information

Article

Leveraging Social Media Data for Enhanced Forecasting of International Student Arrivals in Australia

Ali Abdul Karim, Eric Pardede * and Scott Mann

Department of Computer Science and Information Technology, La Trobe University, Melbourne 3083, Australia; a.abdulkarim@latrobe.edu.au (A.A.K.); s.mann@latrobe.edu.au (S.M.)
* Correspondence: e.pardede@latrobe.edu.au

Abstract: This study examines the extent to which incorporating social media data enhances the predictive accuracy of models forecasting international students' arrivals. Private social media data collected from a public university, along with collected web traffic data and Google Trend data, were used in the forecasting models. Initially, a correlation analysis was conducted, revealing a strong relationship between the institution's international student enrolment and the social media activity, as well as with the overall number of international students arriving in Australia. Building on these insights, features were derived from the collected data for use in the development of the forecasting models. Two XGBoost models were developed: one excluding social media's features and one including them. The model incorporating social media data outperformed the one without it. Furthermore, a feature selection process was applied, resulting in even more accurate forecasts. These findings suggest that integrating social media data can significantly enhance the accuracy of forecasting models for international student arrivals.

Keywords: international students; social media; time series forecasting; web data; web traffic data; Google Trend; students forecasting

Citation: Abdul Karim, A.; Pardede, E.; Mann, S. Leveraging Social Media Data for Enhanced Forecasting of International Student Arrivals in Australia. *Information* **2024**, *15*, 823. https://doi.org/10.3390/info15120823

Academic Editor: Arkaitz Zubiaga

Received: 30 September 2024
Revised: 5 December 2024
Accepted: 18 December 2024
Published: 23 December 2024

1. Introduction

The increasing prevalence of web data, particularly from social media platforms, has opened new avenues for enhancing forecasting models across various domains. Social media data, along with other digital sources like web traffic and Google Trends, captures user interactions with the digital world offering insights into use behaviour and trends. These digital footprints can uncover patterns and predictors not apparent from traditional data sources presenting researchers and practitioners with opportunities to improve forecasting models. In the context of education, analyzing web data allows researchers to uncover patterns that signal shift in students interest and intentions, enabling more dynamic and responsive predictive models.

Our research explores these digital data sources to improve the accuracy of forecasting the number of international students arriving in Australia, which is essential for both government agencies and educational institutions. Accurate predictions enable governments to effectively plan resource allocation, policy development, and infrastructure investments. For universities, reliable forecasts assist in preparing for student intake, managing staffing and facility requirements, and ensuring the overall quality of education and student services. Specifically, we leverage web traffic data to capture the online behaviours of prospective students, Google Trends data to gauge the popularity of search queries related to studying in Australia, and social media data from La Trobe University, a public university in Australia, to assess engagement and interest levels.

This study is significant for several reasons. Initially, it is the first paper to forecast the number of international students arriving in Australia using such a comprehensive approach. Second, it is the first to investigate the contribution of social media data from an

institution in improving the accuracy of these forecasts. By incorporating social media data from an institution, we aim to demonstrate the added value of this non-traditional data source. Additionally, this paper outlines a detailed data extraction process that includes critical steps such as investigating the domain, identifying data sources, and processing the data. By presenting this process, we provide transparency into our methodology and highlight the systematic approach taken to ensure the reliability and relevance of the data used in our forecasting model.

A key aspect of our research is the examination of the correlation between social media data from the institution, its enrolment numbers, and the overall number of international students arriving on student visas. Preliminary analysis shows a strong correlation between social media engagement and enrolment numbers at the institution, which in turn correlates with the number of student visas issued. This finding supports the hypothesis that social media data can serve as a valuable predictor for forecasting international student arrivals at a broader level.

In our methodology, we develop models using the full set of available data and compare them with models that exclude social media data. This comparison helps us evaluating the influence of social media data on model accuracy, underscoring its significance in forecasting. By incorporating a variety of data sources, we aim to enhance forecasting techniques for international student arrivals and offer practical insights for stakeholders in the education sector. Notably, this paper is the first to use university-specific data to forecast a macro-level indicator, making it easier for universities to apply their data to predict both institutional enrolment and the arrival of international students to Australia. Our study emphasizes the critical role of collecting social media data for applications in sectors like education, which is a key industry in Australia.

The following sections will cover the Literature Review, Data Collection and Analysis, Methodology, Experiments and Results, and Conclusions. This comprehensive approach underscores the potential of digital data in enhancing predictive models and supporting strategic decision-making in higher education and government planning.

2. Literature Review

In recent years, the use of alternative data sources, including those generated from web usage and social media, has gained significant attention for enhancing forecasting models across various domains. These data sources have been proven to improve prediction accuracy and offer valuable insights into future trends. This section reviews existing studies that have leveraged social media and web data in forecasting, highlighting their contributions along with the contribution of our study in the forecasting domain.

2.1. Social Media Data in Forecasting

Several studies have explored the predictive potential of social media across different applications. Oliveira et al. [1] investigated the impact of Twitter data on predicting stock market returns, volatility, trading volume, and sentiment indices, demonstrating the effectiveness of social media data in forecasting financial indicators. Zhang [2] showcased the feasibility and effectiveness of leveraging social media data for forecasting economic indicators.

In the education sector, Nguyen et al. [3] emphasized the role of social media engagement in driving enrolment intentions, while Brown et al. [4] explored the potential of social media data in forecasting sports outcomes, using tweets to predict soccer match results. Other studies by scholars such as Wu et al. [5] and Giri et al. [6], have further demonstrated the relevance of social media data in predicting stock market volatility and garment sales, respectively. These studies collectively highlight the growing importance of social media as a rich resource for predictive analytics across various fields.

Building on this foundation, our paper makes a unique contribution by being the first to investigate the impact of using social media data from a university on forecasting international student arrivals to Australia. This study extends the application of social media data beyond typical domains like finance or retail, offering new insights into its

potential within the context of international education. By integrating social media data with other web data, our research aims to enhance the predictive accuracy of international student arrivals that might benefit educational institutions on their planning.

2.2. Web Data in Forecasting

Beyond social media, other web data sources have been extensively utilized for forecasting macro-level indicators. Höpken et al. [7] and Havránek and Zeynalov [8] focused on using Google Trends data to enhance demand forecasting in the tourism sector, illustrating how traveller's online search behaviour can be a valuable predictor of tourist arrivals. These studies reveal the potential of web data, such as Google Trends, to improve forecasting accuracy in the tourism industry. Additionally, Karim et al. [9] showcased that incorporating web traffic data for Australian websites improves the forecasting of short term visitors arriving to Australia.

The data sources have been investigated in forecasting applications in other domains. Smith [10] predicted UK unemployment using Google search data, showcasing the utility of online search behaviour in nowcasting economic trends. Similarly, Zhang et al. [11] employed a Long Short-Term Memory (LSTM) model incorporating internet search index data to forecast hotel accommodation demand, highlighting the significance of machine learning techniques in conjunction with web data for accurate time series forecasting. Additionally, Karim et al. [12] evaluated the performance of machine learning model and SARIMAX when incorporating Google Trend data in forecasting models of two Australian indicators. Wang et al. [13] introduced a fuzzy time series model for forecasting student enrolment, emphasizing the importance of tailored approaches in educational settings.

Our research contributes to this growing body of work by integrating web data, specifically social media data from La Trobe University, into a forecasting model for international student arrivals. While previous studies have highlighted the value of web data in tourism and economic forecasting, our paper is the first to apply this approach to the domain of international education. By leveraging both social media and web data, our study provides a novel framework that could significantly enhance the predictive accuracy of models forecasting international student arrivals, thereby offering valuable implications for decision-making in higher education.

3. Data Collection and Analysis

To accurately forecast the number of international students arriving in Australia on student visas, we collected data from a range of digital and traditional sources. These sources include web traffic data, Google Trends data, the institution's enrolment data, the number of international students arriving in Australia, and social media data from the institution. Our systematic data extraction process involved investigating the domain, identifying relevant data sources, and processing the collected data to ensure its reliability, as illustrated in Figure 1.

Following this overview, Figure 2 provides a more detailed look at the specific steps taken to identify data sources, emphasizing the importance of each type of data in informing our forecasting model. By breaking down the identification of Google Trends, web traffic, and social media data, we highlight how these elements contribute to a comprehensive understanding of the factors influencing potential students' decisions. Given the availability of the social media data starting in 2018, we aligned all data sources from 2018 onwards to ensure consistency in our analysis.

3.1. Web Traffic Data

Web traffic data were sourced from various Australian websites relevant to international students. These data captures the online behaviour of potential students, providing insights into their search patterns and interests related to studying in Australia. The data include numbers representing the actual visitors to each web page. Although web traffic data date back to 2012, we used data from January 2018 onwards for the forecasting exercise

due to the lack of historical social media data before this date. Additionally, we focused on traffic from international internet users by collecting overall traffic to the websites and subtracting traffic from Australia. This approach allows us to monitor trends among international students interested in studying in Australia, excluding local students and international students already residing in Australia. As shown in Figure 2, the identification of relevant websites is a crucial step in our data extraction process, emphasizing the importance of pinpointing sources that accurately reflect the interests and behaviours of prospective international students; a list of websites from which we collected the traffic data are presented in Table 1. Furthermore, Figure 3 illustrates how trends in the number visitors of an Australian university website (La Trobe) are similar to those for the number of international students entering Australia.

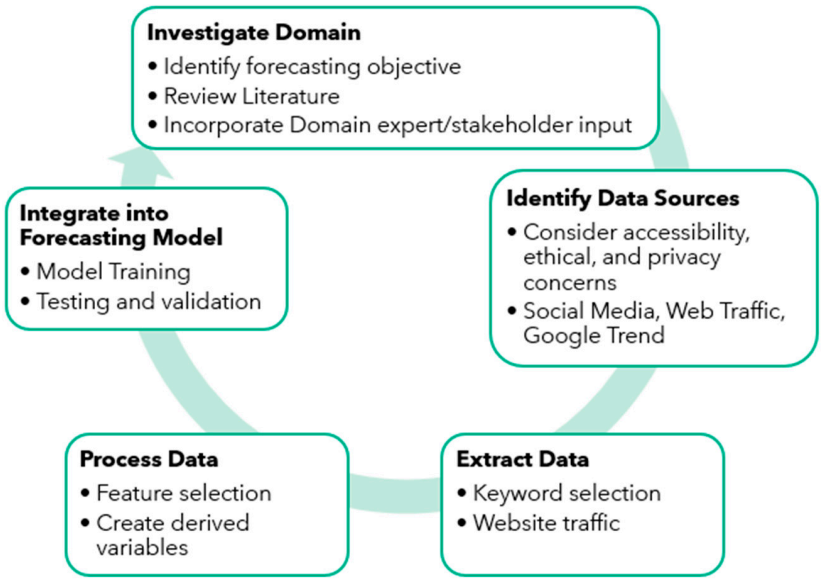

Figure 1. This diagram illustrates the systematic data extraction process employed in this study to forecast the number of international students arriving in Australia, detailing the steps from domain investigation to the integration of various data sources.

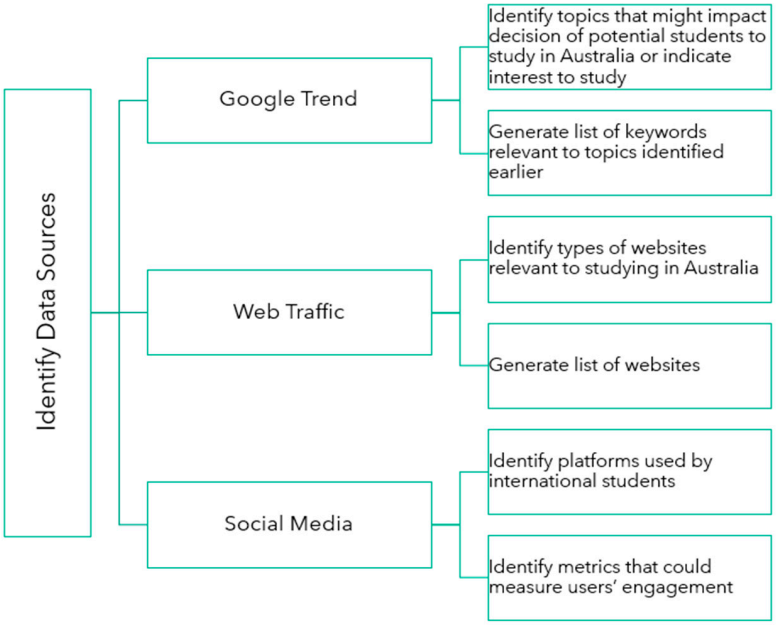

Figure 2. This diagram outlines the process of identifying data sources for forecasting international students arriving in Australia.

Table 1. List of the source of the web traffic data.

Website (accessed 17 December 2024)	Comment
unimelb.edu.au	University of Melbourne website
latrobe.edu.au	La Trobe University website
sydney.edu.au	Sydney University website
anu.edu.au	Australian National University website
monash.edu.au	Monash University
homeaffairs.gov.au	Australian Immigration website
studyassist.gov.au	Information on financial assistance for students in Australia
studyinaustralia.gov.au	Information on studying in Australia

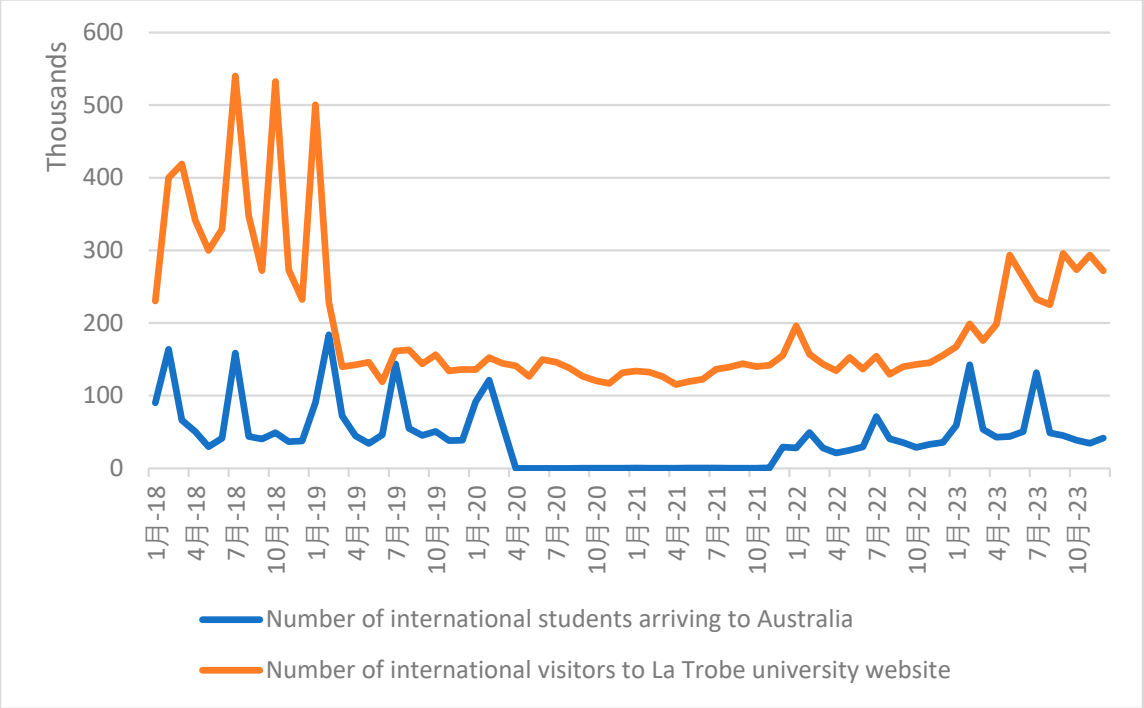

Figure 3. Comparison of a university website visitor and students entering Australia (" 月" is the interpretation of month).

3.2. Google Trends Data

Google Trends data were collected for search queries related to studying in Australia. These data measures the popularity of specific search terms over time, reflecting the level of interest in studying in Australia. While Google Trends data span back to 2004, we used data from 2018 onwards for forecasting international student arrivals.

Google Trends data represent the volume of searches made by users on specific keywords over time, providing insights into public interest and behaviour regarding various topics. The data are typically presented as normalized values ranging from 0 to 100, reflecting the relative search interest. For this study, data were collected for keywords related to studying in Australia, such as "university fees in Australia", "Australian student visa", "study in Australia", and "Australian universities." These data are crucial for forecasting as they reflect the level of interest and intent of prospective students globally. As shown in Figure 2, the identification of relevant keywords is an essential step in our data extraction process, emphasizing how understanding search patterns related to factors that impact students' decisions can enhance the accuracy of our forecasts. Some of these steps were discussed by Moogan [14]. A complete list of search keywords used in this study is provided in Table 2.

Table 2. List of search keywords used in extraction of Google Trends data.

Search Keywords	Comment
university fees in Australia	Indicates interest in the cost of education, which is a crucial factor in a student's decision-making process.
fee structure in Australian universities	
scholarships for international students in Australia	Shows interest in financial support options, a significant factor in attracting international students.
Australian student visa	Highlights concerns about the prerequisites and process for obtaining a student visa, which could affect the number of applicants.
visa requirements for Australia	
study in Australia	A broad search that shows general interest in Australia as a study destination, potentially correlating with initial interest to study in Australia.
study abroad in Australia	
why study in Australia	
English proficiency requirements in Australia	Indicates the need for information on language proficiency standards, essential for admission.
study English in Australia	Targets students looking to improve their English language skills, often a prerequisite for other academic programs.
Australia university ranking	Indicates that students are considering the reputation and quality of Australian universities, which can influence their decision.
best universities in Australia	Reflects students' search for available options of educational institutions, likely impacting their final choice.
top universities in Australia	
Australian universities	
accommodation options in Australia	Indicates exploration of various living arrangements, critical for planning a stay in Australia.
renting in Australia	
Australia student accommodation	
online courses in Australia	Reflects interest in remote learning options, which have become more popular and could impact traditional enrolment numbers.
Australian National University	Focused search that suggests interest in a specific university and academic offering which could indicate higher likelihood of actual enrolment.
University of Melbourne	
La Trobe university	
University of New South Wales	
Monash university	
Melbourne university campus	
University of Sydney	
Melbourne university programs	
Admission requirements for the University of Sydney	Increased interest in applying to a specific university.

For example, an increase in searches for "visa requirements for Australia" indicates that a significant number of prospective students are considering applying for a student visa. Figure 4 shows peaks in Google search for visa requirements associated with peaks of international students arriving to Australia. Additionally, keywords like "study in Australia" and "scholarships for international students in Australia" highlight growing interest in studying in Australia and the financial considerations involved. Tracking these searches allows us to gauge the popularity and demand for Australian education over time.

Analyzing trends in these keywords can identify periods of heightened interest and predict potential surges in student visa applications. For instance, if there is a spike in searches for "Australian student visa" or "admission requirements for the University of Sydney", it suggests more students are inquiring about the application process and are likely to apply soon. This digital footprint offers a real-time, dynamic measure of interest

that complements traditional data sources, providing a more comprehensive and timely understanding of the factors influencing international student enrolment.

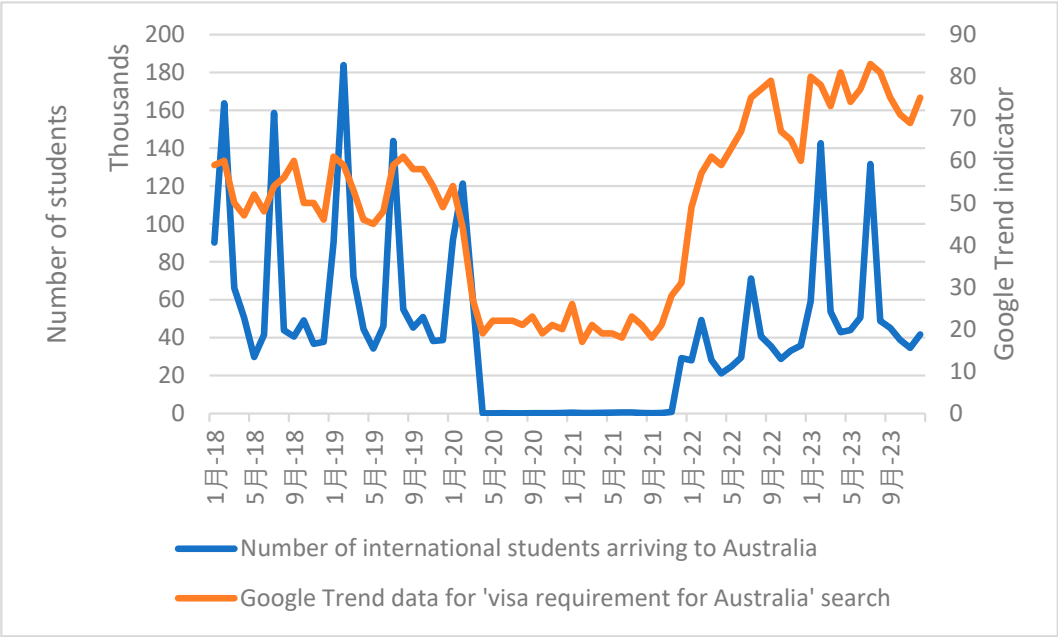

Figure 4. Google Trend data for visa requirement search along with the number of international students ("月" is the interpretation of month).

Understanding these search trends helps educational institutions and policymakers anticipate changes in demand, plan resource allocation, and implement strategies to attract and accommodate future international students. By leveraging Google Trends data, we can make more informed decisions to enhance the international student experience and ensure that Australia remains a top destination for higher education.

3.3. Target Variables

The target variable in this study is the number of international students arriving in Australia on student visas, sourced from the Australian Bureau of Statistics (ABS) with historical data extending beyond the year 2000. This comprehensive dataset provides a robust foundation for analyzing trends and patterns in international student arrivals over an extended period. Additionally, we utilize yearly enrolment data from La Trobe University, dating back to 2004 (Figure 5). These data, obtained directly from the university, play a critical role in our research.

The inclusion of the institution's enrolment data allows for an assessment of their correlation with the broader national data on international student arrivals. If a strong correlation is found between the institution's yearly enrolment numbers and the overall number of international students arriving in Australia, it validates the use of its social media data in the forecasting model. This would mean that social media engagement and interactions at a university level can serve as a reliable indicator of national trends, enhancing the model's accuracy and predictive power. By integrating these data sources, the study aims to leverage the digital footprint of prospective students to improve forecasting outcomes.

3.4. Social Media Data from the Institution

Social media data were gathered from the institution's official social media channels. These data includes metrics such as engagement rates, follower counts, and other interactions indicating the level of interest and engagement with the institution's media content. Social media data are available from 2018 onwards, providing a recent and relevant perspective on student engagement.

Figure 5. La Trobe University enrolment numbers along with the number of international students arriving to Australia.

Since the social media data from the institution is only available from 2018, we have used data from 2018 onwards across all sources to maintain consistency in our analysis. This approach allows us to leverage the most recent and relevant data while ensuring all variables are aligned in terms of the period covered.

By combining these diverse data sources, we aim to build a robust forecasting model that incorporates both traditional and non-traditional indicators of international student interest and arrivals. The integration of digital data sources, such as web traffic and social media data, provides a richer and more nuanced understanding of the factors driving international student mobility.

Social media has become an integral part of modern communication and marketing strategies for educational institutions. With platforms like LinkedIn (LI), Facebook (FB), Twitter (TW), and Instagram (INSTA) playing significant roles in reaching prospective students, it is essential to explore how social media data can aid in forecasting the number of international students enrolling in higher education. This section delves into the types of social media data available and discusses its potential impact on predicting student enrolment.

3.4.1. Type of Social Media Data

The social media data available for analysis include the following key metrics:

- Engagement: The total number of interactions (likes, shares, comments) that the content receives. This metric reflects the overall activity and interest generated by the posts.
- Engagement Percentage: This metric represents the percentage of users who interacted with the content out of the total reach.
- Reach: The total number of unique users who have seen the content. Reach helps in understanding the potential audience size exposed to the institution's messaging.
- Clicks: The number of times users clicked on the links shared in the posts. This metric indicates the level of interest and action taken by the audience towards the content.
- Touchpoint (tp): The number of various interactions or engagements with social media content, such as likes, comments, shares, clicks, or any other form of interaction.

- Enquiries: The number of direct inquiries generated from social media posts. This is a critical metric as it directly correlates with potential leads and interest in the institution's offerings.
- Additionally, Stories—temporary posts that disappear after 24 h—and Reels—short, engaging videos—are features specific to platforms like Instagram and Facebook, designed to enhance user interaction and engagement.

These metrics are collected across multiple social media platforms providing a comprehensive view of the institution's online presence and engagement.

3.4.2. Potential Impact on Forecasting

Social media data can offer valuable insights into the behaviour and preferences of prospective students. By analyzing patterns and trends in the engagement, reach, clicks, and inquiries, educational institutions can gauge the effectiveness of their marketing strategies and adjust their efforts to attract more international students. Several studies demonstrated the significant impact of social media across different domains highlighting its potential in identifying trends and enhance forecasting Gupta and Pandey [15], Li et al. [16], Hewapathirana [17]. Here are some ways in which social media data might enhance forecasting in the education domain:

1. Trend Identification: Social media metrics can help identify trends in student interest and engagement over time. For instance, a surge in engagement or inquiries during specific periods might indicate heightened interest due to marketing campaigns or favourable conditions.
2. Audience Segmentation: Different social media platforms attract different demographics. Analyzing data from multiple platforms allows institutions to segment their audience and tailor their messaging to different groups of prospective students.
3. Predictive Modelling: Incorporating social media data into predictive models is the primary focus of our paper, as it can improve the accuracy of forecasting student enrolment. Metrics like engagement and inquiries can serve as leading indicators of future enrolment numbers, allowing institutions to make more informed decisions.

3.4.3. Preliminary Analysis

Before conducting comprehensive experiments and analyzing results, we performed a preliminary correlation analysis on the provided social media data against the institution's enrolment data and the latter against the yearly number of international students arriving in Australia.

To investigate the relationships between various data sources and the number of international students arriving in Australia on student visas, we performed several correlation analyses.

First, we focused on the institution's social media data. We aggregated the monthly social media data to a yearly level using both the sum and maximum values, then performed a Pearson correlation analysis between the aggregated yearly social media data and the yearly enrolment numbers at the institution from 2018 to 2023. The results indicated a strong positive correlation, with a Pearson correlation coefficient greater than 0.6 (Table 3). This high correlation suggests that increased activity on the institution's social media platforms is closely linked with higher enrolment numbers, highlighting the potential of social media engagement as a predictor of student interest and enrolment decisions.

Next, we analyzed the correlation between the yearly enrolment at the institution and the number of international student visas issued. The visa data, originally provided on a monthly basis by the ABS, were aggregated to a yearly level for consistency. Our analysis, covering data from 2004 onwards, revealed a high positive correlation between the institution's yearly enrolment numbers and the yearly number of international student visas, with a correlation coefficient of 0.83. This strong correlation underscores the link between university enrolment trends and international student arrivals, suggesting that enrolment data can serve as reliable indicators of broader trends in international student mobility.

Table 3. Top 10 correlated social media features with yearly La Trobe enrolment: (a) max represents the maximum value of the feature within a given year; (b) sum represents the total number of interactions that corresponds to the listed feature.

Feature	Correlation Value
sum_reach_STORIES	0.796334
sum_clicks_STORIES	0.78523
max_engage_pct_Reels	0.778304
sum_engage_pct_FB	0.740002
max_enquiries_STORIES	0.718319
sum_reach_LI	0.676359
max_enquiries_Total	0.670123
sum_enquiries_STORIES	0.646459
max_engag_INSTA	0.639729
max_tp_INSTA	0.639618

These preliminary findings suggest that social media data holds promise for enhancing the forecasting of international student enrolment. However, further experiments and detailed analysis are required to validate these insights and determine the exact impact of social media.

4. Approach

In this study, our goal is to forecast the number of international students arriving in Australia on student visas by leveraging diverse data sources such as web traffic, Google Trends, social media metrics from La Trobe University, and historical student arrival data. To achieve this, we followed a comprehensive methodology that included feature engineering, feature selection, model selection, and data splitting. Each step is crucial for building an effective and accurate predictive model.

4.1. Feature Engineering

Feature engineering is a pivotal step in our methodology, as it transforms raw data into meaningful features that capture the underlying patterns essential for improving the performance of machine learning models. In this study, we utilized monthly data from four distinct sources: web traffic data, Google Trend data, social media data from La Trobe University, and the target variable, which is the number of international students arriving in Australia.

To capture temporal dependencies and trends, we created lagged features and rolling averages for each variable. The temporal nature of the data helps in understanding how past values influence future outcomes. Lagged features were created by shifting the data by 1 to 12 months, allowing the model to learn from the delayed effects of these variables on the target variable. This helps in capturing short-term and medium-term dependencies. Rolling averages were computed over windows of 3 to 12 months, these features smooth out short-term fluctuations and highlight longer-term trends, which can improve the stability and reliability of the model.

By engineering these features, we enriched the dataset, making it more suitable for predictive modelling by providing the model with a deeper context and a more comprehensive understanding of temporal trends.

First formula to create a lagged features for a column X where k in the number of months,

$$Lag_k(X) = X_{t-k},$$

$$Rolling\ Average_w(X) = \frac{1}{w}\sum_{i=0}^{w-1} X_{t-i}.$$

The second formula is to create a rolling average for column X where w is the window size (3 to 12).

4.2. Feature Selection

Given the extensive set of features generated through feature engineering, performing feature selection is essential to avoid overfitting and enhance model interpretability.

The importance and purpose of feature selection in our study are multifaceted. First, it helps in reducing overfitting by selecting only the most relevant features, preventing the model from learning noise in the data and thereby improving its generalizability to unseen data. Second, feature selection improves interpretability, as a smaller set of well-chosen features makes the model easier to understand and interpret. Lastly, by focusing on features that are highly correlated with the target variable and less correlated with each other; we mitigate multicollinearity by reducing redundancy and enhancing overall model performance.

We employed a systematic feature selection approach, starting with a grid search to determine the optimal combination of parameters. This search aimed to balance the number of features selected and the threshold for acceptable inter-feature correlation. After evaluating multiple configurations, we settled on m = 30 (limiting the number of features selected to approximately one-thirtieth of the total) and n = 0.5 (the maximum allowable average correlation between newly added features and already-selected ones).

Our method involved calculating the Pearson correlation coefficient between each feature and the target variable. The feature most strongly correlated with the target variable was selected first. Subsequent features were chosen based on their correlation with the target variable and their average correlation with the already-selected features, ensuring that retained features were both informative and minimally redundant. Table 4 presents the number of features used in model training before and after feature selection.

Table 4. Feature count before and after applying feature selection.

	Without Feature Selection		**With Feature Selection**	
Source Type	**All Data**	**Excluding Social Media**	**All Data**	**Excluding Social Media**
Google Trend	594	594	22	15
Social Media	572	0	7	0
Web Traffic	220	220	14	9
Target Features	22	22	3	2
Total	1408	836	46	26

This process ensured that the retained features were both informative and non-redundant, resulting in a refined set of features for model building.

4.3. Model Selection

For the forecasting task, we selected the XGBoost algorithm introduced by Chen and Guestrin [18] due to its robustness and superior performance in various machine learning tasks. XGBoost, or Extreme Gradient Boosting, is a powerful ensemble learning technique that builds multiple decision trees sequentially to improve prediction accuracy.

XGBoost offers several significant advantages that align well with the nature of the data used in this study. Firstly, it is optimized for computational efficiency, making it suitable for handling the large-scale dataset derived from the collected web data quickly and effectively. These datasets often contain complex temporal patterns in which XGBoost excels at managing and handling them delivering state-of-the-art results and high accuracy in forecasts of international student numbers. Another crucial benefit of using XGBoost is its ability to provide insights into feature importance, aiding in model interpretation and helping us understand which time series features contribute most to the predictions. Additionally, XGBoost can inherently manage different data types without requiring transformations such as scaling or log transformation, simplifying the preprocessing steps which in beneficial given the various formats and structures of the web data collected.

We evaluated the model performance using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). RMSE is a common regression metric that measures the square root of the average squared difference between predicted and actual values. MAE measures the average magnitude of the errors in a set of predictions without considering their direction. Lower RMSE and MAE values indicate better model performance, helping us gauge the effectiveness of our forecasting models.

4.4. Training and Testing Split

To validate the performance of our forecasting model, it was crucial to divide the dataset into distinct training and testing sets. The dataset covers the period from 2018 to 2023, incorporating social media data starting from 2018.

The division served several key purposes in our methodology:

Firstly, starting model training from January 2019 ensured that all lagged features were adequately populated. This approach provided the model with comprehensive and meaningful input data, enhancing its ability to capture temporal dependencies and patterns.

Secondly, the initial training phase spanned from January 2019 to June 2023, while the testing phase extended from July 2023 to December 2023. This timeline selection allowed us to validate the model's performance on recent data. By focusing on more recent periods, we aimed to ensure that the model's predictions remained relevant and reflective of current trends in international student enrolment.

Lastly, splitting the data into training and testing sets enabled us to assess the model's generalization capability. Evaluating its performance on unseen data provided a realistic measure of its forecasting accuracy, indicating how well it could predict future trends beyond the training period.

Following this structured methodology, our goal was to develop a robust forecasting model for predicting the influx of international students arriving in Australia on student visas. By leveraging a combination of traditional and digital data sources, including social media metrics, our approach not only aimed to enhance predictive accuracy but also to offer valuable insights into the underlying factors influencing international student mobility and enrolment patterns.

5. Experiments and Results

After creating modelling features from all available data, we built two sets of XGB regression models: one using all features except social media data (Google Trends, Web Traffic and Target variable features), and another using all features including social media data. This was performed to assess the impact of incorporating social media data on model performance. The models were trained using data from January 2019 until June 2023 and tested in the remaining six months of 2023. We built an additional set of models after applying feature selection using the number features presented in Table 4 to train the models.

5.1. Initial Model Performance

The initial model without social media data performed slightly better than the model with social media data. This could be attributed to the large number of features fed into the model and the instability during the COVID-19 period. To further validate these results, we built six additional models, each time shifting the training data back by one month. This allowed for more robust results and consistency checks. The results are summarized in Table 5.

The average RMSE for models including all data were 33,288, compared to 39,240 for models excluding social media data. This indicates that, on average, models incorporating social media data performed slightly better initially (approximately 15% better). However, both types of models showed a significant decrease in performance immediately after the pandemic, which is understandable due to the reopening of Australia's borders and the associated economic and social changes.

Table 5. RMSE and MSE for forecasting models without feature selection.

Test Period	All Data		No Social Media	
	RMSE	**MAE**	**RMSE**	**MAE**
July–Dec 2023	38,179	33,450	39,698	33,895
June–Nov 2023	33,285	26,759	38,544	32,103
May–Oct 2023	23,896	18,610	31,165	24,195
Apr–Sep 2023	23,794	19,039	33,685	22,385
Mar–Aug 2023	22,869	17,271	35,389	21,119
Feb–Jul 2023	46,845	28,830	50,802	32,872
Jan–Jun 2023	44,152	25,666	45,394	26,997
Average	33,288	24,232	39,240	27,652

5.2. Refined Model Performance After Feature Selection

To further investigate the impact of social media data and the large number of features on the previously built models, we performed feature selection. This involved removing features highly correlated with the target variable and other predictors (greater than 0.5). This refined approach demonstrated the superiority of the model with social media data, yielding one average a 12% better RSME with 18,457 for models with all data and 20,869 for models excluding social media data. The results are summarized in Table 6.

Table 6. RMSE and MAE for forecasting models with feature selection.

Test Period	All Data		No Social Media	
	RMSE	**MAE**	**RMSE**	**MAE**
July–Dec 2023	7786	6937	8394	7520
June–Nov 2023	8433	7521	9462	8617
May–Oct 2023	8642	7610	8525	7527
Apr–Sep 2023	3996	3471	15,355	13,693
Mar–Aug 2023	13,464	12,567	15,502	14,039
Feb–Jul 2023	42,941	28,756	42,718	28,822
Jan–Jun 2023	43,936	27,801	46,125	30,624
Average	18,457	13,523	20,869	15,835

The RMSE for model trained on all data sources with feature selection and tested on July–Dec 2023 data are 7786 compared to 38,179 for model trained on all data without feature selection and tested over the same period. This substantial improvement of 79% highlights the importance of feature selection in building effective models and ensuring the contributions of social media data are not overshadowed by other highly predictive features.

Overall, incorporating social media data significantly improved the forecasting model's performance for predicting the number of international students. The initial set of models showed some inconsistencies due to the sheer number of features and the volatility during the COVID-19 period. However, once feature selection was applied, the models with social media data consistently outperformed those without.

These findings underscore the importance of incorporating social media data in forecasting models, as well as the necessity of feature selection to refine and enhance model performance. As presented in Figure 6, the model built on all data sources feature selection predicts the number of international students better than the model built without feature selection. Additionally, the results demonstrate the impact of external factors such as visa requirements and political decisions, which can significantly influence the arrival of international students (Table 7). Future models will benefit from more data and consideration of these external factors to provide more accurate and robust forecasts.

In conclusion, social media data provides valuable contributions to forecasting models for the number of international students. Feature selection is essential to optimize model

performance and interpret the underlying patterns effectively. Collecting more data over time and accounting for external factors will further improve the accuracy and reliability of these models.

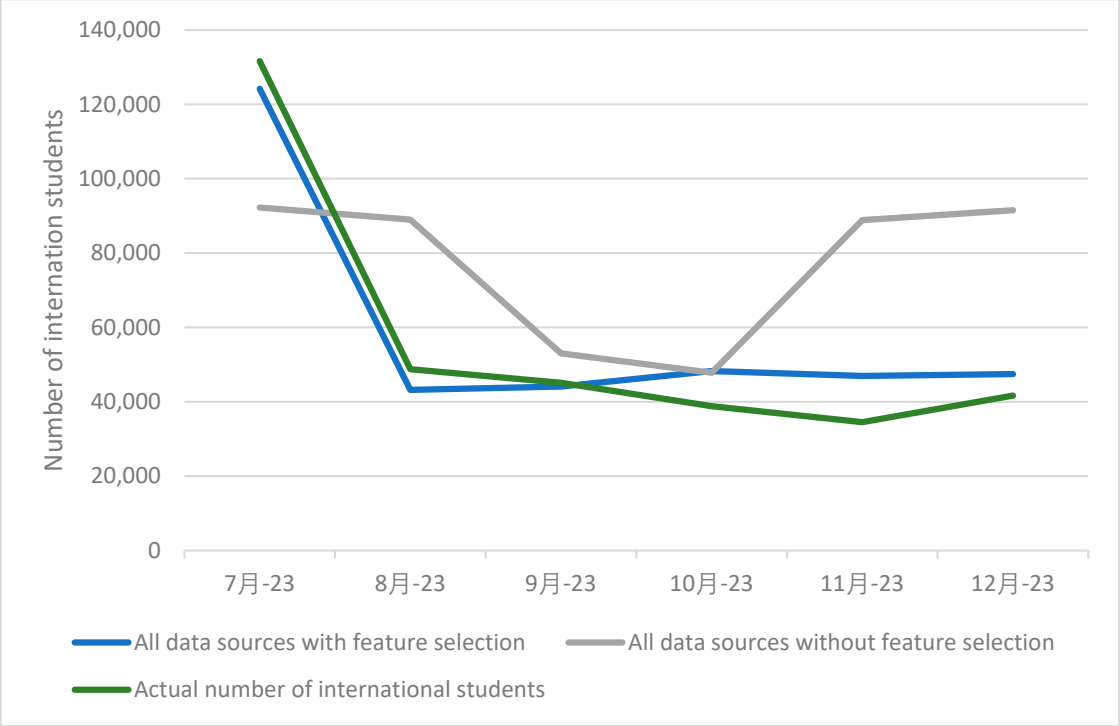

Figure 6. Actual and predicted value of international students arriving to Australia using all data sources with and without feature selection ("月" is the interpretation of month).

Table 7. Top 10 features contribution to models tested on July 2023 to Dec 2023 data. (sm: Social Media, gt: Google Trend, wt: Web Traffic, y: Target features, l: Lag Feature, avg: Average Feature).

Without Feature Selection		With Feature Selection	
All Data	**Excluding Social Media**	**All Data**	**Excluding Social Media**
sm_enquiries_TW_avg_9	wt_studyassist.gov.au_avg_7	gt_study_english_in_australia_l6	wt_latrobe.edu.au_avg_7
gt_study_abroad_in_australia_l2	gt_top_universities_in_australia_l6	y_l12	y_l12
gt_top_universities_in_australia_l6	gt_university_fees_in_australia_l3	wt_latrobe.edu.au_avg_7	gt_top_universities_in_australia_l5
gt_university_of_new_south_wales_l2	gt_university_of_new_south_wales_l2	gt_visa_requirements_for_australia_l1	gt_visa_requirements_for_australia_l1
sm_enquiries_TW_l2	wt_studyinaustralia.gov.au_l1	wt_latrobe.edu.au_avg_6	gt_study_abroad_in_australia_l1
wt_studyinaustralia.gov.au_l1	y_l12	gt_study_abroad_in_australia_l1	gt_australian_national_university_l6
wt_studyassist.gov.au_avg_7	gt_visa_requirements_for_australia_l1	y_l1	gt_study_abroad_in_australia_l6
gt_visa_requirements_for_australia_l1	wt_anu.edu.au_l1	sm_reach_LI_avg_7	wt_latrobe.edu.au_l12
sm_enquiries_FB_l5	wt_anu.edu.au_l9	gt_study_abroad_in_australia_l6	y_l1
sm_reach_INSTA_avg_6	gt_best_universities_in_australia_l12	gt_study_in_australia_l5	wt_latrobe.edu.au_avg_4

6. Conclusions

This study aimed to enhance the forecasting of international student arrivals in Australia by incorporating various data sources, including web traffic data, Google Trend data, social media data from a public university in Australia, and traditional target variable features. Our methodology encompassed extensive feature engineering and selection, leveraging the power of the XGBoost algorithm to build robust predictive models.

Our experiments highlighted the nuanced impact of including social media data in the forecasting models. Initially, models without social media data outperformed those with social media data. However, after applying rigorous feature selection to mitigate the influence of multicollinearity and highly correlated predictors, models incorporating social media data demonstrated superior performance. This underscores the critical role of feature selection in ensuring the effectiveness of predictive models.

The refined models, which accounted for temporal dependencies and trends through lagged features and rolling averages, consistently showed lower MSE values when social media data were included. This finding indicates that social media data can significantly enhance the accuracy of forecasting models by providing additional context and insights into student behaviour and engagement trends.

Our results also revealed the importance of considering external factors, such as economic stability and policy changes, which can dramatically influence international student arrivals. The COVID-19 pandemic, in particular, introduced substantial volatility, affecting model performance during and immediately after this period.

In conclusion, our study demonstrates the substantial benefits of integrating social media data into forecasting models for international student arrivals. Feature selection emerged as a crucial step to optimize model performance and ensure that social media data's contributions are effectively captured. This research highlights the potential of combining traditional and modern data sources to improve predictive analytics in higher education and beyond.

Looking ahead, there are several avenues for future research to build on this study. One direction involves testing the proposed approach on data from other countries to assess its generalizability and adaptability across different contexts. Another avenue focuses on incorporating additional data, such as geopolitical events, immigration policies, and other external factors, which could provide deeper insights and improve the robustness of the forecasting models. Finally, exploring alternative feature selection techniques, which could further enhance model performance, particularly when dealing with the large and diverse datasets collected from web sources. Pursuing these directions will help refine the predictive analytics methodologies and extend their applicability to a broader range of scenarios.

Author Contributions: Conceptualization, A.A.K., E.P. and S.M.; methodology, A.A.K., E.P. and S.M.; software, A.A.K.; validation, A.A.K.; formal analysis, A.A.K.; investigation, A.A.K.; resources, A.A.K.; data curation, A.A.K.; writing—original draft preparation, A.A.K.; writing—review and editing, A.A.K., E.P. and S.M.; supervision, E.P. and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to some data being private institutional data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Oliveira, N.; Cortez, P.; Areal, N. The Impact of Microblogging Data for Stock Market Prediction: Using Twitter to Predict Returns, Volatility, Trading Volume and Survey Sentiment Indices. *Expert Syst. Appl.* **2017**, *73*, 125–144. [CrossRef]
2. Zhang, W.K. Predicting Sales With Social Media Data. *Adv. Mater. Res.* **2014**, *926–930*, 3870–3873. [CrossRef]
3. Nguyen, L.V.T.; Lu, V.N.; Conduit, J.; Tran, T.T.N.; Scholz, B. Driving Enrolment Intention Through Social Media Engagement: A Study of Vietnamese Prospective Students. *High. Educ. Res. Dev.* **2020**, *40*, 1040–1055. [CrossRef]
4. Brown, A.; Rambaccussing, D.; Reade, J.J.; Rossi, G. Forecasting With Social Media: Evidence From Tweets on Soccer Matches. *Econ. Inq.* **2017**, *56*, 1748–1763. [CrossRef]
5. Wu, X.; Wang, X.; Ma, S.; Ye, Q. The Influence of Social Media on Stock Volatility. *Front. Eng. Manag.* **2017**, *4*, 201. [CrossRef]
6. Giri, C.; Thomassey, S.; Zeng, X. Exploitation of Social Network Data for Forecasting Garment Sales. *Int. J. Comput. Intell. Syst.* **2019**, *12*, 1423. [CrossRef]
7. Höpken, W.; Eberle, T.; Fuchs, M.; Lexhagen, M. Google Trends Data for Analysing Tourists' Online Search Behaviour and Improving Demand Forecasting: The Case of Åre, Sweden. *Inf. Technol. Tour.* **2018**, *21*, 45–62. [CrossRef]
8. Havránek, T.; Zeynalov, A. Forecasting Tourist Arrivals: Google Trends Meets Mixed-Frequency Data. *Tour. Econ.* **2019**, *27*, 129–148. [CrossRef]

9. Karim, A.A.; Pardede, E.; Mann, S. A Feature-Based Model Selection Approach Using Web Traffic for Tourism Data. *Int. J. Web Grid Serv.* **2024**, *20*, 342–359. [CrossRef]
10. Smith, P.A. Google's MIDAS Touch: Predicting UK Unemployment With Internet Search Data. *J. Forecast.* **2016**, *35*, 263–284. [CrossRef]
11. Zhang, B.; Pu, Y.; Wang, Y.; Li, J. Forecasting Hotel Accommodation Demand Based on LSTM Model Incorporating Internet Search Index. *Sustainability* **2019**, *11*, 4708. [CrossRef]
12. Karim, A.A.; Pardede, E.; Mann, S. A Model Selection Approach for Time Series Forecasting: Incorporating Google Trends Data in Australian Macro Indicators. *Entropy* **2023**, *25*, 1144. [CrossRef] [PubMed]
13. Wang, H.; Wang, H.; Guo, J.; Feng, H. A Fuzzy Time Series Forecasting Model Based on Yearly Difference of the Student Enrollment Number. In Proceedings of the 2nd International Conference on Soft Computing in Information Communication Technology, Taipei, China, 31 May–1 June 2014. [CrossRef]
14. Moogan, Y.J. An investigation into international postgraduate students' decision-making process. *J. Furth. High. Educ.* **2020**, *44*, 83–99. [CrossRef]
15. Gupta, K.D.; Pandey, R. Emerging Design Trends in Social Media and Its Impact on Business Efficiency and Growth in India. *Shodhkosh J. Vis. Perform. Arts* **2023**, *4*, 1–7. [CrossRef]
16. Li, Y.; Lin, Z.; Xiao, S. Using social media big data for tourist demand forecasting: A new machine learning analytical approach. *J. Digit. Econ.* **2022**, *1*, 32–43. [CrossRef]
17. Hewapathirana, I.U. Advancing Tourism Demand Forecasting in Sri Lanka: Evaluating the Performance of Machine Learning Models and the Impact of social Media Data Integration. *J. Tour. Futures* **2023**. *ahead-of-print*. [CrossRef]
18. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.