Research Project

Master of Computer Science with specialization in Finance
Illinois Institute of Technology - 2023

Gabriel Baranes

Investor sentiment variation across regions: a Natural Language Processing analysis

Outline

1. Introduction

2. Sentiment Analysis

3. Transformers and BERT model

4. Model methodology and analysis

5. Conclusive remarks

1. Introduction

Information and market prices

- Information plays a **fundamental role** in financial markets
  - ➢ One of the main idea :

    - Full information increases market efficiency
    - And, price reflects the true value of the tarded good

  - ➢ Asset prices reflect all available information …
    … in line with the Efficient Market Hypothesis (Fama, 1970)

- A large **literature** studies the link between **information** and **asset prices**
  - ➢ Economic and financial literature
  - ➢ More rencently a growing literature in CS and AI

1. Introduction

News and strategic approaches in investing

- Traders heavily rely on **information** to make **buying** or **selling** decisions
  - ➢ Investors who have better information can strategically use it to make their investment decisions
  - ➢ Can twitter data serve as reliable predictors of daily stock price movements?
- Significant growth of **online financial press** and abundant information on social networks
  - ➢ Need for investors to exploit this amount of information for strategic purposes
  - ➢ and extract market sentiment for strategic purposes
- **Machine Learning** models
  - ➢ Combining sentiment analysis and news to forecast fluctuation in stock prices
  - ➢ Use of NLP models

1. Introduction

The case of multilisted companies

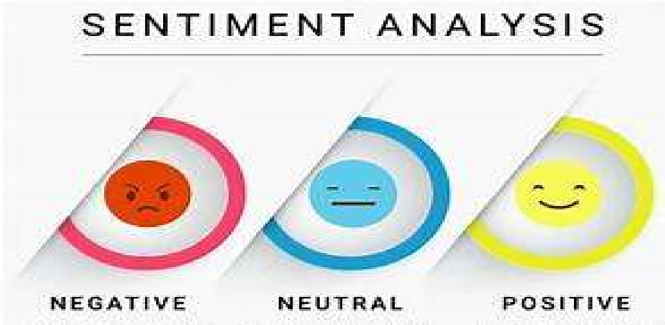
- **Multi-listed** companies (NYSE, Euronext, LSE…) … stock prices **may differ** across stock markets:
  - ➢ Different market conditions and regulations
  - ➢ But also, **different market sentiments** across marketplaces / countries

- The **research project**:
  - ➢ Use a NLP model to conduct a sentiment analysis on press articles related to a specific company
  - ➢ Compare sentiments in two different countries for the same company
  - ➢ Illustrative example: French company LVMH, listed on Euronex and NYSE

2. Sentiment analysis

Text classification

Text classification in NLP involves categorizing and assigning predefined labels or categories to text documents, sentences, or phrases based on their content



- Data Collection
- Text pre- processing
- Sentiment Detection
- Sentiment Classification
- Presentation of Output

2. Sentiment analysis

Text classification

- **Data** colletion
  - ➢ Various sources: news articles, social media, blogs,…
  - ➢ API/ Web scrapping
- Text **pre-processing**
  - ➢ Preparing the text
  - ➢ Removing non informative passages (HTML tags, scripts, advertisments…)
  - ➢ Tokenization
- Sentiment **detection** and **classificaion**
  - ➢ Examining sentences
  - ➢ Categorizing subjective sentences
  - ➢ Classification

2. Sentiment analysis

Classification methods

Based on Lexicon

- Algorithms focus on the words themselves
- Certain words express a positive sentiment, while others strictly express a negative sentiment
- Example:
  - ➢ The corpus-based approach
  - ➢ The dictionary-based approach

Based on AI

- ML approaches use supervised, unsupervised, or reinforcement learning
- Classify words, sentences, or entire documents into sentiment categories
- Example:
  - ➢ MNB classifier
  - ➢ Maximum Entropy classifier
  - ➢ Decision Trees
  - ➢ Artificial Neural Networks

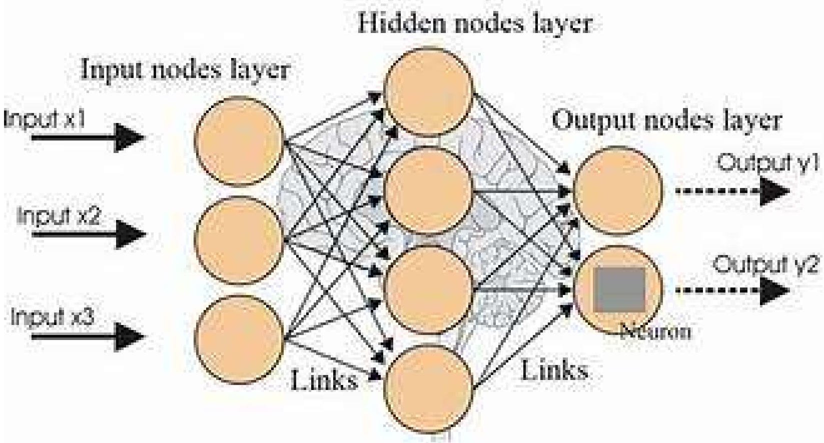
3. Transformers and BERT model

Transformers

- What are **Transformers**?
  - Type of neural network architecture
  - Mainly used in the field of NLP and sequence understanding

- Introduced by **Vaswani** in a paper called "Attention is all you need"

- Transformer is a model that relies on **self-attention**



(Neural Network Architecture)

3. Transformers and BERT model

BERT model

- BERT, which stands for "Bidirectional Encoder Representations from Transformers," is a popular natural language processing (NLP) model developed by **Google AI** in 2018.

- BERT is designed to understand the **contextual relationships** and **nuances of words** in a sentence or a document

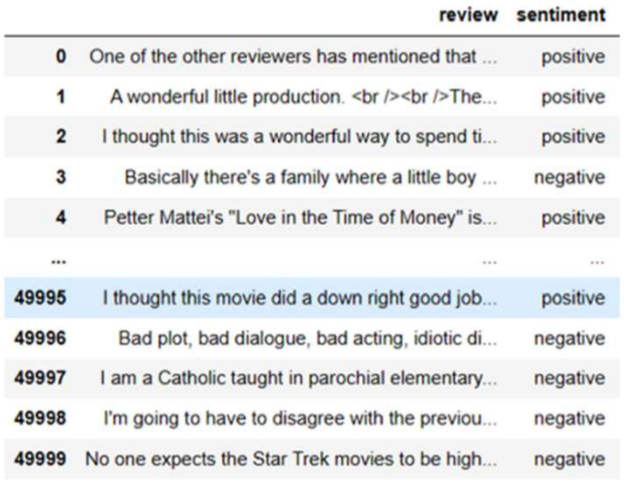
BERT:

- ➢ Transformer Architecture
- ➢ Bidirectional Context
- ➢ Pre-Training
- ➢ Fine Tuning
- ➢ Applications

4. Model methodology and analysis

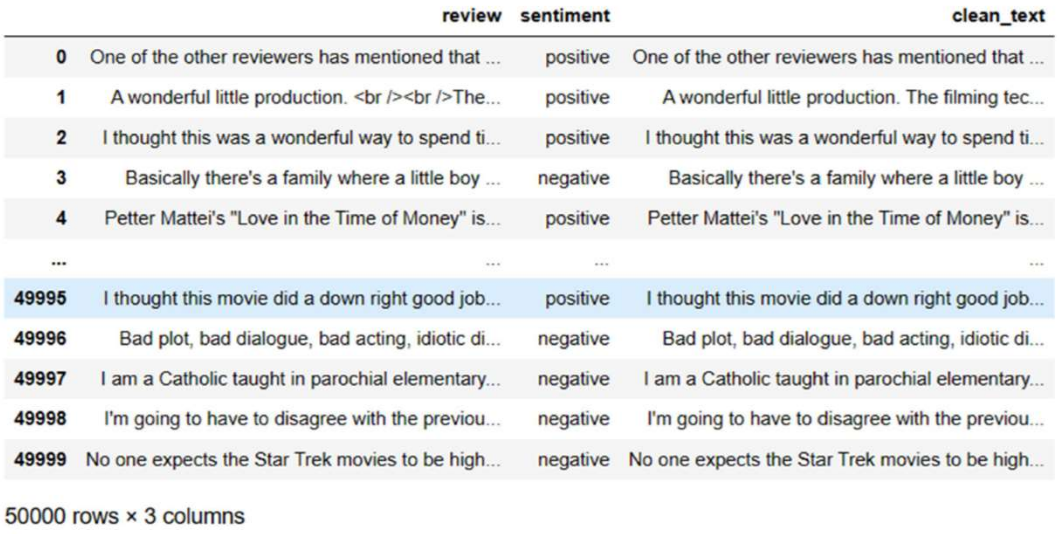
The sentiment analysis (1): Dataset

- IMBD dataset dowloaded from Kaggle
  - 50,000 movie reviews
  - positive/negative sentiment



	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. <br /><br />The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
...	...	...
49995	I thought this movie did a down right good job...	positive
49996	Bad plot, bad dialogue, bad acting, idiotic di...	negative
49997	I am a Catholic taught in parochial elementary...	negative
49998	I'm going to have to disagree with the previou...	negative
49999	No one expects the Star Trek movies to be high...	negative

Figure 1: Extract from the IMDB Dataset



	review	sentiment	clean_text
0	One of the other reviewers has mentioned that ...	positive	One of the other reviewers has mentioned that ...
1	A wonderful little production. <br /><br />The...	positive	A wonderful little production. The filming tec...
2	I thought this was a wonderful way to spend ti...	positive	I thought this was a wonderful way to spend ti...
3	Basically there's a family where a little boy ...	negative	Basically there's a family where a little boy ...
4	Petter Mattei's "Love in the Time of Money" is...	positive	Petter Mattei's "Love in the Time of Money" is...
...	...	...	...
49995	I thought this movie did a down right good job...	positive	I thought this movie did a down right good job...
49996	Bad plot, bad dialogue, bad acting, idiotic di...	negative	Bad plot, bad dialogue, bad acting, idiotic di...
49997	I am a Catholic taught in parochial elementary...	negative	I am a Catholic taught in parochial elementary...
49998	I'm going to have to disagree with the previou...	negative	I'm going to have to disagree with the previou...
49999	No one expects the Star Trek movies to be high...	negative	No one expects the Star Trek movies to be high...

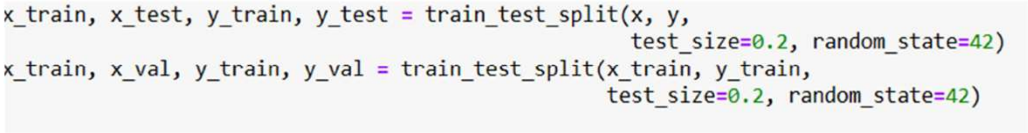
50000 rows × 3 columns

Figure 3: Extract from the IMDB Dataset after text cleaning

4. Model methodology and analysis

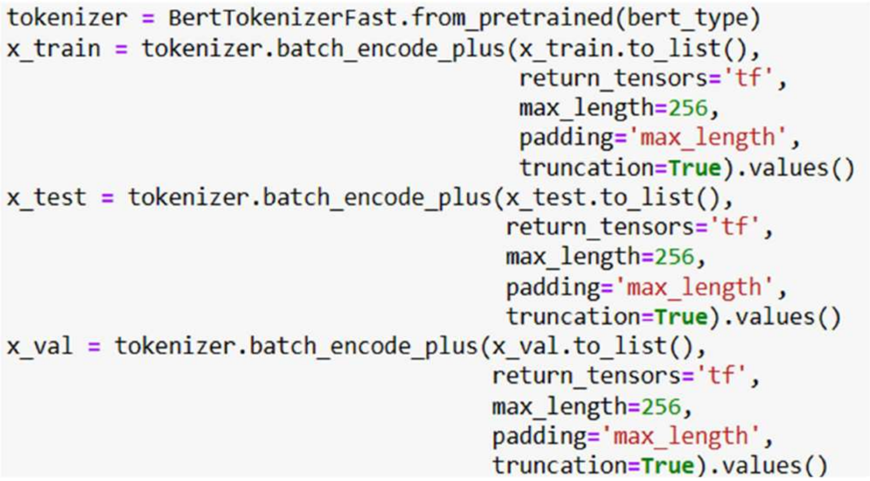
The sentiment analysis (2): Data processing

- Preparing the data to fine tune the model
  - ➢ Data split
  - ➢ Tokenization using the **Bert Tokenizer**
  - ➢ vectorization



```
x_train, x_test, y_train, y_test = train_test_split(x, y,
                                                    test_size=0.2, random_state=42)
x_train, x_val, y_train, y_val = train_test_split(x_train, y_train,
                                                  test_size=0.2, random_state=42)
```

code required for data splitting



```
tokenizer = BertTokenizerFast.from_pretrained(bert_type)
x_train = tokenizer.batch_encode_plus(x_train.to_list(),
                                      return_tensors='tf',
                                      max_length=256,
                                      padding='max_length',
                                      truncation=True).values()
x_test = tokenizer.batch_encode_plus(x_test.to_list(),
                                     return_tensors='tf',
                                     max_length=256,
                                     padding='max_length',
                                     truncation=True).values()
x_val = tokenizer.batch_encode_plus(x_val.to_list(),
                                    return_tensors='tf',
                                    max_length=256,
                                    padding='max_length',
                                    truncation=True).values()
```

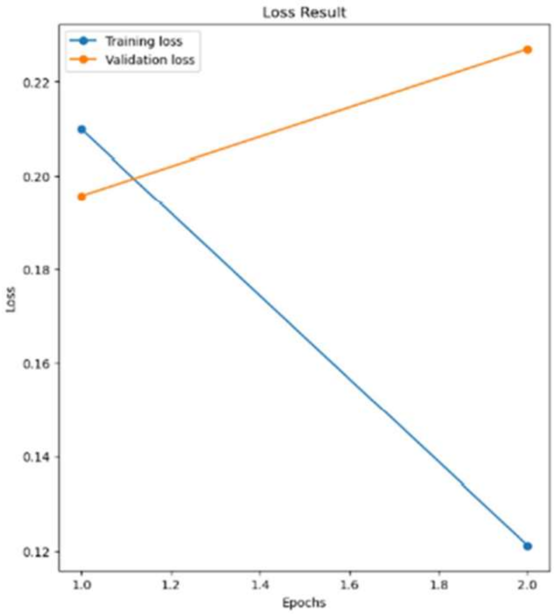
code required for tokenization and vectorization

4. Model methodology and analysis

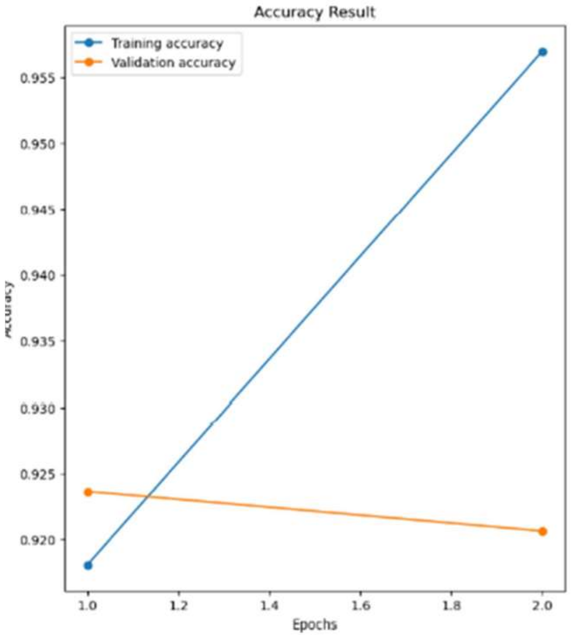
The sentiment analysis (3): The model

- Hyperparameters:
  - ➢ learning rate of 2.10-5
  - ➢ number of epoch=2
  - ➢ batch size = 32

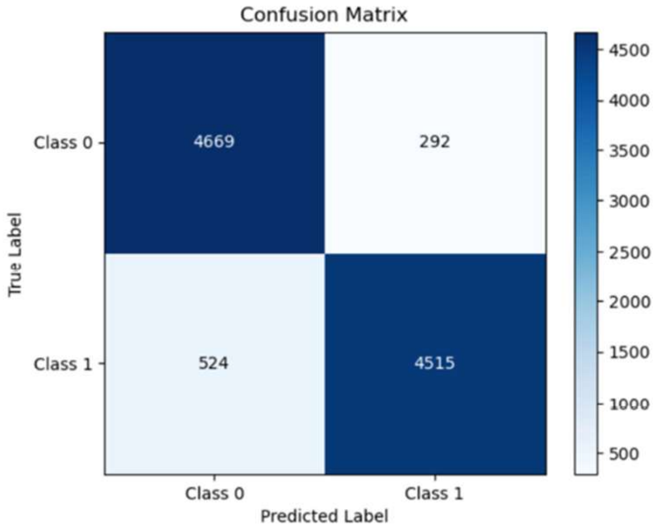
- Model’s Performance:
  - ➢ validation loss = 0.22
  - ➢ validation accuracy = 0.92
  - ➢ F1-score = 0.92



Training loss and validation loss



Training accuracy and validation accuracy



Class 0 = « Negative »

Class 1 = « Positive »

Confusion matrix

4. Model methodology and analysis

Data Scrapping

- Comparative analysis by examining **American articles** in **English** from **American sources** and **French articles** in **French** from **French sources**

- Sentiment analysis model, which operates exclusively on English text
  - ➢ Translation: library Translator from googletrans
  - ➢ Translator has limitations: 500 characters at a time
  - ➢ saving .txt files containing the translated English text

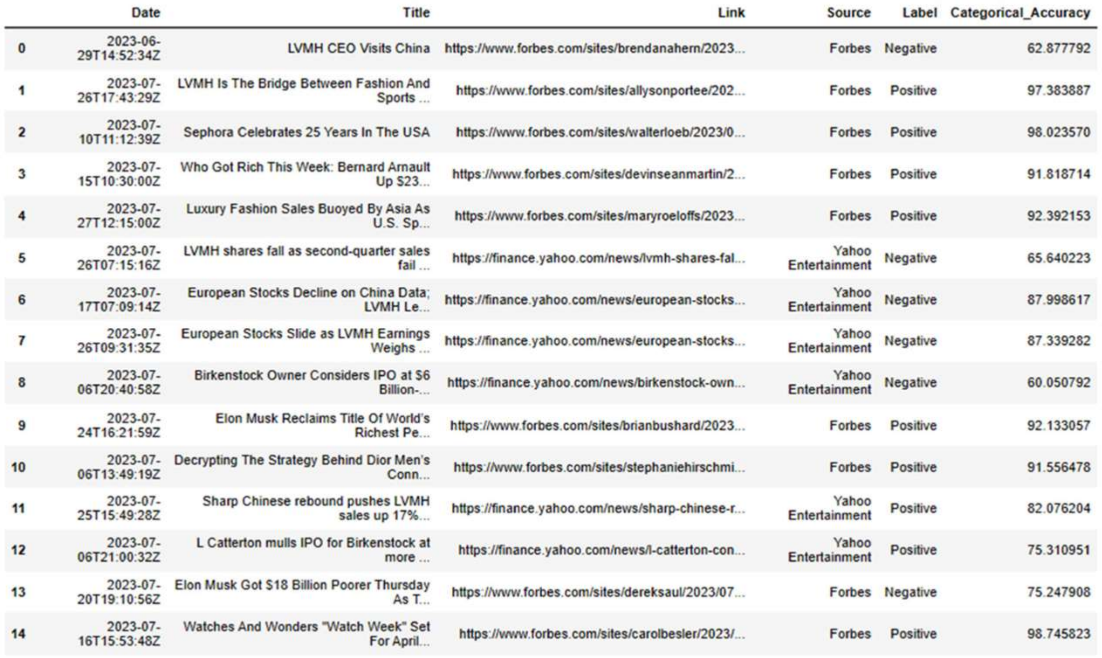
- Translation and data preparation for the sentiment analysis model
  - ➢ News API
  - ➢ Date range (30 days)
  - ➢ List of sources
  - ➢ Extracting text from HTML pages using BeautifulSoup
  - ➢ saving .txt files containing the text from the article

4. Model methodology and analysis

Complete code

- The complete code utilizes the Sentiment Analysis NLP model and the data scraping code

- Key-Word
- Data Scrapping
- Sentiment Analysis evaluation
- Results



	Date	Title	Link	Source	Label	Categorical_Accuracy
0	2023-06-29T14:52:34Z	LVMH CEO Visits China	https://www.forbes.com/sites/brendanahern/2023...	Forbes	Negative	62.877792
1	2023-07-26T17:43:29Z	LVMH Is The Bridge Between Fashion And Sports ...	https://www.forbes.com/sites/allysonportee/202...	Forbes	Positive	97.383887
2	2023-07-10T11:12:39Z	Sephora Celebrates 25 Years In The USA	https://www.forbes.com/sites/walterloeb/2023/0...	Forbes	Positive	98.023570
3	2023-07-15T10:30:00Z	Who Got Rich This Week: Bernard Arnault Up $23...	https://www.forbes.com/sites/devinseanmartin/2...	Forbes	Positive	91.818714
4	2023-07-27T12:15:00Z	Luxury Fashion Sales Buoyed By Asia As U.S. Sp...	https://www.forbes.com/sites/maryroeloffs/2023...	Forbes	Positive	92.392153
5	2023-07-26T07:15:16Z	LVMH shares fall as second-quarter sales fail ...	https://finance.yahoo.com/news/lvmh-shares-fal...	Yahoo Entertainment	Negative	65.640223
6	2023-07-17T07:09:14Z	European Stocks Decline on China Data; LVMH Le...	https://finance.yahoo.com/news/european-stocks...	Yahoo Entertainment	Negative	87.998617
7	2023-07-26T09:31:35Z	European Stocks Slide as LVMH Earnings Weighs ...	https://finance.yahoo.com/news/european-stocks...	Yahoo Entertainment	Negative	87.339282
8	2023-07-06T20:40:58Z	Birkenstock Owner Considers IPO at $6 Billion-...	https://finance.yahoo.com/news/birkenstock-own...	Yahoo Entertainment	Negative	60.050792
9	2023-07-24T16:21:59Z	Elon Musk Reclaims Title Of World's Richest Pe...	https://www.forbes.com/sites/brianbushard/2023...	Forbes	Positive	92.133057
10	2023-07-06T13:49:19Z	Decrypting The Strategy Behind Dior Men's Conn...	https://www.forbes.com/sites/stephaniehirschmi...	Forbes	Positive	91.556478
11	2023-07-25T15:49:28Z	Sharp Chinese rebound pushes LVMH sales up 17%...	https://finance.yahoo.com/news/sharp-chinese-r...	Yahoo Entertainment	Positive	82.076204
12	2023-07-06T21:00:32Z	L Catterton mulls IPO for Birkenstock at more ...	https://finance.yahoo.com/news/l-catterton-con...	Yahoo Entertainment	Positive	75.310951
13	2023-07-20T19:10:56Z	Elon Musk Got $18 Billion Poorer Thursday As T...	https://www.forbes.com/sites/dereksaul/2023/07...	Forbes	Negative	75.247908
14	2023-07-16T15:53:48Z	Watches And Wonders "Watch Week" Set For April...	https://www.forbes.com/sites/carolbesler/2023/...	Forbes	Positive	98.745823

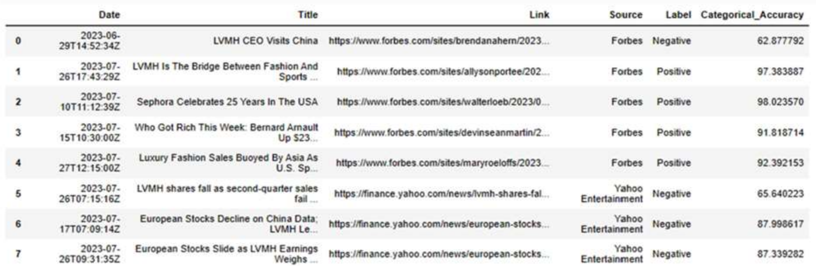
```
Number of ‘Positive’ labels: 9
Number of ‘Negative’ labels: 6
```

Results for American articles on LVMH

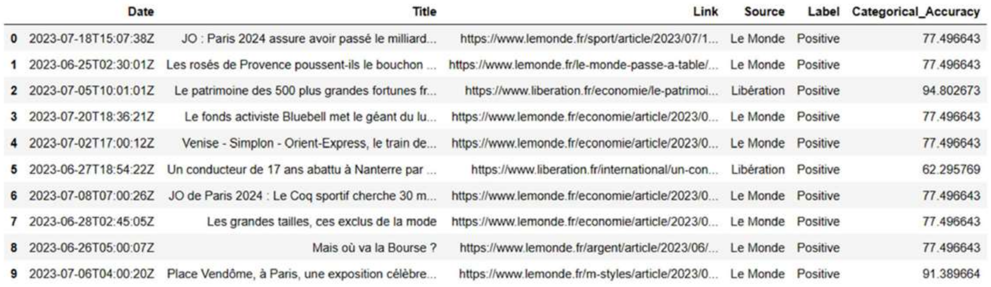
4. Model methodology and analysis

Complete code

- The complete code utilizes the Sentiment Analysis NLP model and the data scraping code



	Date	Title	Link	Source	Label	Categorical_Accuracy
0	2023-06-29T14:52:34Z	LVMH CEO Visits China	https://www.forbes.com/sites/brendanahern/2023...	Forbes	Negative	62.877792
1	2023-07-26T17:43:29Z	LVMH Is The Bridge Between Fashion And Sports ...	https://www.forbes.com/sites/allysonportee/202...	Forbes	Positive	97.383887
2	2023-07-10T11:12:39Z	Sephora Celebrates 25 Years In The USA	https://www.forbes.com/sites/walterloeb/2023/0...	Forbes	Positive	98.023570
3	2023-07-15T10:30:00Z	Who Got Rich This Week: Bernard Arnault Up $23...	https://www.forbes.com/sites/devinseanmartin/2...	Forbes	Positive	91.818714
4	2023-07-27T12:15:00Z	Luxury Fashion Sales Buoyed By Asia As U.S. Sp...	https://www.forbes.com/sites/maryroeloffs/2023...	Forbes	Positive	92.392153
5	2023-07-26T07:15:16Z	LVMH shares fall as second-quarter sales fail ...	https://finance.yahoo.com/news/lvmh-shares-fal...	Yahoo Entertainment	Negative	65.640223
6	2023-07-17T07:09:14Z	European Stocks Decline on China Data; LVMH Le...	https://finance.yahoo.com/news/european-stocks...	Yahoo Entertainment	Negative	87.998617
7	2023-07-26T09:31:35Z	European Stocks Slide as LVMH Earnings Weighs ...	https://finance.yahoo.com/news/european-stocks...	Yahoo Entertainment	Negative	87.339282



	Date	Title	Link	Source	Label	Categorical_Accuracy
0	2023-07-18T15:07:38Z	JO : Paris 2024 assure avoir passé le milliard...	https://www.lemonde.fr/sport/article/2023/07/1...	Le Monde	Positive	77.496643
1	2023-06-25T02:30:01Z	Les rosés de Provence poussent-ils le bouchon ...	https://www.lemonde.fr/le-monde-passe-a-table/...	Le Monde	Positive	77.496643
2	2023-07-05T10:01:01Z	Le patrimoine des 500 plus grandes fortunes fr...	https://www.liberation.fr/economie/le-patrimoi...	Libération	Positive	94.802673
3	2023-07-20T18:36:21Z	Le fonds activiste Bluebell met le géant du lu...	https://www.lemonde.fr/economie/article/2023/0...	Le Monde	Positive	77.496643
4	2023-07-02T17:00:12Z	Venise - Simplon - Orient-Express, le train de...	https://www.lemonde.fr/economie/article/2023/0...	Le Monde	Positive	77.496643
5	2023-06-27T18:54:22Z	Un conducteur de 17 ans abattu à Nanterre par ...	https://www.liberation.fr/international/un-con...	Libération	Positive	62.295769
6	2023-07-08T07:00:26Z	JO de Paris 2024 : Le Coq sportif cherche 30 m...	https://www.lemonde.fr/economie/article/2023/0...	Le Monde	Positive	77.496643
7	2023-06-28T02:45:05Z	Les grandes tailles, ces exclus de la mode	https://www.lemonde.fr/economie/article/2023/0...	Le Monde	Positive	77.496643
8	2023-06-26T05:00:07Z	Mais où va la Bourse ?	https://www.lemonde.fr/argent/article/2023/06/...	Le Monde	Positive	77.496643
9	2023-07-06T04:00:20Z	Place Vendôme, à Paris, une exposition célèbre...	https://www.lemonde.fr/m-styles/article/2023/0...	Le Monde	Positive	91.389664

	Number of article	Positive articles	Negative articles	Mean of positive Categorical Accuracy	Mean of negative Categorical Accuracy
French	10	10	0	79.1%	N/A
American	15	9	6	91%	73%

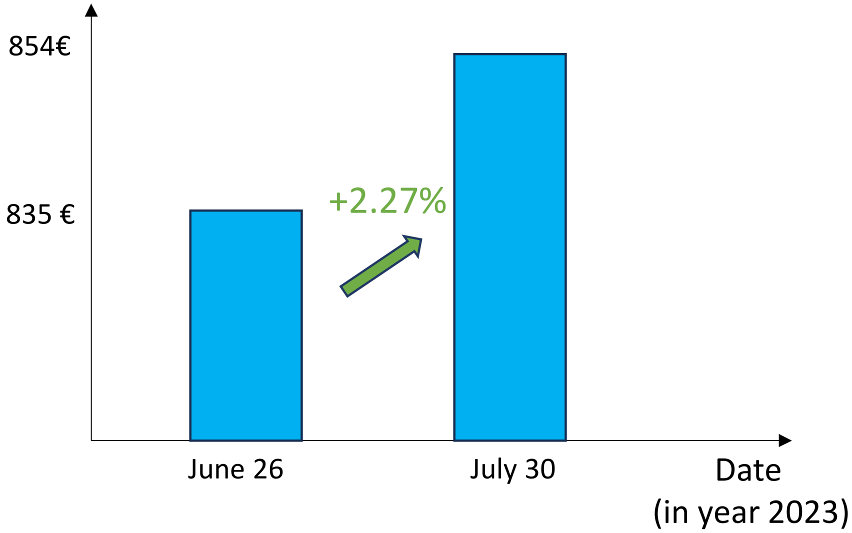
4. Model methodology and analysis

Discussion

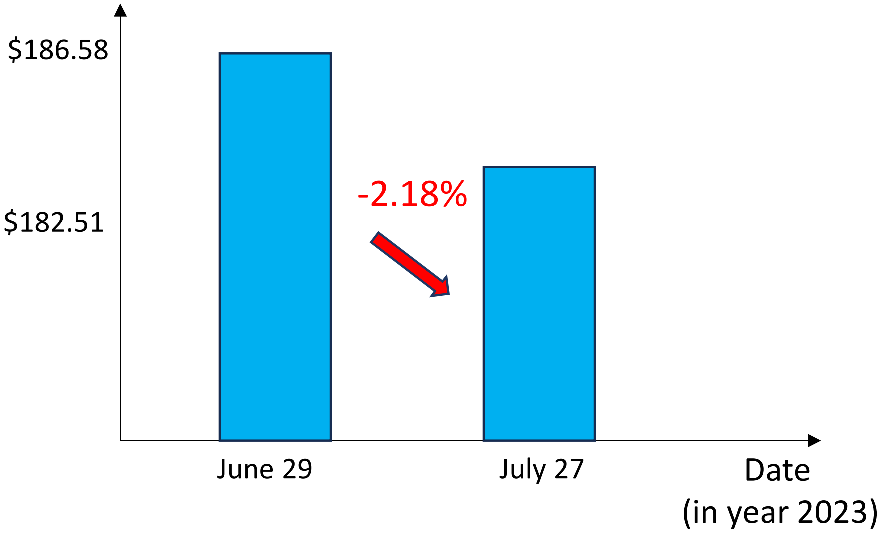
- Investor sentiment can vary across regions
- Different sentiments could contribute to explain why, for the same company the stock price and its variation could differ across stock markets

	Number of article	Positive articles	Negative articles	Mean of positive Categorical Accuracy	Mean of negative Categorical Accuracy
French	10	10	0	79.1%	N/A
American	15	9	6	91%	73%

Stock price of LVMH on EuroNext



Stock price of LVMH on New York Stock Exchange



5. Conclusive remarks

- Certain limitations that offer opportunities for improvement
  - ➢ news APIs
  - ➢ online scrapping
  - ➢ translation

Thank you for your atention!