GROUP WORK PROJECT # 2 MScFE 600: FINANCIAL DATA
GROUP NUMBER: 8000

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Franck Delma Deba Wandji	France	debafranck@gmail.com	
Erastus Gitau Nyoike	Kenya	actuarynyoike@gmail.com	
Youstina Amgad Zaki	Egypt	youstina.amjad@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).

Team member 1	**Franck Delma Deba Wandji**
Team member 2	**Youstina Amgad Zaki**
Team member 3	**Erastus Gitau Nyoike**

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.
Note: You may be required to provide proof of your outreach to non-contributing members upon request.

N/A

GROUP WORK PROJECT # 1 MScFE 600: FINANCIAL DATA
Group Number: 8000

Part 1. Assessing Models with Alternative Data

Q1: The paper uses the following types of data:

1. Historical Stock Prices: These are daily closing prices of stocks.
2. Trading Volume: The number of shares traded in a day.
3. Economic Indicators: Data like interest rates, inflation, and GDP growth rates.
4. News Sentiment Data: Financial news items are analyzed to determine market sentiment using news sentiment data
5. Fundamental Data: Financial metrics of companies, such as earnings and revenue.

Technical indicators are derived from this data by:

- Moving Averages (MA): Calculating the average stock price over a specific period to identify trends.
- Relative Strength Index (RSI): Measuring the speed and change of price movements to find overbought or oversold conditions.
- Bollinger Bands: Using a moving average with upper and lower bands based on standard deviations to indicate volatility.
- MACD (Moving Average Convergence Divergence): Showing the relationship between two moving averages of a stock's price.

Technical indicators are important because they help:

- Identify Trends: Show the direction of the market (up, down, or sideways).
- Best Timing Trades: Determine the best times to buy or sell stocks.
- Manage Risks: Highlight potential overbought or oversold conditions.
- Understand Market Sentiment: Reflect the psychology of market participants.

Q2: IVV is an Exchange-Traded Fund (ETF) that tracks the S&P 500 Index's performance. The 500 biggest publicly traded firms in the US, representing a range of industries such as technology, healthcare, finance, consumer discretionary, and more, make up this index.

Price History: Over the last ten years, the IVV ETF has shown notable growth, which is indicative of the general performance of the US equities market. Strong returns are indicated by the fund's price history, particularly during times of economic expansion. For instance, IVV grew steadily each year between 2010 and 2020, coinciding with the bull market at the time.

Key Statistics:

- **Expense Ratio:** 0.03% - IVV is known for its low expense ratio, making it one of the most cost-efficient ETFs available.
- **Dividend Yield:** Approximately 1.5% - The fund provides regular income to investors through dividends, which are distributed quarterly.

- **Market Capitalization:** Over $200 billion - IVV is one of the largest ETFs by market cap, reflecting its popularity among investors.
- **Performance:** Historically, IVV has delivered strong returns, closely tracking the S&P 500 Index. The fund's long-term performance highlights its effectiveness in providing diversified exposure to the U.S. stock market.

Why classification instead of regression: The authors chose a classification problem instead of a regression problem for several reasons:

- Simplify decisions: Classification provides discrete outcomes that are consistent with actual trading decisions (e.g., will the stock market go up or down), making it easier for investors to act on predictions.
- Handle nonlinear relationships: Classification can more effectively capture and model nonlinear relationships present in financial data than regression.

Alternative classification definitions:

- Threshold-based classification: Instead of a binary classification, the authors could define multiple categories based on specific thresholds of price changes, such as “significant increase,” “moderate increase,” “no change,” “moderate decrease,” and “significant decrease.”
- Volatility-based classification: Another approach might be to classify dates based on market volatility. For example, "high volatility days," "medium volatility days," and "low volatility days," using measures like the VIX index or intraday price range.

Q2: Section 2: Data

Data Collection: Describe sources of historical price, volume, economic indicators, and sentiment data.

Data Processing: Explain steps such as data cleaning, standardization, and missing value handling.

Technical Indicators: Detail how each technical indicator is derived from the raw data.

Section 3: Methodology:

3.1 LASSO Regression

3.2 Architecture of Neural Networks

3.3 Training and Testing Procedures

3.4 Optimization Techniques

Distinguish between descriptive statistics and models:

Descriptive Statistics: Discuss summary statistics such as mean, median, standard deviation, and Pearson correlation coefficient. Models: Focus on algorithms and techniques used for forecasting, such as LASSO regression and neural networks.

The optimization processes for the technical indicators used in the document are grid search and cross-validation to optimize the parameters of the technical indicators to improve their predictive power. This is important because the optimized indicators provide more accurate input features to the neural network model, thereby improving the overall performance and reliability.

Q4: A feature is a single measurable property or characteristic that is used as an input to a predictive model. Examples include technical indicators, volume, and sentiment readings.

To distinguish between features and methods:

- Features: Input variables used for prediction (e.g. RSI, MA, Bollinger Bands).
- Methods: Techniques used to process data and extract features (e.g. sentiment analysis, technical analysis).
- Models: Algorithms used to make predictions based on features (e.g. neural networks, LASSO regression).

Categories of learned features:

- Technical indicators: Derived from historical price and volume data.
- Sentiment scores: Extracted from news and social media analysis.
- Basic indicators: Financial indicators such as profit and sales.

The authors used techniques such as grid search and cross-validation to optimize the parameters of technical indicators to improve their predictive capabilities. The optimized indicators provided more accurate input features, improving the overall performance and reliability of the neural network model.

Q5: By splitting the dataset into training and validation sets, a technique known as cross-validation is used to evaluate a model's performance and make sure it performs well when applied to new data.

Using k-1 folds for training and the remaining fold for validation, K-fold cross-validation divides the dataset into k equal-sized folds. This procedure is repeated k times to guarantee robustness.

By dividing the size of the intersection by the size of the union removed from 1, the Jaccard distance calculates how different two sets are.

Examine two alternative distance measures in comparison to the Jaccard distance.

- along a multidimensional space, **the Euclidean distanc**e is the distance along a straight line between two points
- **The Manhattan Distanc**e calculates the separation between two locations by adding up the absolute coordinate differences between them.

The authors define an optimal solution as the set of model parameters that yield the best predictive accuracy, as determined by evaluation metrics like accuracy, precision, recall, and the area under the ROC curve (AUC).

Step 1: Financial Problem

1. The financial problem the authors aim to solve with their model is assisting investors in market timing and risk management within emerging markets. Their model leverages predictive indicators to help investors determine optimal entry and exit points for ETFs, thereby maximizing returns and minimizing losses. By understanding market dynamics through this approach, investors can swiftly adjust their positions in response to changing market conditions, reducing exposure to volatility and market fluctuations.

2. Key Differences & Model Significance:
    - Cyclic Behavior – Emerging markets rely on price cycles, while developed markets use qualitative factors like sentiment and fundamentals.
    - Volume Indicators – Emerging markets prioritize AOBV (quantitative), while developed markets use PVR (classification-based).
    - Feature Selection – Emerging markets depend on technical indicators, whereas developed markets incorporate broader economic trends.

Models must be tailored to each market's behavior. Emerging market models should focus on cyclic trends, while developed market models require qualitative insights. The study's feature selection approach may improve various ETF models and apply to different asset classes.

Step2 : Application

1. Main Takeaways of the Results:
    - Optimized Feature Selection Improves Prediction,
    - Selecting the right technical indicators significantly enhances stock market movement predictions for emerging markets ETFs,
    - Emerging Markets Depend on Cyclic Trends,
    - Unlike developed markets, emerging markets rely more on price cycles and quantitative indicators for accurate forecasting,

- Volume-Based Indicators Differ by Market Type,
- AOBV is more effective for emerging markets, while PVR works better for developed markets, highlighting key structural differences.

2. Given the Venn diagram of figure 6 in the article, the features that seem to be useful for the study are:
    - AOBV, Archer's on balance volume;
    - BBP, Bollinger band percent;
    - BOP, balance of power;
    - CTI, correlation trend indicator;
    - DEC, decreasing;
    - EBSW, even better SineWave;
    - INC, increasing;
    - STOCHRSI, stochastic relative strength index, and
    - WILLR, Williams % R.

Step 3: Replication

- We chose to work with iShares MSCI Brazil ETF (EWZ).
- We implemented algorithm 2 described in section 2 of the article in a python object class named StockMLPClassifier, designed to predict stock price movements using a Multi-Layer Perceptron (MLP) neural network. The process begins with feature selection, where six statistical measures—low variance, Lasso coefficients, random forest feature importance, PCA variances, mean absolute deviation (MAD), and dispersion ratio—are employed to identify salient features. Features appearing in at least a specified number of subsets (determined by the first quartile of each measure) are retained. An MLP classifier is configured with parameters provided by the article. The model is evaluated using stratified k-fold cross-validation to ensure robustness, and its performance is further assessed on a held-out test set. The implementation emphasizes systematic feature selection, model configuration, and rigorous evaluation to enhance predictive accuracy in stock market forecasting.

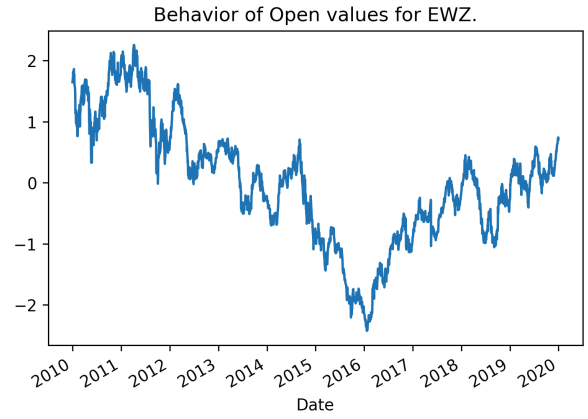

Figure 1: Behavior of Open values for EWZ.

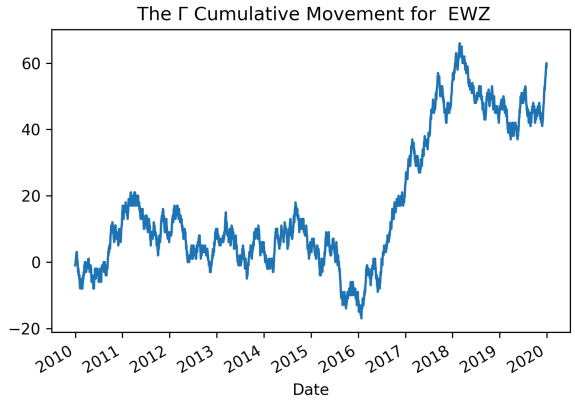

Figure 2: The Gamma Cumulative Movement for EWZ

Figure 1 illustrates the behavior of Open values for EWZ iShares, while figure 2 presents its cumulative movement.

To obtain model predictions, we replicated all technical indicators available in Pandas TA and retained only those without missing values from December 31, 2009, onward. This process resulted in a total of 213 features. We then implemented an MLP model across different n_subsets. The table below summarizes our results:

n_subsets	Features	Accuracy (%)	Training time (s)	Epoch
0	213	77.18	47.99	84
1	169	76.73	44.40	88
2	87	77.34	17*8	115
3	44	75.54	24.09	205
4	16	75.59	40.69	617
5	2	57.14	0.04	46

The analysis reveals a trade-off between the number of selected features and model performance. As the number of subsets (n_subsets) increases, the number of selected features decreases significantly, dropping by 92% (from 213 to 16). Despite this reduction, the median accuracy after cross-validation only decreases by approximately 2.5% (Figure 3), demonstrating

the robustness of the feature selection process. Notably, the results align closely with those reported in the referenced article, even though the chi-squared statistical measure was excluded from the feature selection process.

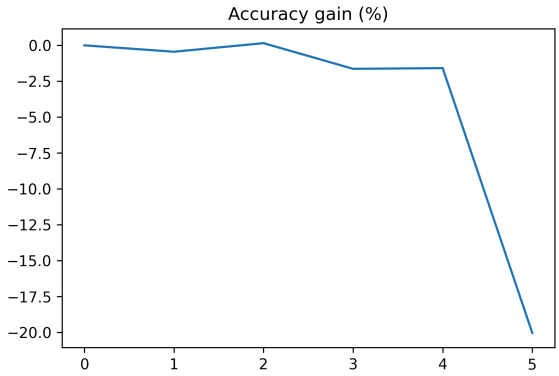

Figure 3: Percentage gain by the selection of features.

Among the selected features, **AOBV, BBP, BOP, and STOCHk** are common with those highlighted in the referenced paper on the **EWZ ETF**. The remaining features, which include **CCI (Commodity Channel Index), AROOND (Aroon Down Indicator), and the Z-scores,** seems to be also useful for predictions. These differences observed here may be explained by the fact that we did not consider chi-square statistical measure for features selection.

These features belong to five categories as illustrated in the following table:

Selected Features (n_subest = 5)	Category
AOBV_LR_2, AOBV_SR_2	Volume
CG_10, STOCHk_14_3_3, BOP, CCI_14_0.015	Momentum
AROOND_14, AROONU_14, DMP_14, AMATe_LR_8_21_2	Trend
open_Z_30_1, close_Z_30_1, high_Z_30_1, low_Z_30_1	Statistics
BBP_5_2.0	Volatility

Part 2. Evaluating One Particular Type of Alternative Data

User Guide: Social Media Data in Finance and Business

Today, the commonly used type of alternative data in finance and business is social media data. Social media data provide real-time insights about customers' opinions and engagements, and public sentiment. These are key information that can be used to make decisions in finance and business. This guide offers a comprehensive overview of social media data as an alternative data source by exploring the following:

1) Sources of Data

Social media data are collected from different online platforms where users engage with each other, share opinions, generate content, and express their opinions. These online platforms include the following. First, Facebook, data includes engagement on financial pages and investors' sentiment in the form of public posts, reactions, and comments. Second, X(Twitter), data includes stock or crypto mentions, new reactions, and financial discussion in the form of tweets, retweets, user interaction, and hashtags. Third, LinkedIn, data includes companies' news, job postings, and professional discussions. Fourth, Reddit, data includes investors' trends and sentiment analysis. Fifth, Stocktwits, data includes stock discussions, which is the core business of the platform. Sixth, YouTube, data includes perspectives on trends and opinions provided by video content, comments, and engagement metrics.

2) Types of Data

Social media data includes both structured and unstructured data formats, which are categorized into the following types. First, textual data, this type of data includes users' tweets, posts, reactions, comments, and reviews. It's mostly used in sentiment analysis, trends, and topics. Second, sentiment data, this data type includes user's opinions, attitudes, and emotions

expressed in text form. It's used in sentiment analysis to measure public perception about a product or market trends. Third, metadata, this data type includes additional information like user profiles, timestamps, geolocation, and engagement metrics (likes, retweets, and shares). This provides context and improves the analysis of social media data. Fourth, multimedia data, this data type includes content like videos, images, and audio shared by social media users that can drive public perceptions. Fifth, network data, this type of data includes information about the relationship between users for example friends, interactions connections, and followers.

3) Quality of Data

To produce accurate and reliable analysis results that can be used in making informed financial decisions, researchers or analysts ought to ensure that social media data is of high quality. The quality of social media data is influenced by the following factors. First, data accuracy. Social media data can be distorted by misinformation, the use of bots, and fake social media accounts. So, filtering out unreliable sources is key to ensuring data validity and accuracy. Second, data relevance. To ensure the relevance of data, social media data needs to align with the business or financial question that's to be addressed. For instance, tweets related to a specific stock are more relevant for investment analysis than general tweets. Third, bias. In some cases, social media users might not be representative of the whole population, this can lead to sampling biases. Therefore, cross-validating information with other relevant data sources can minimize bias issues. Fourth, the completeness of data. Social media data need to be a representative dataset with few or no missing values. Data with too many missing values and incomplete data can adversely affect analysis and thereby the results. Fifth, timeliness. Where market conditions are prone to change rapidly, especially in the stock market and forex market, social media data need to be real-time or near-real-time data to ensure they are valuable for financial application.

Researchers and analysts can ensure high-quality social data by doing the following. First, use of filters to get rid of spam, misinformation, and bot-generated content. Second, using verified APIs to obtain reliable and structured data. Finally, cross-validation of social media data with external sources. Researchers can validate social media findings with reliable external sources to ensure robustness and avoid biases.

4) Ethical Issues

To ensure responsible practices in the use of social media data, researchers ought to shun or mitigate ethical issues that come with its use. First, researchers ought to avoid privacy issues. Most social media users do not consent to the use of their data by third parties. Researchers need to use publicly available data to avoid interference with the privacy of social media users. Second, researchers ought to avoid transparency issues. Researchers ought to communicate clearly and openly to social media users how their data is collected and is to be used. Third, researchers also ought to avoid anonymization issues. This arises when anonymized information is traced back to the owner. To avoid such a problem, researchers ought to use secure anonymization methods to safeguard the identities of social media users. Finally, researchers ought to avoid compliance issues to avoid reputational and legal risks. By adhering to social media's general data protection regulations, researchers can avoid risks that come with non-compliance.

5) Python Code to Import and Structure into Useful Data Structures

In this case, we shall focus on how you can retrieve stock data from X(Twitter) and structure it into useful data.

```
# load required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import tweepy
from textblob import TextBlob
# identify Twitter API credentials
api_key = "replace with your_actual_api_key"
api_secret_key = "replace with your_actual_api_secret_key"
access_token = "replace with your_actual_access_token"
access_token_secret = "replace with your_actual_access_token_secret"

# authenticate with Twitter API
auth = tweepy.OAuth1UserHandler(api_key, api_secret_key, access_token,
access_token_secret)
api = tweepy.API(auth)

# retrieve tweets related to a stock
query = "#Stock Ticker"
tweets = tweepy.Cursor(api.search_tweets, q = query, lang="en",
count=200).items(200)

# structure tweets and  about the a stock into a dataframe
data = []
for tweet in tweets:
    data.append({
        'user': tweet.user.screen_name,
        'location': tweet.user.location,
        'created_at': tweet.created_at,
        'text': tweet.text,
        'retweets': tweet.retweet_count,
        'likes': tweet.favorite_count
    })

stock_data = pd.DataFrame(data)
```

6) Exploratory Data Analysis of Sample Data

To understand the characteristics of the imported data, we need to conduct exploratory data analysis. This will include plot graphs and computation of basic statistics like count, of the key elements of data.

```
# perform sentiment analysis

## create sentiment variable using TextBlob function
stock_data['sentiment'] = stock_data['text'].apply(lambda x:
TextBlob(x).sentiment.polarity)
## categorize sentiments into negative, neutral or positive
stock_data["sentiment category"] = stock_data["sentiment"].apply(lambda x:
"Negative" if x < 0 else "Positive" if x > 0 else "Neutral")
```

```
# display sentiments distribution
print(stock_data["sentiment category"].value_counts())

# plot the sentiments distribution
sns.countplot(x="sentiment category", data=stock_data, palette="coolwarm")
plt.title("Sentiment Distribution of X(Twitter) Data")
plt.xlabel("Sentiment Category")
plt.ylabel("Tweet Count")
plt.show()
```

7) Short Literature Search That Links to Papers Citing Research

Various studies have demonstrated the importance of social media data in finance especially in the investment of stocks and cryptocurrencies. Twitter sentiments can significantly be used to forecast stock price movements, especially for frequently discussed stocks (Sul et al., 23). Bartov et al. found that aggregated opinions from individuals' tweets correlate with firms' future earnings and returns (20). Moreover, other research like Siganos et al., emphasized the relationship between Facebook investor sentiment and stock volatility (1). Finally, the analysis by Kraaijeveld and De Smedt about cryptocurrency discussions on Twitter revealed that there's a strong predictive relationship between price movements and Twitter (X) sentiment(14). All these studies demonstrate the significance of social media data in investment and financial decision-making.

Work Cited

Bartov, Eli, et al. “Can Twitter Help Predict Firm-Level Earnings and Stock Returns?” The Accounting Review, vol. 93, no. 3, 1 July 2017, pp. 25–57, doi:10.2308/accr-51865.

Kraaijeveld, Olivier, and Johannes De Smedt. “The Predictive Power of Public Twitter Sentiment for Forecasting Cryptocurrency Prices.” Journal of International Financial Markets, Institutions and Money, vol. 65, Mar. 2020, p. 101188, doi:10.1016/j.intfin.2020.101188.

Siganos, Antonios, et al. “Facebook’s Daily Sentiment and International Stock Markets.” Journal of Economic Behavior & Organization, vol. 107, Nov. 2014, pp. 730–743, doi:10.1016/j.jebo.2014.06.004.

Sul, Hong Kee, et al. “Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns.” Decision Sciences, vol. 48, no. 3, 23 June 2016, pp. 454–488, doi:10.1111/deci.12229.

Sagaceta Mejia, C., et al. (2022). An Intelligent Approach for Predicting Stock Market Movements in Emerging Markets Using Optimized Technical Indicators and Neural Networks. *De Gruyter*. Retrieved from https://www.degruyter.com/document/doi/10.1515/econ-2022-0073/html