**Erasmus University, Learning from Big Data: Module 1 - Final Assignment**

# Relationship between Tweets sentiment and Stock price

Author: Semir Arabo

Date: 09/22/2024

**Data Sources:**

**UGC dataset** (Tweets about Nvidia) was retrieved from here and uploaded as "Nvidia-Tweets.csv"

**Corpus** used for Lexicon creation (a dataset of labaled Tweets) was retrieved from here and uploaded as "training.160000000.processed.noemoticon.csv"

**Financial data** (Nvidia stock price) were retrieved from Yahoo finance using "yfinance" library

## #1 Introduction

Behavioral finance theory expects that the emotions and moods in society play a role in investment decision-making, thus influencing the financial markets. Researchers suggest that the social media buzz is a proxy of short-term public sentiment that could help to predict changes in stock prices (Rao et al., 2012). Nowadays, data-processing techniques allow us to retrieve this sentiment from social media posts. One of those techniques is the Naive Bayesian Classifier (NBC).

## # 2 Research Question

This work tries to examine the use of Naive Bayesian Classifier Sentiment Analysis to capture this public sentiment from social media and test if it is a relevant predictor for financial markets. By examining the relationship between tweets related to Nvidia Corporation and stock price fluctuations of this company we answer the question „Does capturing public sentiment by NBC Sentiment Analysis (from Twitter data) predict the stock price fluctuations of Nvidia Corporation?“.

By evaluating the NBC for sentiment analysis, this research aims to establish its potential in accurately forecasting financial market trends. Demonstrating the predictive power of social media sentiment on stock prices could offer practical insights for investors and contribute to the research on the intersection of technology, finance, and behavioral economics.

To express the stock price fluctuations we will utilize the changes in prices on the closed (Nasdaq) stock market since these changes are influenced by new „events“ happening when the market is closed (Miwa, 2020). This allows us to separate concrete time periods that we can match with tweets happening at that time. For tweets, we are interested only in their „Positive/Negative Sentiment“ since this dimension is expected to be the most universal and precise predictor of stock price change. This will be done by conducting NBC based on the relevant corpus Sentiment140 (positive and negative tweets dataset). The research question will be then answered by conducting linear regression with Stock Market Changes as the independent variable and NBC Sentiment Analysis (public sentiment) as the dependent variable.

## # 3 Data

We explore the concepts from this project using a dataset of tweets about Nvidia Corporation. The tweets were collected from November 2022 to February 2023, with corresponding financial data from the same period (Yahoo finance dataset).

Each observation in the dataset represents a single tweet. The dataset includes the following variables for each tweet:

- **Tweet Text**: The textual content of the tweet.
- **Timestamp**: The exact time when the tweet was published.
- **Tweet Sentiment**: A constructed variable representing the probability that a tweet has a positive sentiment, determined using our NBC Sentiment Analysis.
- **Market Condition**: A constructed variable indicating the market condition (positive, negative, or neutral price change) during which the tweet was published.

The financial dataset includes variables for each day of the Opening and Close price of Nvidia stock. The Close price is adjusted to include any dividends, stock splits, and other corporate actions that occurred before the next day's opening, providing a more accurate reflection of the stock's value.

The dataset Sentiment140 consists from tweets collected in 2009 but thanks to its size and relevance it is suitable as a corpus. From this dataset, we will use the text of each tweet and its polarity – variable for positive or negative sentiment which was constructed using distant supervision (Bhayani et al., 2009).

Those datasets also include variables such as Tweet ID etc. that are not relevant for us.

## # 4 Data aggregation and formating

To ensure the data addresses the research question, "Does capturing public sentiment by NBC sentiment analysis from Twitter data predict the stock price fluctuations of Nvidia Corporation?", we made the following decisions:

1. **Filtering Tweets**: We filtered the dataset at first from advertisments which are not UGC and our research question. They may cause a bias since we would expect mostly positive sentiment in those tweets no matter the market conditions. We drop tweets from our dataset that include any common marketing phrases such as „Limited offer“ or „Top analyst price...“ (the second-mentioned observed from data).
   We filtered data to include only tweets posted during periods when the Nasdaq stock market was closed  (during all three types of market changes – positive, negaive or neutral change). This resulted in a final dataset of 300 tweets. This sample size is large enough for meaningful analysis while remaining manageable within our processing power limitations (required processing time tested on smaller samples).

2. **Constructing Variables**:
   - **Tweet Sentiment**: Using NBC Sentiment Analysis, each tweet is assigned a probability score indicating its positive sentiment. The process is described in part #6 Supervised learning - the Naive Bayes classifier.
   - **Market Condition**: The market condition during the time a tweet was posted is categorized based on the percentual change in Nvidia's stock price. If the percentual change between the closed adjusted price and open market price exceeds an arbitrary threshold of ±2 percentage points, the market condition is labeled accordingly.
     $Percentual\ Change = 100 \times (Close\ Price(D-1) \div Open\ Price(D) - 1)$
     This threshold ensures a balanced distribution of positive and negative changes and aligns with common research practices.

The decision to filter tweets to those posted during non-market hours aims to capture sentiments that could influence the market once it reopens. By constructing the "Tweet Sentiment" and "Market Condition" variables, we can directly correlate public sentiment with subsequent stock price movements, providing a robust framework for our analysis. This structuring ‘per tweet‘ not ‘per day‘ takes into account our dataset limitations such as that our dataset does not capture all tweets in the given day while allowing us to test our hypothesis effectively.

## #5 Supervised learning - the Naive Bayes classifier

### #5.0 Creating a lexicon

In our NBC Analysis, we are interested in the dimensions of Positive and Negative Sentiment. Therefore, we need to transform our corpus into a lexicon consisting of columns: ‘word’, ‘word count’ (the number of times the given word appears in the entire corpus), and ‘positive count’ (the number of times the given word appears in the positive portion of the corpus). We also retain the ‘negative count’ to construct negative likelihood and probability as a sanity check (both probabilities should add up to 1).

To achieve this, we must transform our corpus dataset. First, we convert all strings to lowercase since we do not expect any difference between lower- and uppercase in sentiment (differences average out to zero in this sample size). Next, we remove string punctuation and tokenize the text. We observe that positive tweets significantly outnumber negative ones in our dataset. Since we assume an equal prior probability of positive and negative sentiment for tweets, we create a corpus with an equal number of positive and negative tweets.

Then, we iterate over all words in positive tweets to count the ‘positive count’ for each word in our corpus. We do the same for negative tweets and then remove unnecessary stop words that would only increase the processing power needed. With these steps completed, our lexicon is ready.

**#5.1 Likelihoods**

Following the creation of our lexicon, we proceeded to calculate the likelihoods for each word in our corpus. These likelihoods represent the probability of a word appearing in a positive or negative tweet. In our implementation, we employed additive smoothing to handle words that might not appear in a positive (or negative) category. This approach ensures that no word has a zero probability, which could potentially disrupt our NBC calculations. We need to note that NBC Analysis has its limitations in bag-of-words problem or language ambiguity. One example is the word ‘study‘ which has two meanings – ‘to study‘ and ‘academic paper‘ for which we would assume different likelihoods. However our (positive) likelihood for ‘study‘ is 0.36 which indicates rather negative sentiment. Despite this, it is still a relevant tool and we expect most of those errors to average out.

**#5.2 NBC**

For each tweet, we take the prior probability of 0.5 to being positive (or negative). Then we iterate over all words in the given tweet using the NBC function we created. This takes the prior probability, and based on likelihoods updates this probability. Once it iterates over all words in that text we get the final posterior probability that indicates the probability of the given tweet to be positive. This is the variable that we will use to test our research question.

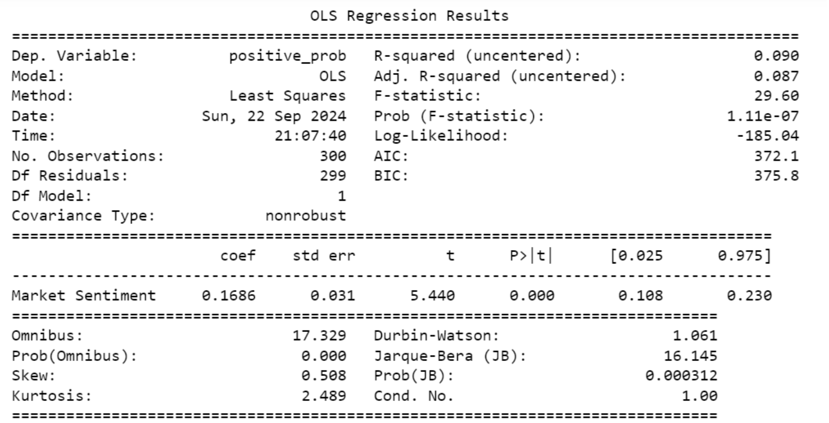
## # 6 Analysis (and NBC performance)

In this section, we evaluate the performance of the Naive Bayes Classifier (NBC) in predicting stock price fluctuations of Nvidia Corporation based on public sentiment captured from Twitter data. The main objective is to determine whether the sentiment variable created using NBC has predictive power over stock price movements.

To formally test this, we conducted a linear regression analysis with Market Condition as the independent variable and Tweet Sentiment as the dependent variable.

$$Stock\ Market\ Changes = \beta_0 + \beta_1 \times Tweet\ Sentiment + \epsilon$$

The regression results are summarized below:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:       positive_prob   R-squared (uncentered):             0.090
Model:                         OLS   Adj. R-squared (uncentered):        0.087
Method:              Least Squares   F-statistic:                        29.60
Date:             Sun, 22 Sep 2024   Prob (F-statistic):              1.11e-07
Time:                     21:07:40   Log-Likelihood:                   -185.04
No. Observations:              300   AIC:                                372.1
Df Residuals:                  299   BIC:                                375.8
Df Model:                        1
Covariance Type:         nonrobust
=============================================================================
                      coef   std err          t      P>|t|      [0.025    0.975]
-----------------------------------------------------------------------------
Market Sentiment    0.1686     0.031      5.440      0.000      0.108     0.230
=============================================================================
Omnibus:                    17.329   Durbin-Watson:               1.061
Prob(Omnibus):               0.000   Jarque-Bera (JB):           16.145
Skew:                        0.508   Prob(JB):                 0.000312
Kurtosis:                    2.489   Cond. No.                     1.00
======================================================================
```

The p-value for the Market Sentiment (Market Conditon) coefficient is less than 0.001, indicating that the relationship between Market Condition and Tweet Sentiment is strongly statistically significant. This result supports our hypothesis that public sentiment, as captured by NBC sentiment analysis, is a significant predictor of Nvidia's stock price fluctuations.

While our primary focus is the linear regression analysis, it is also important to evaluate the performance of the NBC. The classifier's performance is reflected in the accurate classification of tweet sentiments, which in turn impacts the regression results. Given that our regression analysis shows strongly significant results, it indirectly validates the effectiveness of the NBC in capturing relevant sentiment from Twitter data.

## #7 Conclusion

The analysis demonstrates that the Naive Bayes Classifier, when applied to Twitter data, can effectively capture public sentiment and that this sentiment is a significant predictor of Nvidia Corporation's stock price movements. This finding aligns with the expectations of behavioral finance theory, which posits that emotions and moods in society influence investment decisions and financial markets.

Overall, this project demonstrates the effectiveness of the NBC in sentiment analysis and its relevance to financial market predictions.

### Citations:

1. Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using Twitter sentiment analysis. *Advances in Social Networks Analysis and Mining*, 119–123.
2. Miwa, K. Market Closures and Cross-sectional Stock Returns. *Asia-Pac Financ Markets* 27, 1–33 (2020).
3. Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford, 1(2009), p.12*.