**CCT College Dublin**

|  |  |
| --- | --- |
| **Module Title:** | Advanced-Data Analysis |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | David McQuaid |
| **Student Full Name:** | Haider Riaz Ghuman |
| **Student Number:** | 2022564 |
| **Assessment Due Date:** | 08/05/2024 |
| **Date of Submission:** | 05/05/2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my work and that all material from third parties has been appropriately referenced. I further confirm that this work has yet to be submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Introduction**

This project's objective is to perform sentiment analysis together with time-series forecasting to gain insights and predict future trends based on already provided data. Sentiment analysis is critical in Natural Language Processing (NLP) which involves determining the sentiment expressed as text data. Time-series forecasting, unlike NLP, focuses more on predicting future values based on patterns and trends observed from the historical data that has already been obtained.

In the current generation, large data to be used for the analysis can easily be obtained from social media, sentiment analysis plays a very huge role in understanding customer feedback as well as public opinion. We can analyze sentiments presented as textual data in organizations to rate if there is customer satisfaction, take note of emerging trends, and make decisions based on the results of the analysis to improve the products and services offered. Sentiment analysis is applied in various fields e.g. finance, marketing, monitoring, social media, and managing customer relationships.

In addition, time-series forecasting is crucial for businesses to predict future trends which help them make decisions. By analyzing pre-existing data, organizations can point out patterns, seasons, and trends, allowing them to forecast future trends and plan accordingly. Time-series forecasting is largely used in finance, sales and forecasting demand, inventory management as well as allocation of resources.

By combining both time-series forecasting and sentiment analysis the project is aimed at providing crucial insights into sentiment trends, enabling accurate decision-making, and helping organizations in laying their strategies based on the prediction.

**Sentiment Extraction Technique.**  
To achieve sentiment extraction, we have used SentimentIntensityAnalyzer from the NLTK library. This is a technique that utilizes a lexicon-based approach by assigning sentiment scores to a text based on the words in the text. The words can be positive negative or neutral. The SentimentIntensityAnalyzer also calculates the compound sentiment score which represents the whole sentiment expressed in a given text. I have chosen this technique because of its efficiency, simplicity, and excellent performance in capturing sentiment in various fields. (Liu, 2012)

For the justification of the choice of the SentimentIntensityAnalyzer, I conducted experiments comparing SentimentIntensityAnalyzer with other sentiment analysis models, including rule-based approaches and machine learning models. It consistently demonstrated efficient performance in terms of speed and accuracy across multiple datasets that I used. In addition, its lexicon-based approach allows quality and clear interpretability of the sentiment scores, which is important in understanding the sentiment patterns in a dataset.

**Time-Series Forecasting Method**

In this project, I have explored some popular time-series forecasting methods: ARIMA (Autoregressive Integrated Moving Average) and SARIMA (Seasonal ARIMA).

ARIMA is widely used for univariate time-series forecasting and combines differencing (I), autoregressive (AR) as well as moving average (MA) techniques to capture dependencies and patterns. Its effectiveness is seen when data with stationary behavior is used and becomes more valuable in the absence of seasonality.

SARIMA extends ARIMA for consideration of seasonal patterns. It is used appropriately when the dataset being used has both trend and time. By incorporating time, SARIMA can provide accurate forecasts on data with recurring trends over fixed time intervals.

Both ARIMA and SARIMA have been used in this project based on their suitability for capturing complexities when performing sentiment scores and time series. ARIMA is well known for autocorrelation and trend, while on the other hand, SARIMA handles seasonality and trend. (Masters,1995)

In this analysis, I have compared the performance of SARIMA and ARIMA models using multiple parameters and validation techniques. I have also used Evaluation metrics such as RMSE, MSE, and forecasting accuracy. Both models have demonstrated stiff performance in capturing sentiment trends as well as predicting future scores.

In conclusion, ARIMA and SARIMA have proved to be effective methods to be used for forecasting and sentiment analysis. They have provided accurate and reliable forecasts; this forecast may be used to improve service providence.

**Final Analysis and Forecasting**  
For the final analysis, I’ve chosen the SentimentIntensityAnalyzer for sentiment extraction and SARIMA and ARIMA for the time-series forecasting. This has allowed me to exploit the simplicity and clarity of the sentiment extraction technique and employ forecasting models that best capture the complex trends in the sentiment score time series. (Livera et al,2011)

Based on the pre-existing sentiment scores, I was able to generate forecasts for one week, one month, and 3 months into the future. The forecasted sentiment scores provide awareness of the expected sentiment trends over these periods. The two models (ARIMA and SARIMA), I trained them trained on the already existing sentiment scores and evaluated their accuracy and reliability using appropriate validation techniques.

The forecasts for 1 week, 1 month, and 3 months are as follows:

* ARIMA Forecasted Sentiment Score at 1 week: -0.1006969689140514
* SARIMA Forecasted Sentiment Score at 1 week: -0.11431992496658845
* ARIMA Forecasted Sentiment Score at 1 month: -0.10069696891405194
* SARIMA Forecasted Sentiment Score at 1 month: -0.1055080325602983
* ARIMA Forecasted Sentiment Score at 3 months: -0.10069696891405194
* SARIMA Forecasted Sentiment Score at 3 months: -0.10550881910868427

I have assessed the reliability and accuracy of the forecasts by comparing their results with the actual sentiment scores. The evaluation metrics that I employed included root mean squared error, mean squared error and finally forecasting the forecasting accuracy. These metrics that I used provided evidence of the forecasting models their performance and their ability to capture the sentiment trends and patterns in the data provided.

**Dynamic and Interactive Dashboard**  
To present the results of the analysis, I have developed a dynamic and interactive dashboard that adheres to Tufts principles. The dashboard is very user-friendly and has a user-friendly interface with visualizations of the results also has very easy-to-use controls. The design format behind the dashboard focuses more on accessibility, usability, and eye appeal.

The dashboard is composed of various interactive features, which include interactive graphs and charts that allow users to explore sentiment trends with time. The users are allowed to select specific time ranges, zoom in and out, also view very detailed sentiment scores for specific dates or intervals according to how they choose. The visuals are designed in such a way that they provide a clear understanding of the sentiment patterns and trends thus making i making it easy for users to extract awareness from the data.

**Conclusion**In conclusion, in this project, I made use of the SentimentIntensityAnalyzer for sentiment extraction and employed ARIMA and SARIMA models for time-series forecasting. The choice of these techniques was justified based on the performance of each, interpretability, and ability to capture the sentiment trends in the data that had been provided.

The forecasts for 1 week, 1 month, and 3 months provide valuable awareness of the expected sentiment patterns in the coming future. The evaluation of the forecasting models indicates their accuracy, reliability, and clarity in predicting sentiment scores over different time horizons as time goes by.

The dynamic and interactive dashboard that I developed is meant to enhance the user experience as well as facilitate the exploration of sentiment trends. It also provides the users with an inborn interface for analyzing sentiment scores and gaining awareness of the data.

For future improvements, I would recommend the exploration of additional sentiment extraction techniques, such as deep learning models, to further enhance sentiment analysis accuracy and flexibility altogether. Additionally, combining more advanced time-series forecasting methods, such as LSTM (Long Short-Term Memory) or Prophet, may potentially improve the forecasting accuracy and also capture more complex patterns in the sentiment score time series.

Overall, this project successfully made use of sentiment analysis and time-series forecasting techniques and provided valuable awareness of sentiment trends and predicted future sentiment scores. The dynamic and interactive dashboard enhances the user experience and facilitates the interpretation of the results.

**References**

* Liu, B. (2012). Sentiment Analysis. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.
* Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2008). Time Series Analysis: Forecasting and Control. John Wiley & Sons.
* De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting Time Series with Complex Seasonal Patterns Using Exponential Smoothing. Journal of the American Statistical Association, 106(496), 1513-1527.
* Wei, W. W. (2006). Time Series Analysis: Univariate and Multivariate Methods (2nd ed.). Pearson Education.
* Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts.
* Brockwell, P. J., & Davis, R. A. (2016). Introduction to Time Series and Forecasting (3rd ed.). Springer.
* Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). Forecasting: Methods and Applications (3rd ed.). John Wiley & Sons.
* Chatfield, C. (2000). Time-Series Forecasting (2nd ed.). CRC Press.