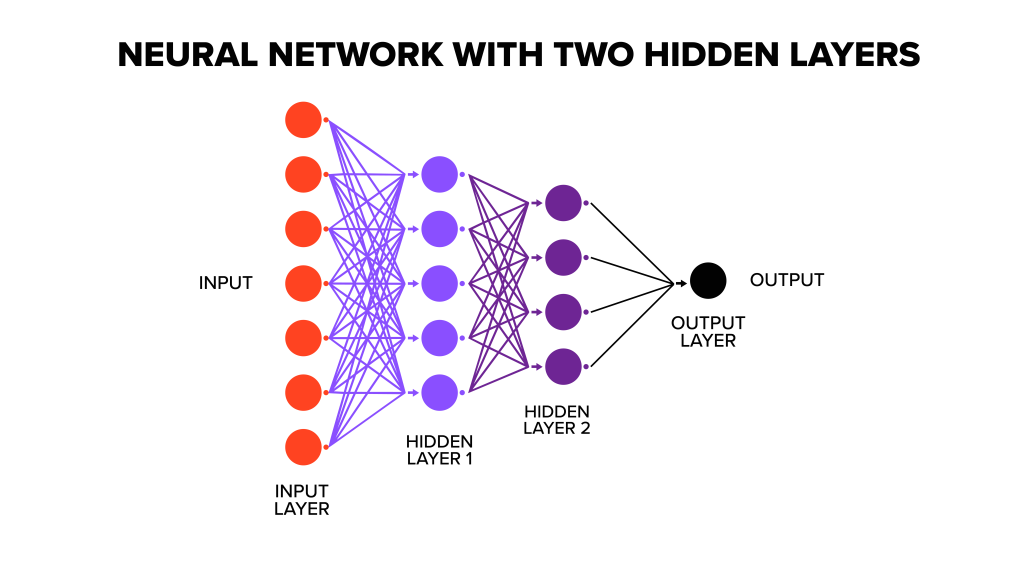
### 3.3.4. PhoBERT

PhoBERT is a pre-trained language model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically designed for the Vietnamese language. It captures deep contextual representations of words and phrases, making it highly effective for Vietnamese text classification tasks. PhoBERT was ultimately selected in this study for its superior performance in accurately classifying financial news sentiment.

### 3.3.5. Logistic Regression

Logistic Regression is a linear model used for both binary and multi-class classification problems. It predicts the probability of categorical outcomes using a logistic function. Due to its simplicity and interpretability, Logistic Regression is effective for classifying text data into sentiment categories.

### 3.3.6. Deep Neural Network

Deep Neural Network (DNN) is an artificial neural network with multiple hidden layers between the input and output layers. Each layer consists of interconnected neurons that transform input data through weighted connections and activation functions. 

**Figure 3.3: Configuration of DNN**

## 3.4. Metric evaluation

Evaluating the performance of sentiment classification models is crucial to understanding their accuracy and effectiveness. This study utilizes several standard classification metrics derived from the confusion matrix to measure how well the models classify financial news sentiment.

### 3.4.1. Classification Problem

The classification task involves predicting whether a financial news article conveys Positive, Neutral, or Negative sentiment. This process is evaluated using a confusion matrix that consists of four main components:

* True Positive (TP): Instances where the model correctly predicts a positive sentiment.
* True Negative (TN): Instances where the model correctly predicts a negative sentiment.
* False Positive (FP): Instances where the model incorrectly predicts a negative sentiment as positive.
* False Negative (FN): Instances where the model incorrectly predicts a positive sentiment as negative.

The confusion matrix visually categorizes predictions into four groups, providing insights into model performance:

The confusion matrix visually categorizes predictions into four groups, providing insights into model performance:

**Table 3.1: Classification Metrics in a Confusion Matrix Visualization**

|  |  |
| --- | --- |
| **Component** | **Description** |
| True Positive (TP) | Correctly identified positive cases. |
| True Negative (TN) | Correctly identified negative cases. |
| False Positive (FP) | Incorrectly identified negative cases as positive. |
| False Negative (FN) | Incorrectly identified positive cases as negative. |
| Classified as Positive | Total instances predicted as positive (TP + FP). |
| Positive/Negative | Reflects the true class distribution in the dataset. |

### 3.4.2. Classification Metrics

The following metrics are used to quantitatively evaluate model performance:

**Table 3.2: Classification Metrics: Definitions and Formulas**

|  |  |  |
| --- | --- | --- |
| **Metrics** | **Formula** | **Description** |
| Accuracy (A) | TP + TN / (TP + TN + FP + FN) | The ratio of correctly predicted sentiments (both positive and negative) over the total predictions. |
| Recall (R) | TP / (TP + FP) | The model's ability to correctly identify all actual positive sentiment cases (Sensitivity). |
| Precision (P) | TP / (TP + FN) | The ratio of correctly predicted positive sentiments over all instances predicted as positive. |
| F1-Score (F1) | (2 x P x R) / (P + R) | The harmonic mean of Precision and Recall, balancing both metrics. |

* Accuracy evaluates the overall correctness of the model but may not reflect performance well in imbalanced datasets.
* Recall emphasizes identifying all positive cases, reducing false negatives.
* Precision focuses on minimizing false positives, ensuring that predicted positive cases are truly positive.
* F1-Score balances precision and recall, making it a reliable metric when dealing with uneven class distributions.

These metrics provide a comprehensive assessment of how well the selected models classify sentiment in financial news and their impact on predicting stock price movements.

## 3.5. Power BI

Data visualization is the process of using charts, graphs, and images to represent data in an easily comprehensible manner. This process is crucial as it enables viewers to better interpret information through clear visuals, simplifies complex data, helps identify trends and patterns, and supports decision-making. For our research, data visualization is essential as it allows us to explore data more effectively. To create dynamic and efficient visualizations, we chose Power BI as the supporting software. Visualizing data with Power BI involves several steps to achieve the overall objective of our research: exploring data and understanding the relationship between sentiment analysis and Vietnam's financial market. First, we design a data model to establish connections between datasets, ensuring consistency and strong integration. Second, we visualize the data using effective and practical charts in Power BI. Finally, we analyze these visualizations to uncover insights and extract meaningful findings that contribute to our study.

## 3.6. Streamlit

To implement the sentiment classification model, we chose Streamlit, a prominent open-source framework that enables the rapid and flexible development of web applications. Utilizing Streamlit, we created an intuitive user interface that allows users to easily input data, such as titles and descriptions of financial articles.

The application we developed features key functions such as data entry and sentiment prediction based on a pre-trained model. When users provide the necessary information, the application quickly uses the model to classify the sentiment into categories: Positive, Neutral, or Negative. The predicted results are displayed instantly, offering users a clear and comprehensible view of the sentiments derived from financial articles.

The use of Streamlit enhances not only the efficiency of the analytical process but also improves user accessibility and interaction. This ultimately supports investors and analysts in making more informed decisions based on emotional insights from financial news and visual data analyses.

# EXPERIMENT

## 4.1. Dataset

The dataset utilized in this study was meticulously collected from CafeF.vn, one of the most reputable and widely used financial news platforms in Vietnam. Established as a leading source of financial information, CafeF.vn provides comprehensive coverage of economic news, stock market updates, banking sector developments, and corporate financial reports. The platform aggregates data from various trusted sources and offers real-time updates on market trends, making it a vital resource for investors, financial analysts, and policymakers.

CafeF.vn is renowned for its detailed reports on company earnings, market movements, and industry-specific developments, particularly in the banking sector, which is a critical pillar of Vietnam's growing economy. The data extracted from CafeF.vn in this study primarily focuses on news articles related to the operations, performance, and strategic decisions of various banks in Vietnam. This rich source of information serves as a solid foundation for conducting sentiment analysis and investigating the impact of news sentiment on stock price movements in the Vietnamese financial market.

### 4.1.1. News.csv



Figure 4.1: 15 head rows samples of News.csv

This dataset comprises comprehensive financial news related to various banks in Vietnam.

Data Collection Period:

* Start Date: January 1, 2000
* End Date: April 20, 2025

Dataset Scale:

* Total Records: 20,224 data entries.
* Number of Banks: 53 different banks.

Attributes in the Dataset:

**Table 4.1: Attributes in file News**

|  |  |
| --- | --- |
| Bank | The name of the bank mentioned in the news article |
| Link | The URL linking to the original article on CafeF.vn |
| Title | The headline of the news article, summarizing its main content |
| Description | A brief description providing more context about the article |
| Time-published | The publication date and time of the article (in day-month-year and hour-minute format) |
| Predict | Human labeled sentiment analysis(Positive, Negative, Neutral) |

With this dataset, we manually evaluated the sentiment (Positive, Negative, Neutral) of the first 8,000 records based on the Title and Description. We compared the sentiment classifications generated by ChatGPT with human annotations and observed that ChatGPT produced several inaccuracies. Therefore, we decided to manually label the sentiment for the first 8,000 records to ensure higher accuracy. The remaining 12,000 records will be automatically labeled using a trained machine learning model based on the manually evaluated data.

### 4.1.2. Full News.csv

After experimenting with various models on the first 8,000 manually labeled records, PhoBERT emerged as the most accurate and reliable model for sentiment classification. Recognizing its superior performance, we leveraged the PhoBERT model to efficiently and accurately annotate the sentiment for the remaining 12,000 records in the dataset. This approach ensures both consistency and precision in the overall sentiment analysis.

### 4.1.3. Stocks.csv

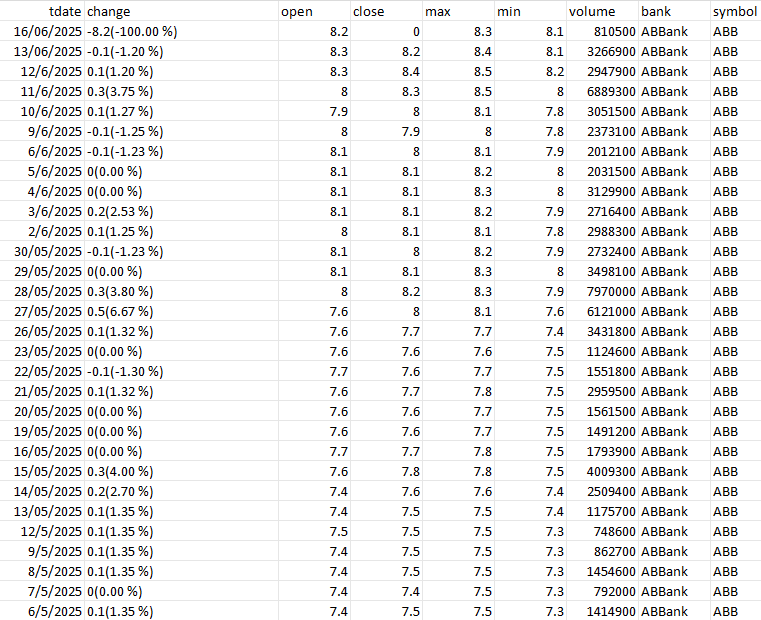


Figure 4.2: 15 head rows samples of Stocks.csv

This dataset contains detailed stock price information of various banks in Vietnam. It includes daily trading data such as stock price movements, trading volumes, and price fluctuations. The data is crucial for analyzing the impact of market sentiment on stock performance in the banking sector.

Data Collection Period: The dataset captures daily stock prices and trading data.

* Start Date: January 8, 2000
* End Date: June 16, 2025

Dataset Scale:

* Total Records: The dataset contains a total of 102,925 records.
* Number of Banks: 53 different banks.

Attributes in the Dataset:

**Table 4.2: Attributes in Stocks.csv**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Data Type** |
| **Bank** | The name of the bank whose stock data is recorded. | Categorical |
| **Tdate** | The trading date in the format (day/month/year). | Date |
| **Change** | The change in stock price along with the percentage change. | String |
| **Open** | The opening stock price on the trading day. | Float |
| **Close** | The closing stock price on the trading day (may have missing values). | Float |
| **Max** | The highest stock price during the trading day. | Float |
| **Min** | The lowest stock price during the trading day. | Float |
| **Volume** | The total number of shares traded on that day. | Integer |

This dataset allows for a comprehensive analysis of daily stock price movements in Vietnam’s banking sector.

By combining this stock data with sentiment analysis from financial news, researchers can investigate how market sentiment influences stock price fluctuations.

The dataset supports predictive modeling and strategic investment decisions based on both historical price trends and market sentiment.

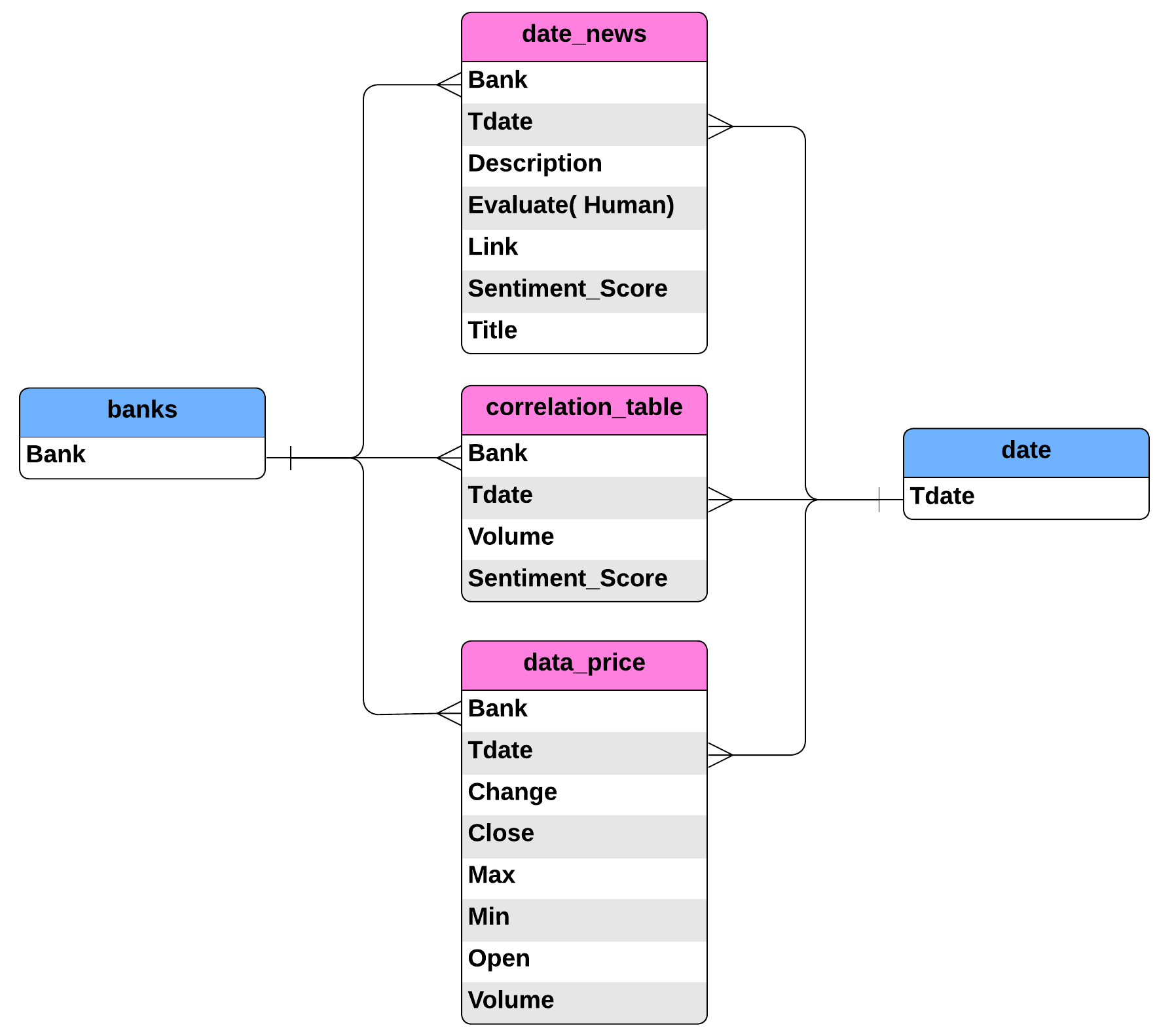
In this dataset, we systematically collected financial news and articles related to 41 banks and financial institutions operating in Vietnam. These entities play a significant role in the country's financial market, and their activities often have a direct impact on stock price movements and market sentiment. By gathering comprehensive news data from these banks and financial companies, the study aims to provide deeper insights into how financial information influences investor behavior and stock performance. This diverse and extensive dataset serves as a robust foundation for analyzing the correlation between market sentiment and stock price fluctuations in Vietnam's dynamic financial sector.

**Table 4.3: Banks and financial institutions operating in Vietnam**

|  |  |  |
| --- | --- | --- |
| **No.** | **Bank** | **Company Names** |
| 1 | ABBank | An Binh Commercial Joint Stock Bank |
| 2 | ACB | Asia Commercial Bank |
| 3 | Agribank | Vietnam Bank for Agriculture and Rural Development |
| 4 | BIDV | Bank for Investment and Development of Vietnam |
| 5 | Bac A Bank | North Asia Commercial Joint Stock Bank |
| 6 | CB | Construction Bank |
| 7 | DaiA Bank | Dai A Commercial Joint Stock Bank |
| 8 | DongA Bank | Dong A Commercial Joint Stock Bank |
| 9 | EVN Finance | EVN Finance Joint Stock Company |
| 10 | Eximbank | Vietnam Export Import Commercial Joint Stock Bank |
| 11 | FE CREDIT | FE Credit Financial Company |
| 12 | HDBank | Ho Chi Minh City Development Joint Stock Commercial Bank |
| 13 | HSBC | Hongkong and Shanghai Banking Corporation |
| 14 | Home Credit | Home Credit Vietnam Finance Company |
| 15 | Lienvietpostbank | Lien Viet Post Joint Stock Commercial Bank |
| 16 | MB | Military Commercial Joint Stock Bank |
| 17 | MSB | Maritime Bank |
| 18 | Mcredit | MB Shinsei Finance Limited Liability Company |
| 19 | Mirae Asset | Mirae Asset Finance Company |
| 20 | Nam A Bank | Nam A Commercial Joint Stock Bank |
| 21 | OCB | Orient Commercial Joint Stock Bank |
| 22 | OceanBank | Ocean Commercial Joint Stock Bank |
| 23 | PVcomBank | Vietnam Public Joint Stock Commercial Bank |
| 24 | SCB | Saigon Commercial Bank |
| 25 | SHB | Saigon-Hanoi Commercial Joint Stock Bank |
| 26 | Sacombank | Saigon Thuong Tin Commercial Joint Stock Bank |
| 27 | SeABank | Southeast Asia Commercial Joint Stock Bank |
| 28 | TPBank | Tien Phong Commercial Joint Stock Bank |
| 29 | Techcombank | Vietnam Technological and Commercial Joint Stock Bank |
| 30 | VBSP | Vietnam Bank for Social Policies |
| 31 | VDB | Vietnam Development Bank |
| 32 | VIB | Vietnam International Commercial Joint Stock Bank |
| 33 | VPBank | Vietnam Prosperity Joint Stock Commercial Bank |
| 34 | Viet Capital Bank | Viet Capital Commercial Joint Stock Bank |
| 35 | VietCredit | VietCredit Finance Company |
| 36 | Vietbank | Vietnam Thuong Tin Commercial Joint Stock Bank |
| 37 | Vietcombank | Joint Stock Commercial Bank for Foreign Trade of Vietnam |
| 38 | Vietinbank | Vietnam Joint Stock Commercial Bank for Industry and Trade |
| 39 | Woori | Woori Bank Vietnam |
| 40 | Hana Bank | Keb Hana Bank Vietnam |
| 41 | HSBC Vietnam | HSBC Bank (Vietnam) Ltd. |
| 42 | Indovina Bank | Indovina Bank Ltd (IVB) |
| 43 | PGBank | Prosperity And Growth Commercial Joint Stock Bank |
| 44 | HDSaiSon | HD SAISON Finance Co., Ltd |
| 45 | Prudential Finance | Prudential Vietnam Finance Company Ltd |
| 46 | SaigonBank | Saigon Bank for Industry and Trade |
| 47 | Standard Chartered | Standard Chartered Bank (Vietnam) Limited |
| 48 | Easy Credit | Easy Credit – Consumer Finance Division of EVN Finance |
| 49 | Mizuho Bank | Mizuho Bank Ltd. – Ho Chi Minh City Branch |
| 50 | KienlongBank | Kienlong Commercial Joint Stock Bank |
| 51 | Công Ty Tài chính TNHH MTV Ngân hàng TMCP Sài Gòn | Saigon Commercial Bank Finance One Member Limited Liability Company |
| 52 | EVNFinance | Vietnam Electricity Finance Joint Stock Company |
| 53 | Citibank | Citibank N.A. – Ho Chi Minh City Branch |

## 4.2. Relationship designing

The data model or data schema refers to the organization of data in the form of tables, where each table represents an entity that interacts and maintains relationships with other entities. Each entity contains attributes, which may share similarities or exhibit differences. A data schema includes these components along with primary keys, foreign keys, and relationships between the entities. Our dataset comprises five entities: banks (containing the names of banks and financial companies in Vietnam), date (including the publication time of news or timestamps associated with financial market stock information), data\_news (containing data related to news), data\_price (including data related to stocks), and the correlation table (containing data to explore the relationships between banks and their respective stocks). A detailed description of this dataset is presented in the Data Schema table below.

**

**Figure 4.3: Data Schema for Data Visualization**

The schema outlines the relationships between four primary tables: banks, date\_news, data\_price, and correlation\_table, which are interconnected through the primary keys Bank and Tdate. The banks table and the date table are connected to the remaining tables with a one to many relationship. The banks table serves as a reference entity, storing unique identifiers for individual banks. The date\_news table associates financial news with specific banks and dates, providing details such as a sentiment score, title, description, and external links for each news item. This table uses Bank and Tdate as foreign keys to link back to the banks and date tables, respectively. Similarly, the data\_price table tracks stock market metrics for each bank on specific dates, such as opening and closing prices, daily highs and lows, changes, and trade volumes. The primary keys Bank and Tdate in this table ensure a direct correlation to the banks and date tables. The correlation\_table consolidates data by combining the Bank and Tdate keys with metrics like trade volume and sentiment scores. It acts as a bridge between financial data from data\_price and the sentiment analysis derived from date\_news, allowing for comparative analysis across temporal and financial dimensions.

Thus, by building a data model as above, we have been able to build tight connections to help organize data effectively, providing good support for the data visualization process.

## 4.3. Data visualization

### 4.3.1. Structure of report in Power BI

Reports in Power BI are the data visualizations. Here we have three sheets including news data visualization, stock data visualization, and news-stock relationship data visualization. This gives us an overview of both news and stock data, and also a deeper look at the relationship between them. More specifically about each sheet, the structure of each sheet is quite similar, including the content of the title, basic information of the data, adjustments, and charts. The title helps the viewer have the most general view of the topic we are working on. The basic information presents an overview of the data such as the number of news, the number of banks, etc. In addition, there are adjustments for time, bank selection, emotion selection (Positive, Negative, Neutral), etc. The charts include Pie chart, Stacked bar chart, Stacked column chart, Treemap, etc. Each chart has different characteristics, giving us different purposes of use in an effective and meaningful way.

### 4.3.2. Functions of sections in a sheet

**Table 4.4: Function of each component in the sheet**

|  |  |  |
| --- | --- | --- |
| **Part** | **Component** | **Function** |
| Title and Basic Information | News sheet: title, team information, number of banks, number of news. | Provides an overview of news data, helping to understand the data clearly and simply. |
| Stock sheet: title, team information, highest close price, lowest close price, lowest volume, number of stock. | Provides an overview of stock data, helping to understand the types of prices in financial markets, the volume of transactions. |
| News & Stock sheet: title, number of stock, number of news | Provides an overview of stock data, and news data |
| Adjustments | News sheet: time from 1/1/2000 to 18/5/2025, bank choice, sentiment choice | * Helps analyze data over time * Helps analyze data by each bank or financial company * Helps analyze data by each sentiment choice |
| Stock sheet: time from 1/1/2000 to 16/6/2025, stock choice | * Helps analyze data over time * Helps analyze data by each stock |
| News & Stock sheet: time from 1/1/2000 to 18/11/2025, stock choice, sentiment choice | * Helps analyze data over time * Helps analyze data by each stock * Helps analyze data by each sentiment choice |
| Tables and charts | News sheet: Top 5 Banks have the Most News, Top 5 Banks have the Most Positive News, Analysis News and Sentiment Frequency Over Time, Analysis Sentiment Distribution in News, Annual News Volume Analysis | This research utilizes various charts to analyze news and sentiment trends in the banking sector. Key visualizations include the top five banks receiving the most news and positive coverage, sentiment distribution in news, and trends in news and sentiment frequency over time. Additionally, annual news volume analysis highlights long-term media attention patterns, providing a comprehensive understanding of public and media perceptions of banks. |
| Stock sheet: Analysis of Opening Prices Across Time, Change Price Frequency by Banks, data table, Change Price Frequency by Banks, Quarterly Transaction Volume Analysis of Banks | This study employs various visualizations to analyze banking data comprehensively. The "Analysis of Opening Prices Across Time" examines trends in stock opening prices, while the "Change Price Frequency by Banks" highlights price fluctuations across institutions. A data table complements these charts, offering detailed numerical insights. Additionally, the "Quarterly Transaction Volume Analysis of Banks" sheds light on transaction trends over time, providing a deeper understanding of market activities. |
| News & Stock sheet: Correlation Analysis Between Transaction Volume and Sentiment Score, Daily Transaction Volume and News Frequency, Daily Sentiment Distribution and Closing Price Analysis | This research utilizes charts to explore relationships between financial and sentiment data. The "Correlation Analysis Between Transaction Volume and Sentiment Score" examines how sentiment influences market activity. The "Daily Transaction Volume and News Frequency" chart highlights connections between news coverage and trading behavior. Lastly, the "Daily Sentiment Distribution and Closing Price Analysis" investigates the impact of sentiment on stock prices, providing insights into market dynamics. These visualizations offer a comprehensive view of the interplay between sentiment and financial performance. |

## 4.4. Data processing

### 4.4.1. Data Cleaning

* Removing special characters and numbers and drop missing values: special characters, punctuation marks, and numeric values were removed from the text using regular expressions to eliminate noise in the data.
* Lowercasing: All text in the Title and Description columns was converted to lowercase to ensure uniformity.

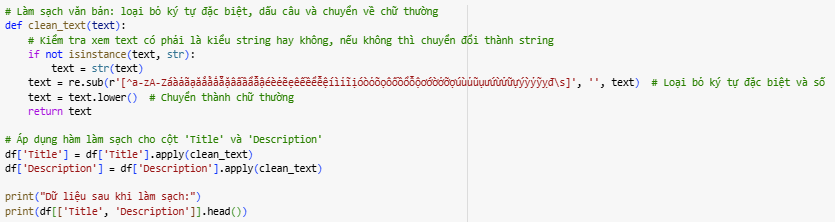
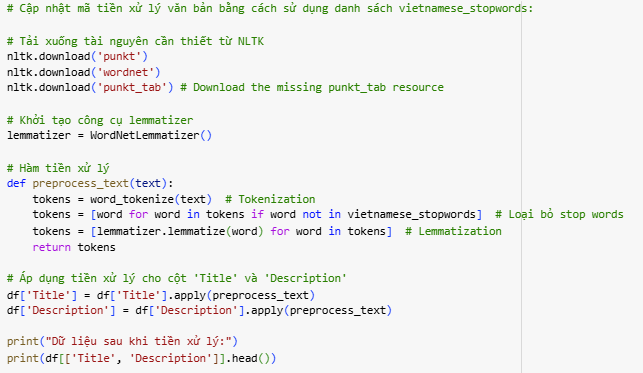


Figure 4.4: Data Cleaning

### 4.4.2. Text Preprocessing

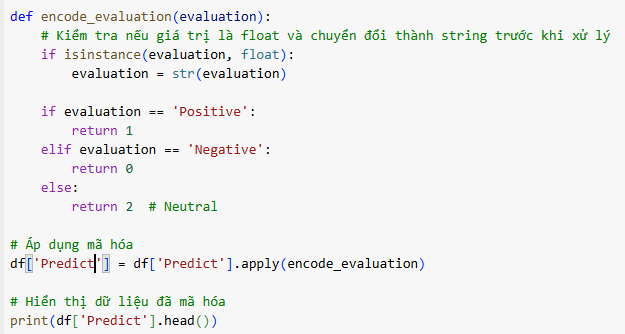
* Tokenization**:** Text was split into individual words (tokens) using the word\_tokenize function from the NLTK library.
* Stop words Removal**:** A Vietnamese stop words list was imported and applied to remove common, non-informative words (e.g., "và", "của") that do not contribute meaningfully to sentiment classification.
* Lemmatization: Words were reduced to their root forms using WordNetLemmatizer to standardize text and reduce variability in word forms.



**Figure 4.5: Text Preprocessing**

### 4.4.3. Sentiment Label Encoding

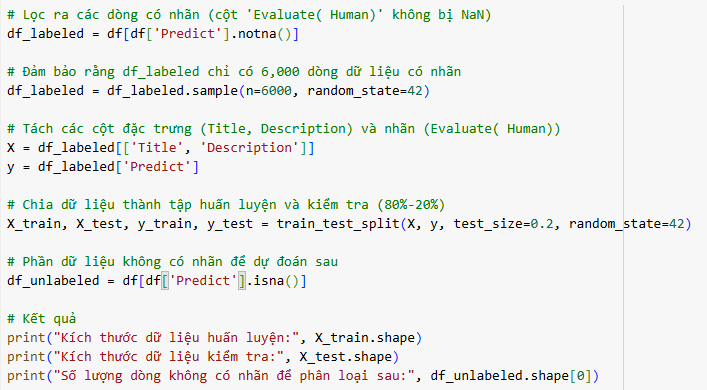
* Manual Labeling: The first 8,000 records were manually labeled as Positive, Neutral, or Negative.
* Label Encoding: Sentiment labels were numerically encoded for model training:
* Positive → 1
* Neutral → 2
* Negative → 0



**Figure 4.6: Label Encoding**

### 4.4.4. Data Splitting

* Training and Testing Sets: The manually labeled data was split into training (80%) and testing (20%) sets to evaluate model performance.
* Unlabeled Data for Prediction: The remaining 12,000 records were automatically labeled using the best-performing model (PhoBERT).



**Figure 4.7: Data Splitting**

### 4.4.5. Feature Extraction

* TF-IDF Vectorization: The Title and Description fields were transformed into numerical vectors using TF-IDF (Term Frequency-Inverse Document Frequency) to capture word importance.
* Word Embeddings: Advanced embedding models like Word2Vec, GloVe, and PhoBERT were employed to generate contextual word representations for more accurate sentiment analysis.

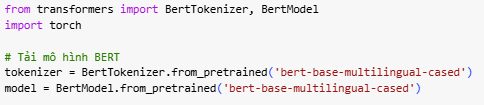


**Figure 4.8: Feature Extraction**

### 4.4.6. Text Representation with BERT

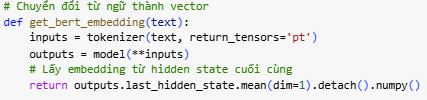
To capture deep contextual relationships in the Vietnamese text data, BERT was applied as a feature extraction technique.

* Tokenizer and Model Loading: The pre-trained bert-base-multilingual-cased model was used to process Vietnamese text.



**Figure 4.9: Tokenizer BERT**

* Text-to-Vector Conversion: The text was tokenized and passed through BERT to extract dense, meaningful vector representations.



**Figure 4.10: Text-to-Vector conversion**

This embedding captures semantic meaning, providing rich features for sentiment classification.

### 4.4.7. Smote

The dataset used for sentiment classification exhibited class imbalance, where certain sentiment classes (*Positive*, *Neutral*, *Negative*) were underrepresented compared to others. To solve this problem, SMOTE was applied to balance the dataset and improve model performance.

SMOTE is an over-sampling technique that generates synthetic data points for the minority class by interpolating between existing samples. This helps to balance the distribution of classes, which is critical for preventing machine learning models from being biased toward the majority class.

SMOTE was applied to the training set to balance the distribution of sentiment classes. Synthetic samples were generated for the minority classes to match the number of samples in the majority class.

Impact of SMOTE:

* Balanced Class Distribution: SMOTE ensured that all sentiment classes (Positive, Neutral, Negative) were equally represented in the training data, reducing model bias toward the majority class.
* Improved Model Performance: Balancing the dataset helped the classification models perform better, especially in correctly predicting minority sentiment classes, leading to improved accuracy, precision, recall, and F1-Score.
* Robust Training: The model trained on the balanced dataset became more robust and generalized better on unseen data.

## 4.5. Model

### 4.5.1. Support Vector Machine

Support Vector Machine(SVM) model was implemented with a linear kernel to efficiently classify news sentiment by finding the optimal hyperplane separating different sentiment classes.

* Preprocessing: Text data was vectorized using TF-IDF to transform words into numerical features.
* Training Configuration:
* The model was trained for 100 iterations with optimized hyperparameters.
* It was trained on the manually labeled dataset and tested on the 20% hold-out set.

### 4.5.2. Convolutional Neural Network

A Convolutional Neural Network (CNN) was implemented to capture local patterns and key features in the financial news text for sentiment classification.

* Model Architecture:
* Embedding Layer: Converts words into dense vector representations.
* Convolutional Layer (Conv1D): Detects important local text patterns.
* MaxPooling Layer: Reduces feature dimensions.
* Dense Layer: Outputs sentiment predictions (Positive, Neutral, or Negative).
* Training Configuration: The model was trained for 10 epochs with a batch size of 64.

### 4.5.3. Long Short-Term Memory

The LSTM model was implemented to capture long-term dependencies and contextual information in financial news text, which is crucial for sentiment analysis.

* Model Architecture:
* Embedding Layer: Converts text into vector representations.
* LSTM Layer (50 units): Captures sequential patterns and dependencies.
* Dropout Layer: Prevents overfitting by randomly deactivating neurons.
* Dense Layer: Outputs sentiment classification results.
* Training Configuration:
* The model was trained for 50 epochs with a batch size of 32.
* Used the Adam optimizer and categorical cross-entropy as the loss function.

### 4.5.4. PhoBERT

PhoBERT, a pre-trained language model specifically designed for Vietnamese text, was utilized for sentiment classification. This model leverages the BERT architecture, providing deep contextual understanding of Vietnamese language patterns.

* Preprocessing:
* Financial news articles were tokenized using the PhoBERT tokenizer.
* Contextual embeddings were generated from text for model input.
* Training Configuration:
* The model was fine-tuned on the manually labeled dataset of 8,000 records.
* PhoBERT was then used to automatically classify the remaining 12,000 records.

### 4.5.5. Logistic Regression

The Logistic Regression model was implemented as a baseline classifier for sentiment analysis due to its simplicity and effectiveness for binary and multi-class classification tasks. Essentially, it estimates the likelihood that a given input falls into each possible category.

* Preprocessing: TF-IDF Vectorization was applied to convert the cleaned text data into numerical vectors suitable for classification.
* Training Configuration:
* The model was trained on the manually labeled dataset with an 80-20 split for training and testing.
* The model used the 'liblinear' solver, which is suitable for small datasets and multi-class classification.

### 4.5.6. Deep Neural Network

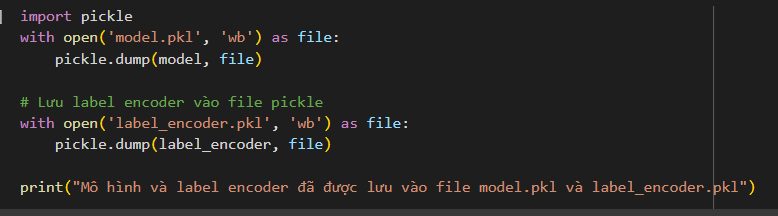
The Deep Neural Network (DNN) model was implemented to classify the sentiment of financial news articles. DNNs are highly effective for text classification tasks because it can pick up intricate patterns and build layered representations from the data.

* Preprocessing:
* The Title and Description fields were cleaned and combined into a single text column.
* Text data was vectorized using TF-IDF Vectorization with a maximum of 1,000 features.
* To address class imbalance, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied.
* Model Architecture:
* Input Layer: Fully connected layer with 128 neurons and ReLU activation.
* Dropout Layer: Applied with a rate of 0.5 to prevent overfitting.
* Hidden Layer: Fully connected layer with 64 neurons and ReLU activation.
* Dropout Layer: Another dropout layer with a rate of 0.5.
* Output Layer: Fully connected layer with a Softmax activation to classify news into sentiment categories.
* Training Configuration:
* The model was compiled using the Adam optimizer and Sparse Categorical Cross Entropy as the loss function.
* Early Stopping was implemented to halt training when the validation accuracy stopped improving.
* ReduceLROnPlateau was used to reduce the learning rate when the validation loss plateaued.
* The model was trained for 50 epochs with a batch size of 32 and a 20% validation split.

## 4.6. Fill data

First, I chose the PhoBERT model to handle missing data in my dataset. After selecting the model, I saved it to a pickle file for easy access and future use. This step ensures that I can load the model efficiently without needing to retrain it every time.

To fill in the null values, I implemented a process where I preprocess the input data. This involves combining the title and content of each entry into a single string, which is then tokenized using the PhoBERT tokenizer. The input is padded and truncated to fit within a maximum length of 512 tokens, making it compatible with the model's requirements.

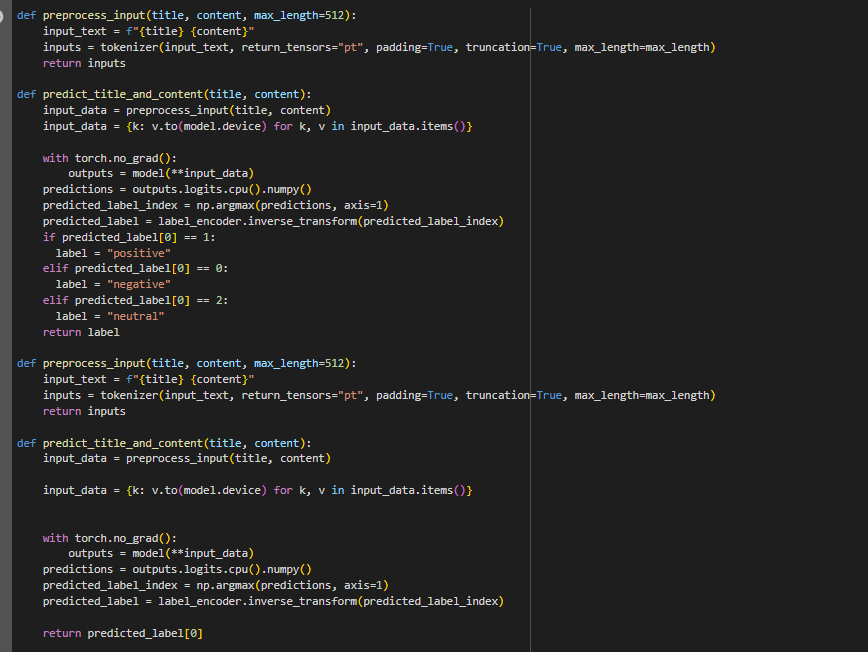


**Figure 4.11: Saving label encoder**

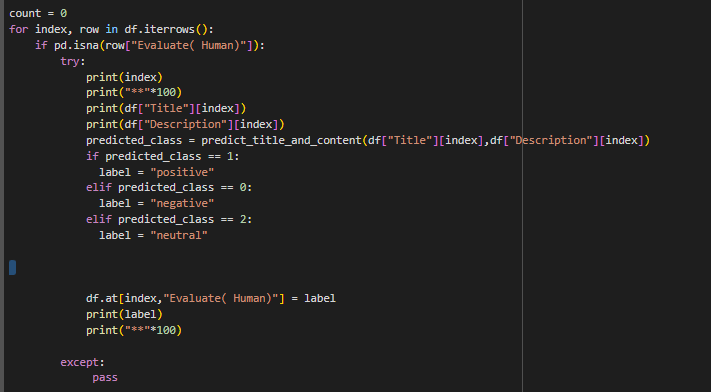
Next, I created a function called predict\_title\_and\_content. This function takes the title and content as inputs, preprocesses them, and then feeds them into the PhoBERT model to obtain predictions. The model outputs logits, which I convert into probabilities. The predicted class is determined by finding the index of the maximum probability, which is then mapped back to human-readable labels using a label encoder.

I then iterated through the DataFrame containing my dataset, checking for any null values in the "Predict" column. For each entry with missing data, I used the predict\_title\_and\_content function to classify the sentiment as positive, negative, or neutral. The results were saved back into the DataFrame, effectively filling in the missing values

Iterate through the DataFrame, predict the sentiment for rows with null evaluations, and update the DataFrame accordingly



**Figure 4.12: The function for filling the data**



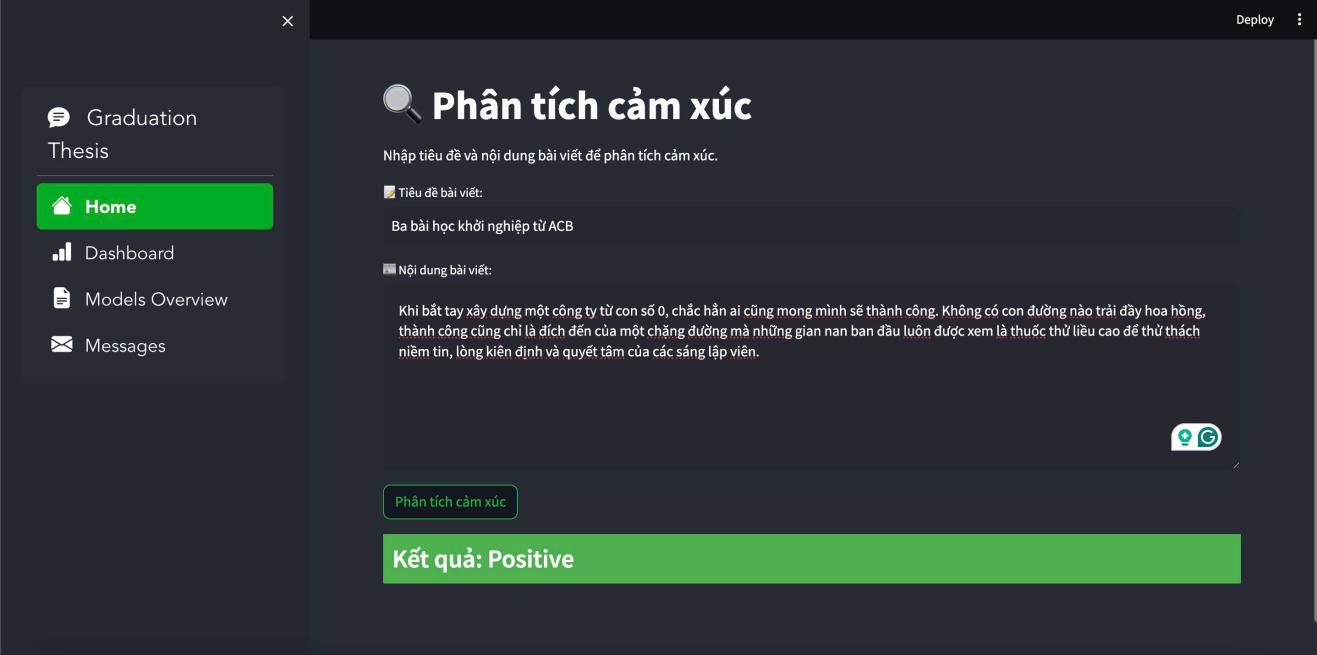
**Figure 4.13: Filling data**

# EXPERIMENTAL RESULTS

## 5.1. Demo

### 5.1.1. Home

The Home section of the "Sentiment Insight" application serves as the user-friendly interface where users can input news headlines and content for sentiment analysis. This section features a clear title that indicates its purpose, along with a brief introduction explaining the application's goal of providing insights into emotional sentiment from textual data. Users are presented with text input fields for entering the news title and a larger area for the detailed content. After submitting the information, users can click the "Predict" button, which triggers the sentiment analysis process. The application utilizes a pre-trained PhoBERT model for this task, ensuring accurate sentiment predictions tailored to Vietnamese text. The model, along with the label encoder and vectorizer, has been securely stored in files using Python's pickle library. Specifically, the model is loaded from model.pkl, the label encoder from label\_encoder.pkl, and the vectorizer from vectorizer.pkl. Once the prediction is made, the results are displayed in a dedicated area, giving users immediate feedback on the sentiment classification whether it is positive, negative, or neutral.



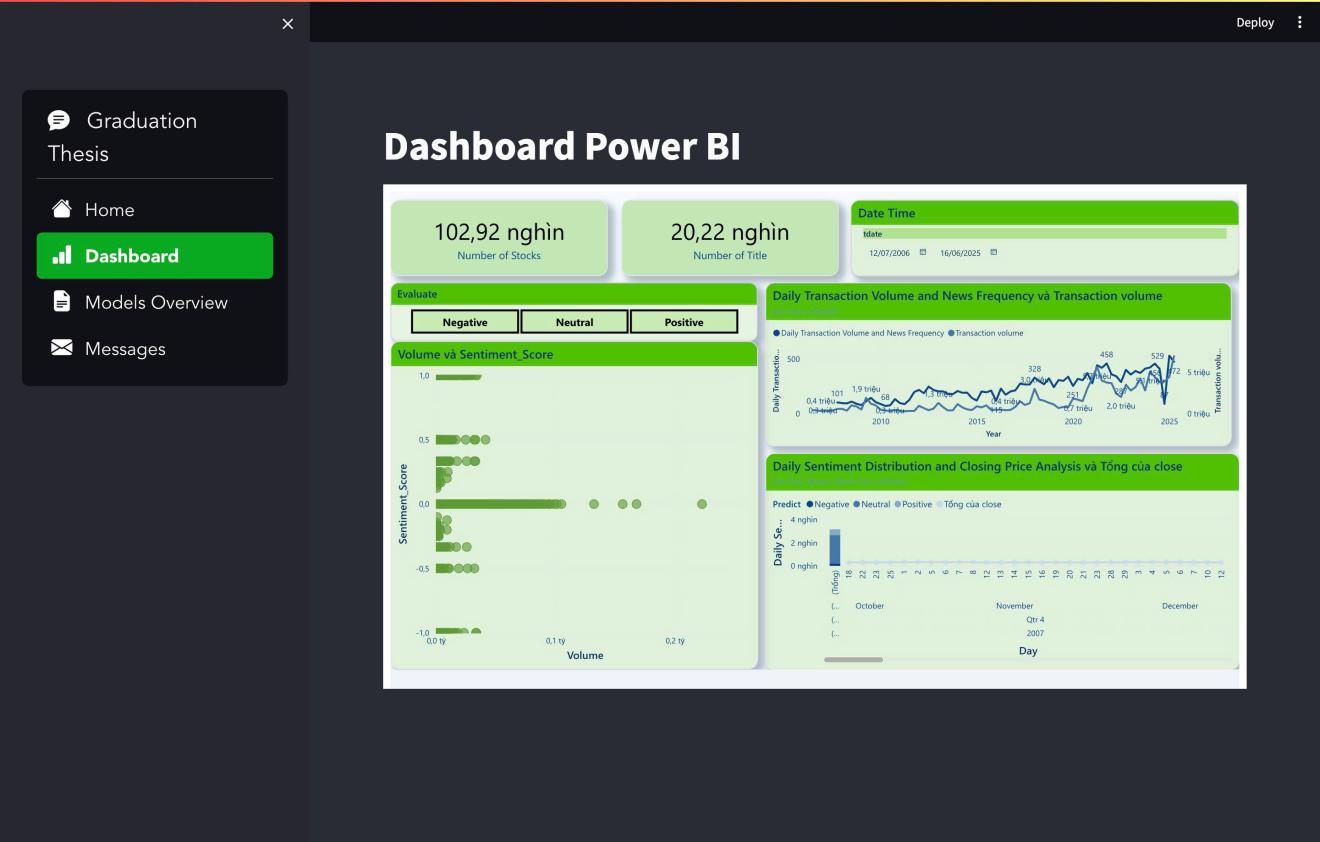
**Figure 5.1: Sentiment insight demo**

The prediction results will be displayed immediately on the interface, making it easy for users to recognize the emotional sentiment of the news they have entered. The Home section not only facilitates data entry but also provides an overview of how the system operates, from model storage to analysis and result display.

### 5.1.2. Dashboard

The Dashboard of the "Sentiment Insight" application offers a comprehensive and in-depth analysis of investor sentiment and its correlation with stock price movements in Vietnam's financial markets. This interactive dashboard is built using Power BI and is seamlessly embedded into our web application through an iframe, providing a smooth experience for users.

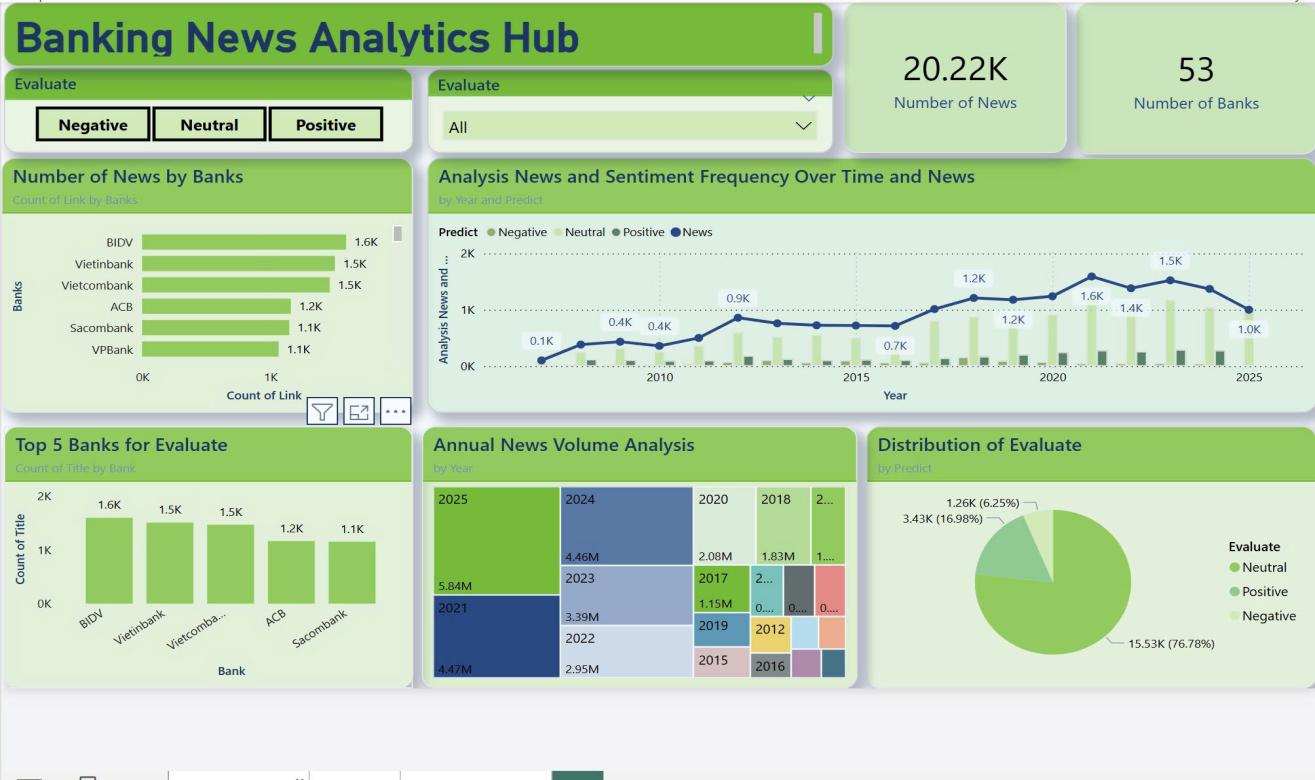
Within the dashboard, users will find various visualizations, including bar charts and line graphs, displaying key metrics such as sentiment scores, market trends, and stock price fluctuations over time. Users can easily interact with these visual elements to gain insights into how investor sentiment impacts the dynamics of the financial market.



**Figure 5.2: Overview dashboard in streamlit**

We have prioritized a user-friendly and intuitive design, enabling both analysts and general users to access and utilize sentiment analysis information effectively. Notably, the integration of the dashboard into the web application not only enhances its functionality but also creates a powerful platform for data exploration and informed investment decisions.

#### **5.1.2.1. News**



**Figure 5.3: Overview News**

Our news data from 53 Vietnamese banks and financial companies, including about 20.224 thousand news items.



**Figure 5.4: Top 5 banks with the most news through all time**

**Purpose:** This chart helps identify banks that appear most frequently in the media, reflecting public and market interest in each bank.

**Insights**

- BIDV, Vietinbank, and Vietcombank are state-owned banks, playing a crucial role in Vietnam’s financial system. As a result, they are frequently mentioned in the media regarding macroeconomic policies, major financial activities, or directives from the State Bank of Vietnam.

- Sacombank and ACB are private banks with large-scale operations, especially in retail banking, credit, and investment. Their presence in the media may stem from expansion strategies, financial events, or stock market movements.

- BIDV has the highest number of news articles (1,6k), indicating its key role in major financial projects, credit policies, and economic support programs.

- Vietinbank and Vietcombank also have a high volume of news coverage, emphasizing their significant roles in the banking sector and their frequent engagement in activities that attract public attention.

- Sacombank and ACB are the two most-mentioned private banks, likely due to their dynamic business activities, expansion strategies, or major internal structural changes.

- VPBank also has 1,1k bank, which is the most lowest news in top 5 banks with the highest news through all the time

**Strategic**

- For state-owned banks: It is crucial to ensure that media coverage accurately reflects their strategies and policies, particularly in relation to government decisions.

- For private banks: A high volume of news can enhance brand visibility, but it also requires careful media management, especially concerning financial matters, stock performance, or internal conflicts.

- Extensive media coverage presents an opportunity to strengthen brand awareness but, if not managed properly, can lead to a public relations crisis, impacting customer and investor trust.



**Figure 5.5: Top 5 bank have the most positive/ negative/ neutral news through all time**

**Purpose:**

This chart serves to analyze the media coverage of banks, categorizing them into positive, negative, and neutral news. It provides valuable insights into how different banks are perceived by the public and the media, helping to assess:

- Brand & Communication Strategy: By identifying which banks receive the most positive coverage, we can evaluate which institutions are effectively building their image and managing their media presence.

- Trust & Reputation: A high volume of positive news can indicate that a bank is viewed as more reliable by both customers and investors, suggesting strong trust and reputation in the market.

- Market Sentiment Impact: Positive media coverage can have a significant effect on investment decisions, capital mobilization, and customer loyalty, thus directly influencing a bank's financial performance and growth.

- Reputational Risks: Banks with a high volume of negative coverage are at risk of public distrust, reduced customer engagement, and increased regulatory scrutiny. This can damage their image and harm their long-term growth prospects.

- Assessing State-owned vs. Private Banks: The chart also highlights the differences in how state-owned and private banks manage their public image, especially in terms of crisis response and media perception. This comparison can provide valuable insights into the unique challenges faced by each banking model.



**Figure 5.6: Top 5 bank have the most positive news through all time**

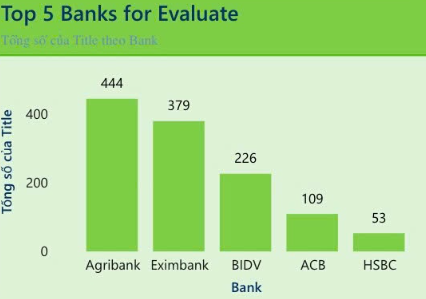
**Insights**

BIDV (625 news) are the two state-owned banks with the highest number of positive news articles. Also, private banks such as ACB(801 News), HDBank (520 news), HSBC (444 news), and Eximbank (292 news) also have high levels of positive news, indicating strong growth and effective communication strategies. So:

- State-owned banks (BIDV) have the most positive news → This may reflect customer confidence in government-backed banks and their sustainable development strategies.

- Private banks are also performing well → ACB, HDBank, HSBC, and Eximbank have made breakthroughs in products, services, and media efforts, attracting strong market attention.

- Differences between state-owned and private banks → State-owned banks are often perceived as stable and trustworthy, whereas private banks gain positive media coverage through innovation, technology, and customer service.



**Figure 5.7: Top 5 bank have the most negative news through all time**

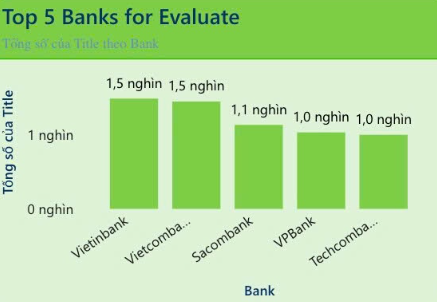
**Insights**

Agribank (444 news) have the highest negative news volume and BIDV(226 News) have the most average negative news volume, and also two of banks I mentioned about is state-owned banks. This suggests that even government-backed banks are not immune to public criticism and operational risks. ACB, Eximbank and HSB**C**, three;’;; private/commercial banks, also appear in the top 5, indicating challenges in governance, operational issues, or public perception problems. So:

**- State-owned banks face more scrutiny:** Due to their size, role in the economy, and government ties, they attract more media attention, making negative news more impactful.

**- Negative news volume does not necessarily mean poor performance:** Some of these banks (e.g., BIDV) also rank high in positive news, suggesting they are actively engaged in the market but face criticism alongside success.

**- Reputation risk is a major issue for all banks:** Both state-owned and private banks must proactively manage their public image to mitigate negative sentiment.



**Figure 5.8: Top 5 bank have the most neutral news through all time**

**Insights:**

VietinBank and Vietcombank are leading with the highest number of neutral articles (1.5 thousand), followed by VPBank, Sacombank and Techcombank. The neutral tone in coverage may indicate that Vietinbank and Vietcombank( state-owned banks) have always received considerable media attention and

**Strategies:**

**Banks with the Most Positive News**

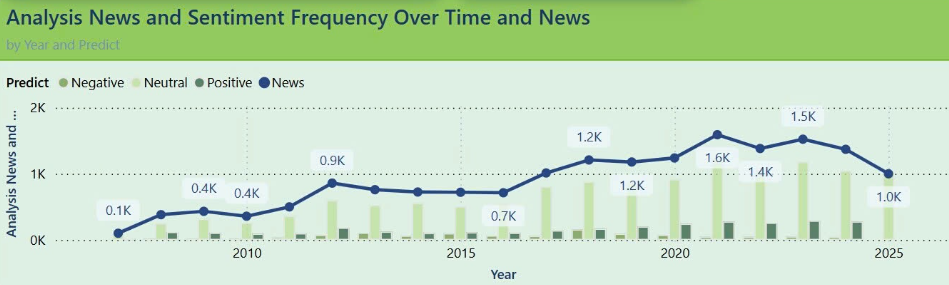
* Maintain and Optimize Positive Media Strategies: Banks with the most positive media coverage, especially state-owned banks like BIDV, can leverage media attention to continue building a stable, transparent, and trustworthy image. They should focus on developing attractive financial products while maximizing the impact of positive media coverage to strengthen brand reputation.
* Private banks, such as ACB, Eximbank, HSBC and HDBank should intensify PR campaigns focusing on innovation, technology, and customer experience to maintain media attention, while asserting their ability to innovate and be creative in the market.
* Leverage Positive News for Marketing & Investment Attraction: Banks can use positive media coverage as part of their marketing campaigns to attract customers and investors. Additionally, they should build strong relationships with the press to ensure that shared information is transparent and accurate, maintaining a solid brand image.

**Banks with the Most Negative News**

* Manage Media Risks & Control Negative News: While having a high amount of positive news is beneficial, banks must also focus on crisis communication management to prevent negative news from harming customer trust. They should guide media narratives towards beneficial topics such as growth, innovation, and corporate social responsibility (CSR).
* Strengthen Crisis Communication & Media Management: State-owned banks like Agribank and BIDV need to focus on transparent communication, addressing public concerns quickly, and ensuring regulatory compliance to maintain trust. Private banks such as Eximbank, HSBC and ACB need to enhance public relations efforts and improve customer service to reduce negative media attention.
* Investigate Root Causes of Negative News: Conduct sentiment analysis to understand common themes in negative news (e.g., poor customer service, financial instability, or regulatory issues). Banks should address internal weaknesses and create proactive solutions rather than reactive damage control.

**Banks with the Most Neutral News**

* Balance Media Coverage with Positive News: Banks receiving a high volume of neutral news, like Vietinbank and Vietcombank, need to find ways to convert neutral coverage into positive news. This can be achieved through stronger media campaigns that highlight achievements, innovation, and community impact.
* Develop Strategic Partnerships with Media Outlets: These banks should build strong relationships with media outlets to ensure fair and accurate reporting, which will help maintain a positive brand image.
* Strengthen Risk Management & Internal Governance: Improving risk management and internal governance is crucial for banks with a high volume of neutral news. This will help prevent any issues from escalating into a major PR crisis.



**Figure 5.9: Analysis News and Sentiment Frequency over time period**

**Purpose**

The goal of this chart is to provide a clear view of the bank's media sentiment over time, illustrating the balance of positive, negative, and neutral news. By analyzing this, we can:

- Monitor Public Perception: Understanding the ratio of positive and negative news helps evaluate how the market views the bank.

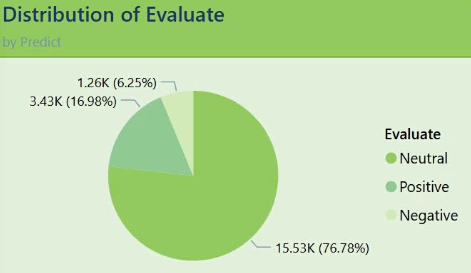
- Assess Risks and Opportunities: Identifying periods with high negative news volume enables the bank to take proactive steps to manage its image.

**Insight**

- From 2008 to 2021, the frequency of the topic/banking in the press increased 10 times.

- News is increasingly multidimensional: The period 2020–2021 recorded many positive and negative news side by side, reflecting the financial-banking environment with both opportunities and challenges.

- Stabilization after the peak: From 2022 to the forecast 2025, the amount of news tends to decrease slightly and the sentiment distribution returns to a more balanced state.



**Figure 5.10: Analysis Sentiment Distribution in News**

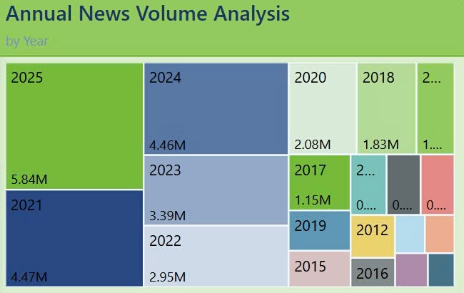
**Purpose**

This chart shows the general trend in the work being done about banks, including positive, negative and neutral news. The goal is to help banks identify the image they can use in the utility media and from there build an appropriate communication strategy.

**Insight**

The chart shows that positive news overwhelmingly dominates, accounting for 16.98%. This could be the result of communication strategies used by banks to manage public opinion and attract customers. Banks often release positive information to build a strong brand image and foster trust with the public.

However, banks with a higher percentage of negative news should actively work on improving their situation, as negative coverage can impact their reputation and brand image.



**Figure 5.11: Annual News Volume Analysis**

**Purpose**

This chart analyzes the annual news volume related to banks, helping assess the trend of information coverage over the years. The purpose of the chart is to help banks understand the changes in media coverage over time, thereby optimizing communication strategies and tracking the industry's development.

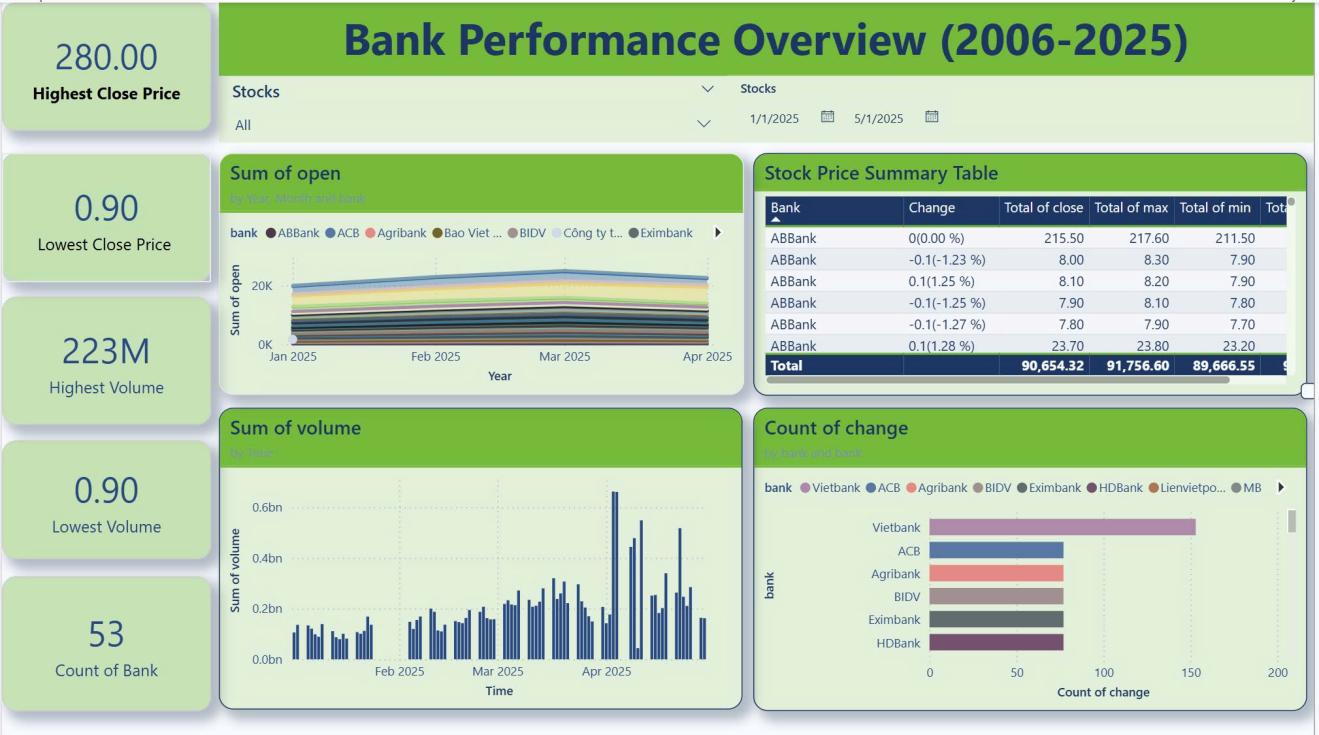
**Insight**

- The 2012-2016 accounted for only a few hundred thousand articles, each under 0.3 million

- From 2017 (1.15 M) to 2020 (2.08 M), news volume roughly doubled, signaling growing media attention.

- The years 2022 (2.95 M) and 2023 (3.39 M) show steady, year‑on‑year increases (~15–20%), indicating sustained interest rather than a one‑off spike. The trend of increasing news over time can be explained by the rapid growth of digital media, more detailed financial reports, and growing interest in issues within the banking sector.

#### **4.1.2.2. Bank Performance Overview**

**

**Figure 5.12: Overview stocks**

**General**



**Figure 5.13: Change price frequency by banks**

**Purpose**

- Identify banks with the highest price change frequency highlighting key players that experienced significant stock price fluctuations.

- Evaluate whether these price changes were driven by positive factors or negative factors

- Provide insights into the overall stability and sentiment of the banking sector by analyzing price volatility trends.

**Insight**

The banks with the most significant price changes in 2025 include Sacombank, ACB, Standard Chartered, followed by Công ty Tài Chính TNHH MTC Sài Gòn, VietCapital Bank, VietCredit. These frequent price changes indicate high trading activity but also raise concerns about market volatility for these banks.

Potential Causes

* Positive Impact: High price changes might reflect market expectations for future growth, strategic expansions, or robust financial performance.
* Negative Impact: Conversely, they could signal market pessimism due to underperformance, declining investor trust, or external factors like regulatory challenges and economic uncertainties.

Recommendations:

* These banks should analyze the root causes of price fluctuations and focus on addressing negative investor sentiment through better communication, transparency, and strategic planning.
* Stabilizing stock performance is essential, and banks should aim to rebuild trust through strong operational results, improved governance, and consistent financial reporting.



**Figure 5.14: Quarterly Transaction Volume Analysis of Banks in 2024**

**Purpose**

- Analyze the quarterly transaction volume trend of banks in 2024 to understand investor behavior over time.

- Identify periods of high or low trading activity, providing insights into market sentiment and potential factors influencing these trends.

- Assess the relationship between trading volumes and macroeconomic conditions, such as interest rates, inflation, or regulatory policies.

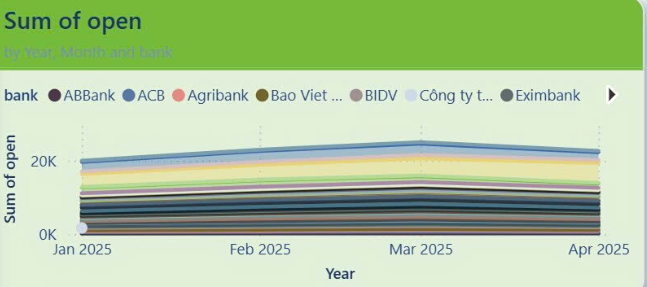
**Insight**

**Overall Upward Trend:**

As we seen in the chart, the trend analysis has the way of overall upward trend here. According to this chart, we have stable start in January–February (around **0.10–0.15 billion** per day), volume steadily climbed through March, peaking regularly above **0.25 billion**. This suggests growing market interest or heightened news flow as we moved into Q2.

Also in April, several days topped **0.50 billion**, with the single largest spike close to **0.65 billion**. From the chart, we conclude that volatility increasing volume in 2025. The variability of daily volume widened in mid‑March onward: low days around **0.15 billion** contrast sharply with high days above **0.40 billion**.

This higher volatility again points to an environment of more frequent catalyst‑driven activity.



**Figure 5.15: Analysis of Opening Prices Across Time**

**Purpose**

- To analyze the trend of opening prices of Sacombank's stock over the period from 1/2025 to 4/2025.

- To identify fluctuations in stock prices during historical periods, providing an overview of Sacombank's performance in the financial market.

- To support the assessment of the impact of macroeconomic factors, internal banking situations, or market sentiment on stock prices.

**Insight**

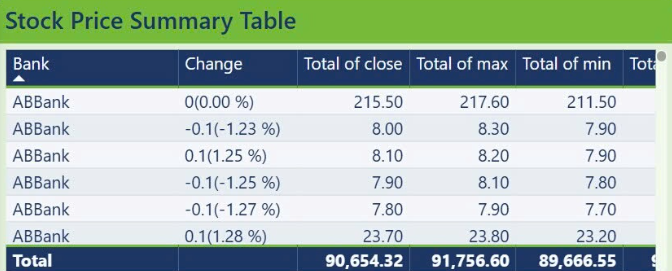
When I look at how much “opening volume” each bank brings in from January through April 2025, a few things jump out:

Think of total opening volume as a balloon that gently inflates from about 20K in January to roughly 21.5 K by March—it doesn’t suddenly pop or deflate. By April, it’s held there, hovering around the same level.

**Old Friends Still in Charge:** If my banks were a classroom, the big players—say BIDV, Vietcombank, VietinBank, Techcombank—are the “popular kids” who always sit up front. Their slice of the pie hardly ever changes month to month. No newcomers have jumped ahead, and nobody’s fallen way behind.

The smaller banks (ABBank, ACB, Agribank, Eximbank, etc.) are like the rest of the class—each one shows up reliably, takes its seat, but doesn’t suddenly stand on the desk or disappear.

In this chart, I will use ABBank as a example so we can understand more about the chart.



**Figure 5.16: Stock Price Summary Table**

**Purpose**

I wanted to peek under the hood of ABBank’s trading lately—how often its share price moves each day, how big those swings are, and what that tells us about investor mood.

**Insight**

ABBank’s share price barely budges more than about ±1.25% on any given day. Over the whole period, the total of all closing prices (90,654) actually sits closer to the bottom‑of‑day lows (89,667) than the highs (91,757). In plain English, folks aren’t making wild bets—and there might be a little more selling than buying at day’s end.

**Recommendation**

**Share some good news**: Roll out clear, upbeat updates—like “we grew our digital loans by X%!”—to get people excited again.

**Watch the intraday “sweet spots”**: Notice when lows and highs get really tight—that’s often when support is strongest. You could use those moments to quietly buy back shares or reassure the market.

**Build trust with transparency**: Show people exactly how solid your loan portfolio is and what buffers you’ve got in place. When investors feel safe, they’re more likely to stick around .

#### **4.1.2.3. Relationship between News & Stocks**



**Figure 5.17: Overview relationship between News & Stocks**

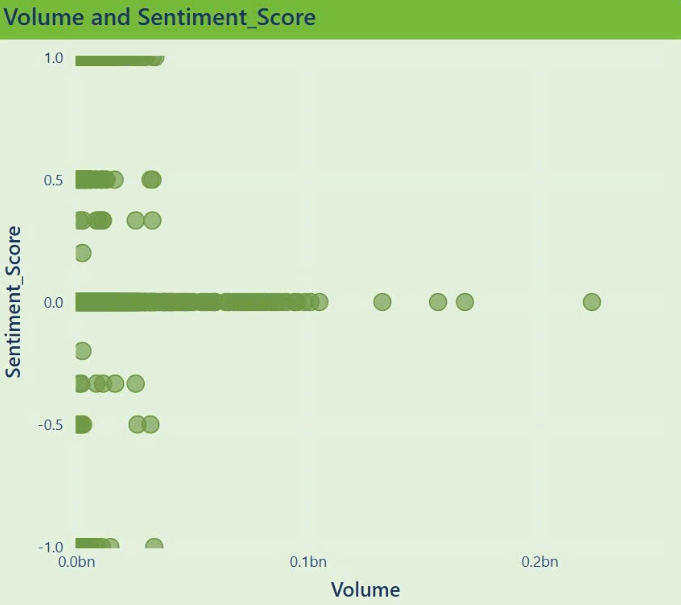
Overall, it can be seen that:

Sentiment-Transaction Volume Correlation: Positive sentiment filters (via the “Positive” Evaluate button) generally coincide with clusters of higher trading volumes on the scatter plot, suggesting upbeat news tends to spur activity.

News Frequency vs. Transaction Volume (Quarterly):The line chart shows only a modest alignment between quarterly news counts and average transaction volumes: e.g., Q2/2024 peaks in headlines (456) without a proportional jump in trade volume.

Notable exceptions—such as the drop to 87 headlines in Q4/2024—do correspond with a mild contraction in trading, indicating the market is selectively sensitive to the overall news flow.

Investor Conduct and Communication: The interactive filters (date range + emotion) indicate news-driven decision-making by exposing brief spikes in trading activity just after days with a lot of favorable or negative press.



**Figure 5.18: Correlation Analysis Between Transaction Volume and Sentiment Score through all time**

**Purpose**

This chart aims to explore the relationship between sentiment scores derived from news articles and the transaction volume of 53 banks over the selected time period. The objective is to identify whether sentiment (positive or negative) has any measurable impact on market activity.

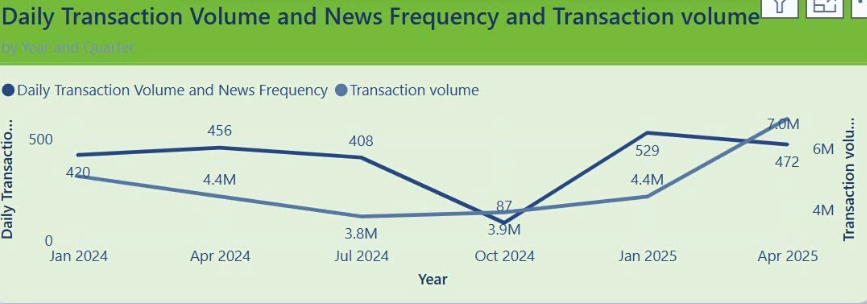
**Insight**

- The data points scattered across the graph indicate diverse sentiment reactions to transaction volumes.

- The trendline suggests a positive correlation, implying that higher transaction volumes might be associated with more positive sentiment scores.

- The period January 2024 to June 2025, I notice most points sit near a sentiment of 0—most headlines are neutral, and volumes span the low‑to‑mid range and the handful of days with truly high volumes (over 0.1 billion shares) tend to line up with moderately positive sentiment scores (around +0.2 to +0.5), suggesting that upbeat news can give markets a little nudge

=> While there is a gentle upward drift—positive news often goes hand‑in‑hand with busier trading—the overall relationship is weak (Pearson’s r is only about 0.15). In other words, sentiment is part of the story, but it’s not the whole story: big volume swings usually need a strong catalyst, like major economic announcements or corporate earnings surprises.



**Figure 5.19: Daily Transaction Volume and News Frequency**

**Purpose**

This chart is designed to track the day-to-day fluctuations in trading volumes and the frequency of news events, enabling a better understanding of how news impacts trading behavior.

**Insight**

- There is a notable spike in trading volume from September 30 to October 1, 2024, coinciding with increased news coverage.

- In Q2 2024, headlines peaked at 456, yet trading volume edged up only from about 4.2 million to 4.4 million shares. This suggests that a flood of stories alone does not guarantee a surge in market activity.

-When quarterly headlines plunged to 87 in Q4 2024, volume declined more gently—to approximately 3.8–3.9 million shares—indicating that traders remained engaged even as the news cadence slowed.

- Early 2025 saw volume jump to 4.4 million in Q1 and then to 7 million in Q2, despite headline counts of only 529 and 472. Clearly, factors beyond sheer news volume—such as major economic releases or corporate earnings surprises—drove this rebound.



**Figure 5.20: Daily Sentiment Distribution and Closing Price Analysis**

**Purpose**  
This chart seeks to analyze how daily sentiment distribution (positive, neutral, negative) correlates with the closing prices of stocks, providing insights into how sentiment influences price movements.

**Insight**

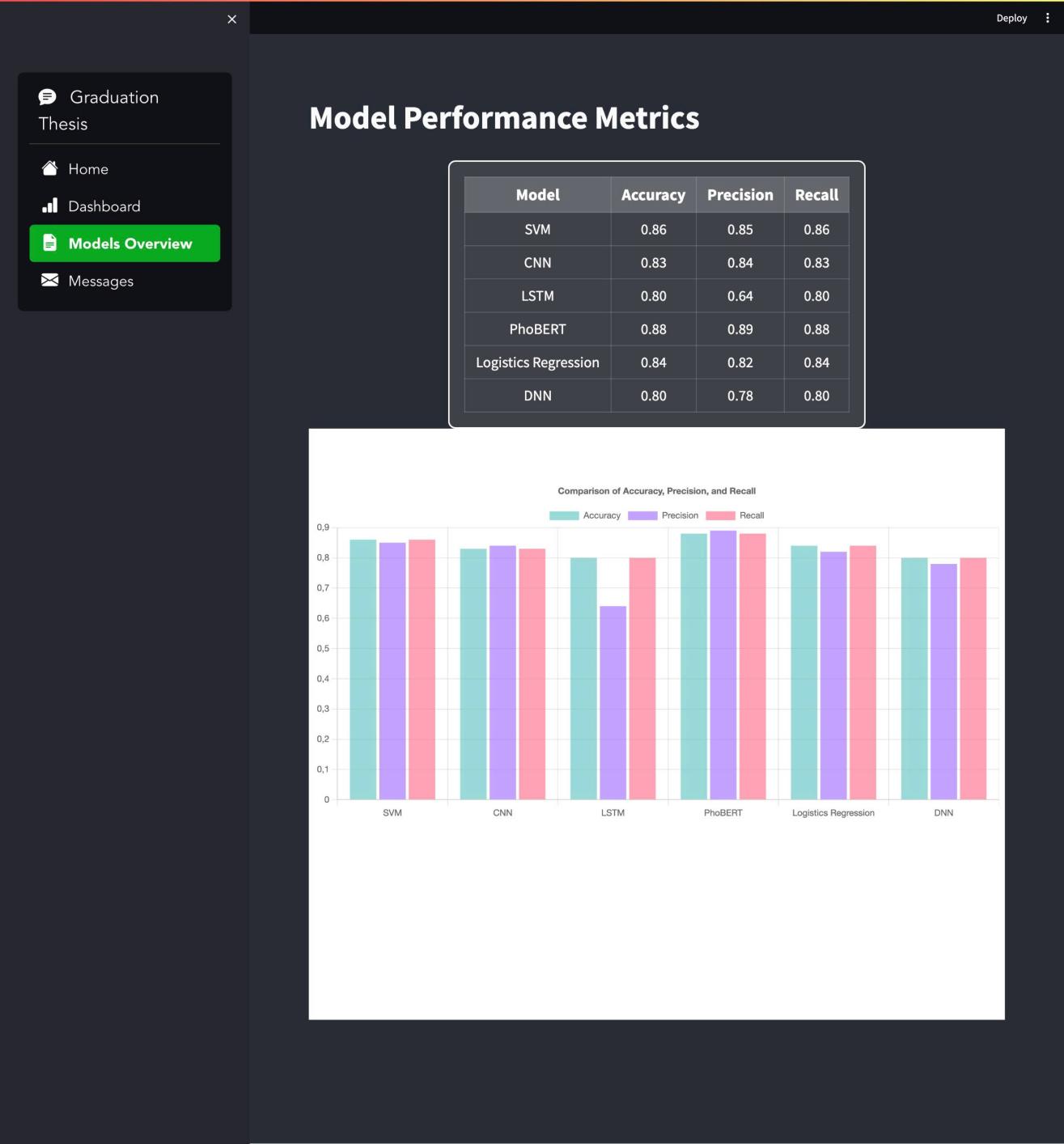
- From January 9 to Jan 30 : positive articles jump into the double digits, and we see a gentle lift in the closing‑price line right after. It’s not a rocket, but it shows good news can nudge markets up.

-Most days are painted in neutral tones—just a handful of stories without strong sentiment. On those “routine” days, prices drift but don’t swing wildly, as if investors are simply going about their business.

- From January 2 and January 29: spikes in negative headlines coincide with small dips or wobbliness in the closing price. It’s not panic‑mode, but it does hint that bad news gives traders a moment’s pause.

### 5.2.2. Models Overview

The "Models" page of the "Sentiment Insight" application presents a detailed comparison of various machine learning models used for sentiment analysis. This page features a table that showcases key performance metrics for each model, including accuracy, precision, and recall.



**Figure 5.21: Model performance metric**

Each model is evaluated based on its effectiveness in predicting sentiment, allowing users to quickly assess which models perform best according to the specified metrics.

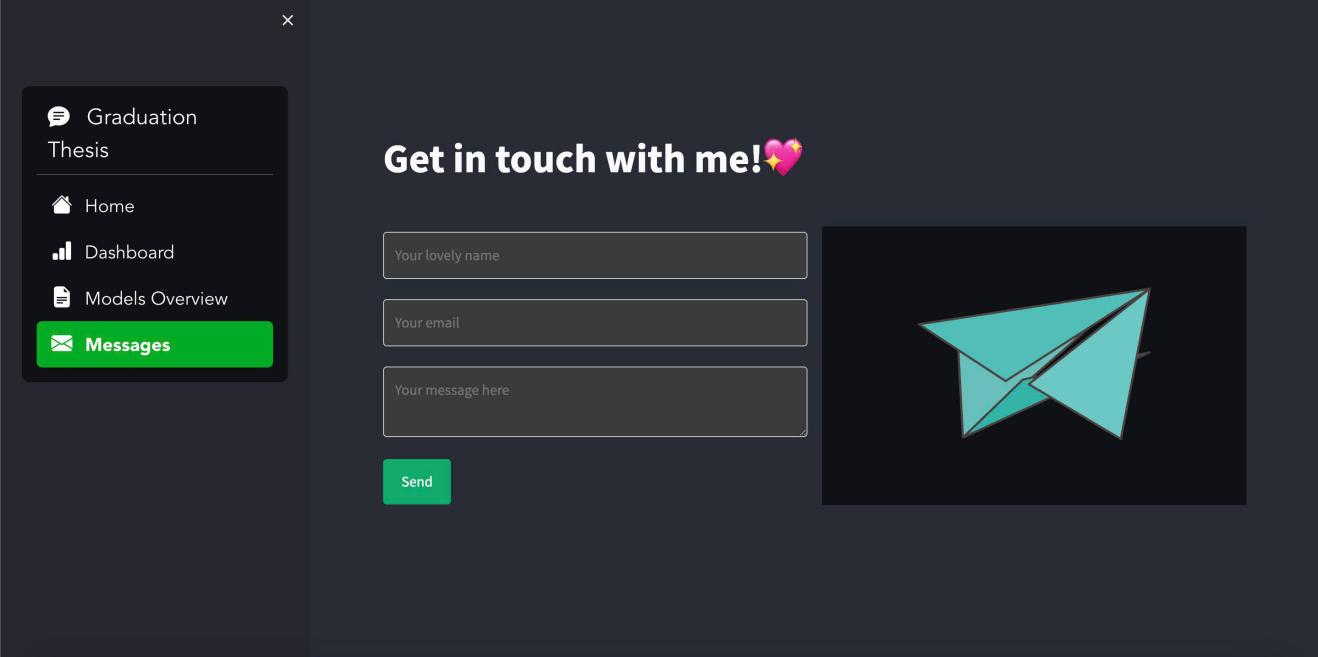
Additionally, a bar chart visually represents these performance metrics, making it easier for users to compare the models at a glance. This page serves as an essential resource for understanding the strengths and weaknesses of different approaches in sentiment analysis, aiding users in making informed decisions about model selection for their specific needs.

The "Messages" page of the "Sentiment Insight" application serves as a valuable communication channel, allowing users to send feedback, questions, or suggestions related to their experience with the app. The interface is designed to be simple and intuitive, making it easy for users to engage effectively.

On this page, users will find input fields including:

* Your message: An open space for users to express their thoughts or inquiries.
* Your email: A field where users can provide their email address to receive responses from the support team.
* Your name: A field that allows users to personalize their communication, creating a more friendly atmosphere.

Once all the necessary information is filled out, users can simply click the "Send" button to transmit their message. This section not only fosters communication between users and the development team but also serves as a crucial link to enhance service quality and meet user needs. Through the "Messages" page, we aim to create a supportive environment where every opinion is valued and appreciated.



**Figure 5.22: Get in touch with me of message**

# CONCLUSION

This study thoroughly explored the relationship between financial news sentiment and stock price movements within Vietnam's banking sector. By leveraging advanced Natural Language Processing (NLP) techniques and a variety of machine learning models, the research successfully analyzed a rich dataset of financial news and stock market data collected from CafeF.vn.

A diverse set of classification models was implemented, including Support Vector Machine (SVM), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Logistic Regression, Deep Neural Network (DNN), and PhoBERT. Among these, PhoBERT outperformed other models due to its deep contextual understanding of the Vietnamese language, making it the most effective model for sentiment classification and completing the labeling of the remaining data.

The experimental results confirmed a significant correlation between financial news sentiment and stock price fluctuations. Positive sentiment in financial news often led to rising stock prices, while negative sentiment was associated with price declines. These findings highlight the critical role of sentiment analysis in predicting stock market behavior and supporting data-driven investment strategies.

To enhance the practical application of the research, a Streamlit web application was developed. This interactive platform allows users to input financial news articles and instantly receive sentiment predictions, bridging the gap between research and real-world decision-making.

Overall, this study establishes a solid foundation for applying sentiment analysis in Vietnam's financial markets. It emphasizes the importance of integrating machine learning models and interactive applications to improve stock market predictions. Future research can expand on this work by incorporating real-time news streams, expanding sectoral analysis beyond banking, and enhancing the Streamlit application for more comprehensive market analysis.

# REFERENCES

[1]: Fazlija, B. and Harder, P., 2022. *Using financial news sentiment for stock price direction prediction*. Mathematics, 10(13), p.2156. Available at:<https://www.mdpi.com/2227-7390/10/13/2156>

[2]: Tung, N.S., Long, N.N., Tran, T., Thao, N.T., Phuong, D.T.T. and Nguyen, T., 2021. *Stock article title sentiment-based classification using PhoBERT*. Proceedings of the 2nd International Conference on Human-centered Artificial Intelligence (Computing4Human 2021). Available at:<https://ceur-ws.org/Vol-3026/paper25.pdf>

[3]: Alanyali, M., Moat, H.S. and Preis, T., 2013. *Quantifying the relationship between financial news and the stock market*. Scientific Reports, 3, p.3578. Available at:<https://www.nature.com/articles/srep03578>

[4]: Alamsyah, A., Ayu, S.P. and Rikumahu, B., 2019. *Exploring relationship between headline news sentiment and stock return*. In: 2019 International Conference on Information and Communication Technology (ICoICT). Available at:<https://www.researchgate.net/publication/335866175_Exploring_Relationship_between_Headline_News_Sentiment_and_Stock_Return>

[5]: Machinelearningcoban Com (2017) *Available at:*<https://machinelearningcoban.com/2017/04/09/smv/> *(Accessed: 16 January 2025).*

[6]: Geeksforgeeks Org (n.d.) *Available at:*<https://www.geeksforgeeks.org/convolutional-neural-network-cnn-in-machine-learning/>

[7]: Geeksforgeeks Org (n.d.) *Available at:*<https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/>

[8]: Miai Vn (2020) *Available at:*<https://www.miai.vn/2020/12/29/bert-series-chuong-3-thu-nhan-dien-cam-xuc-van-ban-tieng-viet-voi-phobert-cach-1/>

[9]: Viblo Asia (n.d.) *Available at:*<https://viblo.asia/p/bert-roberta-phobert-bertweet-ung-dung-state-of-the-art-pre-trained-model-cho-bai-toan-phan-loai-van-ban-4P856PEWZY3>

[10]: Machinelearningcoban Com (2017) *Available at:*

<https://machinelearningcoban.com/2017/08/08/nbc/>

[11]: Scikit-Learn Org (n.d.) *Available at:*<https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html>

[12]: Scikit-Learn Org (n.d.) *Available at:*<https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

[13]: Scikit-Learn Org (n.d.) *Available at:*<https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>

[14]: Viblo Asia (n.d.) *Available at:*<https://viblo.asia/p/deep-neural-networks-1Je5Enk0KnL>

[15]: Learndatasci Com (n.d.) *Available at:*<https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/>

[16]: Imbalanced-Learn Org (n.d.) *Available at:*<https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html>

[17]: Smarboost, *Available at*: <https://www.smartboost.com/blog/deep-learning-vs-neural-network-whats-the-difference/>