National College of Ireland

Ireland tourist demand forecasting using social media bigdata with a new machine learning analytical approach

MSc Research Project
Data Analytics

Shylesh Veeraraghavan Govindarajulu
Student ID: x21219249

School of Computing
National College of Ireland

Supervisor: Dr. Syed Muslim Jameel

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Shylesh Veeraraghavan Govindarajulu
Student ID:	x21219249
Programme:	M.Sc. Data Analytics **Year:** 2022-23
Module:	MSc Research Project
Supervisor:	Dr. Syed Muslim Jameel
Submission Due Date:	15.08.2023
Project Title:	Ireland demand forecasting using social media big data with a new machine analytical approach
Page Count:	23 pages
Word Count:	6628 words

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Shylesh Veeraraghavan Govindarajulu

Date: 14/08/2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Ireland tourist demand forecasting using social media bigdata with a new machine learning analytical approach

Shylesh Veeraraghavan Govindarajulu
x21219249

Abstract

With the growth of social media users and business's adopting it for marketing, this research presents a method to estimate tourist needs using data from Twitter, now called X. Twitter conversations serve as the primary data. From these conversations, main topics get identified and assigned a sentiment score using the LDA model. Using these scores, predictive models get developed to estimate the number of tourists visiting Ireland. The outcome of this study is a new tool for the travel industry, offering insights into tourist behavior and aiding businesses in planning effective marketing campaigns.

1. Introduction

The tourism sector's contribution to a country's GDP is comparatively high compared to other sectors. Ireland, as an English-speaking nation in the European Union with its beauty, has experienced remarkable growth, in its tourism sector in recent times.

Tourism drives a country's economy extensively across the globe. On a yearly basis, a steady stream of tourists travels to Ireland since it is a popular vacation location. It will certainly help the government and the companies to prepare themselves based on the tourist inflow prediction. Traditional prediction methods had previously been employed to estimate the number of tourists. Due to advances in technology and excessive use of social media data, there is a scope to use social media data extensively and make the prediction more accurate.

Social networking platforms have lately developed as common places in which individuals may express their opinions, experiences, and beliefs. Particularly Twitter stands out as a real-time data treasure trove, providing insights into the prevailing attitudes, propensities, and even behaviors of the general population. The success of tourism marketing efforts can be significantly improved by using this data from Twitter to assess the sentiments among prospective tourists.

1.1. Key goals

This study is a combination of machine learning with the people's opinions and views on Twitter to generate a fresh formula and also to determine the number of travelers whom will pack their luggage for Ireland. The traditional method examines charts and graphs from the past. Yet, the tourism sector continues to change rapidly, thus these outdated charts could miss the most recent developments. Consider having an enchanted sphere that can pick out trends in the flood of tweets and updates, refining your forecasts and preparing you for when the world changes.

1.2. Study goals

The main focus of this research is to bring forth a new strategy to forecast the tourist arrival numbers in Ireland and also to get insights based on the emotional evaluation of the conversations pulled from Twitter (Yulei Li, 2022). To reach this goal, a few specific objectives find a place:

Drawing out pertinent topics from Twitter chats about Ireland to better comprehend what potential tourists might be interested in or prefer. Computing the average sentiment score for each topic that comes to light, thereby getting a read on the general sentiment tied to various aspects of Ireland as a place for tourists. Bringing in a Latent Dirichlet Allocation (LDA) model to automatically cull out hidden topics from the unstructured data on Twitter.

Constructing machine learning models, considering the average sentiment scores for each topic and data on tourist arrivals from the past, to forecast tourist numbers in Ireland's future. Comparing the machine learning method (Mehrbakhsh Nilashi, 2017) proposed here with traditional methods used for forecasting, and evaluating how accurate and efficient it is.

1.3. Importance and Pertinence

In the vast tapestry of tourism prognostication methodologies, this treatise emanates as a beacon of avant-gardism, elucidating an ostensibly novel paradigm by which one might prognosticate the ebb and flow of sojourner congregations within the verdant confines of Ireland. Moreover, extrapolating this modus operandi to encompass other illustrious global sojourn hotspots appears to be within the realms of feasibility. By adroitly amalgamating the multifarious granules of insights gleaned from the vast expanse of social media platforms with the nuanced art of sentiment dissection, the scholarly endeavour seeks to bequeath to the academic and industry conclave a cornucopia of groundbreaking elucidations and enhancements.

Better Prediction Accuracy: By looking at what people talk about on social media, we can use advanced technology to better predict travel trends. To better plan and utilize their resources to suit the needs of travelers, the travel industry benefits from this.

Real-time Insights: Conventional methods of forecasting tourism trends often depend on data from the past. These methods may not respond quickly to trends and sentiments that change at a fast pace. Using data from Twitter enables real-time tracking and quick responses to shifts in potential tourists' preferences.

Customized Marketing Strategies: When the sentiments and interests of tourists come to light from their Twitter chats, the tourism sector can shape marketing strategies that strike a chord with their target groups.

The techniques used to extract topics and attitudes from Twitter discussions will be thoroughly examined in the following chapters. It is discussed how to build machine learning models to generate predictions, then the results are carefully analyzed. These methods' disadvantages are also discussed and also the other possible add-ons that can enhance the framework are also discussed. The primary focus is to use social media data to the maximum potential to forecast tourist numbers.

1.3. Research question

Predict the number of tourists coming to Ireland by analyzing their sentiments and preferences shared on social media and utilize machine learning models to estimate the forthcoming tourist counts.

2. Related Work

2.1. Exploring the Drivers of tourism demand

2.1.1. Economic factors influencing the tourism demand

This paper (Muhammad shafiullah, 2018) deals with finding the determinant of international tourism demand. He has conducted the panel and time series econometric techniques by using Australia and its state data. The tourist demand forecasting carried out by econometric models include static factors that influence the international tourists coming to the respective places. Models are created by considering a variety of econometric elements, such as global income, transit costs at the state level, the population of people who were born abroad, and currency rates. The author arrives at the conclusion that instead of having a general policy that applies to all states and territories, there should be policymaking for each individual state or territory that takes into account both economic and sociopsychological factors.

Multiple experiments with econometric models were conducted, and this research (Jean Max Tavares, 2016) utilized a pooling least squares estimator combined with an approach of moment system estimators. The main objective was to identify the econometric components using Brazil as the case study. For the purpose of estimating tourist demand, the study highlights the

importance of factors like shared borders and proximity to major cities. This essay (Jean Max Tavares, 2016) convincingly demonstrates the effectiveness of the "Aquarela plan 2020" in accurately predicting the requirements of foreign tourists.

In the intricate mélange of economic determinants modulating global sojourn propensities, the ramifications of currency valuations vis-à-vis tourist demand loom large, as expounded upon in this erudite composition (De Vita, 2013). This disquisition furnishes a plethora of incontrovertible attestations corroborating the phenomenon wherein a nation's fiscal denomination's depreciation inexorably catalyzes an augmentation in the conflux of alien excursionists. Through an assiduous and labyrinthine theoretical dissection, it's postulated that globe-trotters, in their geographic predilections, frequently resort to the foreign exchange metric as a surrogate barometer, eschewing a mere consideration of quotidian expenditures or hospice tariffs. Predicated upon the extrapolations derived from the employed econometric paradigm, the treatise culminates in the affirmation that currency fluctuations wield a preponderant influence upon transnational peripatetics' locational inclinations. Consequently, this variable's palpable pertinence in dictating alien excursionist ingress mandates its paramountcy in the annals of touristic strategization and scrutiny.

2.1.1. Non-Economic factors influencing the tourism demand

Numerous non-economic factors, usually divided into two categories: external influences and social-behavioral factors, might affect the demand for tourism. Terrorism, the destination nation's economic policies, and political stability are some of the most important external issues. Terrorist incidents have a big influence on travelers' decisions when selecting their trip destinations. This paper (Liu, 2017) describes the adverse effects of terrorism activities that could possibly affect a country's tourism to a large extent. A research has been conducted 95 nations data in relation to these incidents. The findings of the study allowed for the division of these nations into two groups depending on the impact of terrorism on tourism: those with short-term effects and those with long-term consequences.

Both the Policymakers and business stakeholders can get important insights from understanding how different nations have responded to similar occurrences and also with the aid of these insights, well-informed strategies and backup plans can be created to lessen the adverse effects. This can help in developing the nation's tourism.

According to the study, wealthy nations experienced a decrease in short-term visitor numbers following a significant terrorist incident. However, it took longer for people to wish to return to poorer nations. This paper (Liu, 2017) shows that if a place is politically calm and safe, more tourists want to go there. How tourists feel and what they like can change based on how safe a place feels. Knowing this helps those in charge make plans to keep tourism safe and good, even when bad things happen.

Old places and things are really important for getting tourists to visit a country. A study (Cho V. , 2010) looked at if old places in China make more people want to visit. The results (Cho

V. , 2010) said that old places do make tourists want to visit. The paper (Cho V. , 2010) said that because every country looks after its old places differently, it's hard to guess how many tourists will visit using only info from old places. To make sure these old spots help get tourists, they need to be looked after in special ways.
For both the government and airport operators, accurate forecasting of visitor arrivals at airports is important. To determine the impact of external factors including shocks, fluctuations, flight services, and economic conditions on passenger arrivals. A research (Wai Hong Kan Tsui, 2016) has been conducted that explored into the data of eight key airports in Australia.

2.2. Methods for predicting tourist demand: An overview

Researchers have tried a wide range of strategies throughout the years to figure out the intricacies involved in forecasting visitor demand. Fascinatingly, a study (Haiyan song, 2019) has revealed a stunning compilation of roughly 211 illustrious research publications, covering the years 1968 to 2018, all carefully focused on the enigmatic field of tourism demand forecasting.

When trying to guess how many tourists will come, there are some favorite ways people like to use. They like using things called time series models, econometric models, machine learning, and another thing we're still learning about called artificial intelligence. These ways are picked the most by people when they want to know about future tourists.

2.2.1 Time series method

A study (Cho V. , 2003) compared time series approaches and artificial neural networks to examine the prediction of tourists arriving in Hong Kong from various nations. Along with artificial intelligence technologies like Elman's neural network model, the research included time series techniques like exponential smoothing and the Univariate ARIMA model. The results showed that neural networks outperformed conventional time series approaches in properly estimating visitor numbers. A notable flaw in the time series technique, though, was that it ignored customer behavioral factors, a critical component in thorough tourism demand forecasting. Therefore, even though neural networks showed more prediction ability, behavioral analysis must still be included for a more complete knowledge of the dynamics of tourism demand.

2.2.2 Econometric method

Since it considers more factors than only information on visitor arrivals, the econometric methodology is preferred to the time series method as a tool for forecasting tourism demand. A study (Yuan-Yuan Liu, 2018) that examined the relationships between visitor arrival figures and other crucial factors like weekends, holidays, weather, and web searches relevant to the target area provided evidence of this tactic. With the aid of big data datasets and vector autoregressive models, the research comprehensively analyses these linkages. The study

found out some big things. The way the study did this can show people who make choices and others in the field how to check and make things even better for tourists in the future.

2.2.3 Machine learning method

Machine learning methods could yield much more accurate results in comparison to the traditional methods discussed above. As discussed above the artificial neural network models have proven results of predicting the tourism demand. Research (Wu, 2019) has been done and it has been found it is not easy to build an ANN model by considering both independent variables and tourism demand. Scholars have used multiple machine learning models and obtained better results in predicting the tourism demand.

2.3. Research niche

This paper talks about what makes people want to travel. Some reasons are about money and some aren't. We look at ways people try to guess why tourists come. Reasons about money can be things like how much stuff costs in other places, how much money people have, and what rules places have about money. Then there are other reasons, like if a place is safe, if it has old stuff to see, and what people say online or where they book hotels. We also talk about ways to guess how many tourists will come. Some ways are old, some are new, and some use computers in a cool way. But a lot of what's been written doesn't think about what travelers really feel. So, this study looks at what people say on the internet to see what they feel about traveling. Another study (Yuan-Yuan Liu, 2018) did this too. They checked what travelers think using what they say online and cool computer stuff.

3. Research Methodology

This section gives a picture of how the domain of tourism is selected and the how the thesis framework is proposed, and the process involved in the entire journey of the research. The figure 1 will give a brief idea how research has passed through various phases starting from the proposal to getting reviews from the guide and how the social media platform twitter is selected, and all other sources of data obtained. The forthcoming parts of this section will give the methods and techniques involved in building and executing this framework. Various papers has been referred to build this framework as mentioned in the previous section.

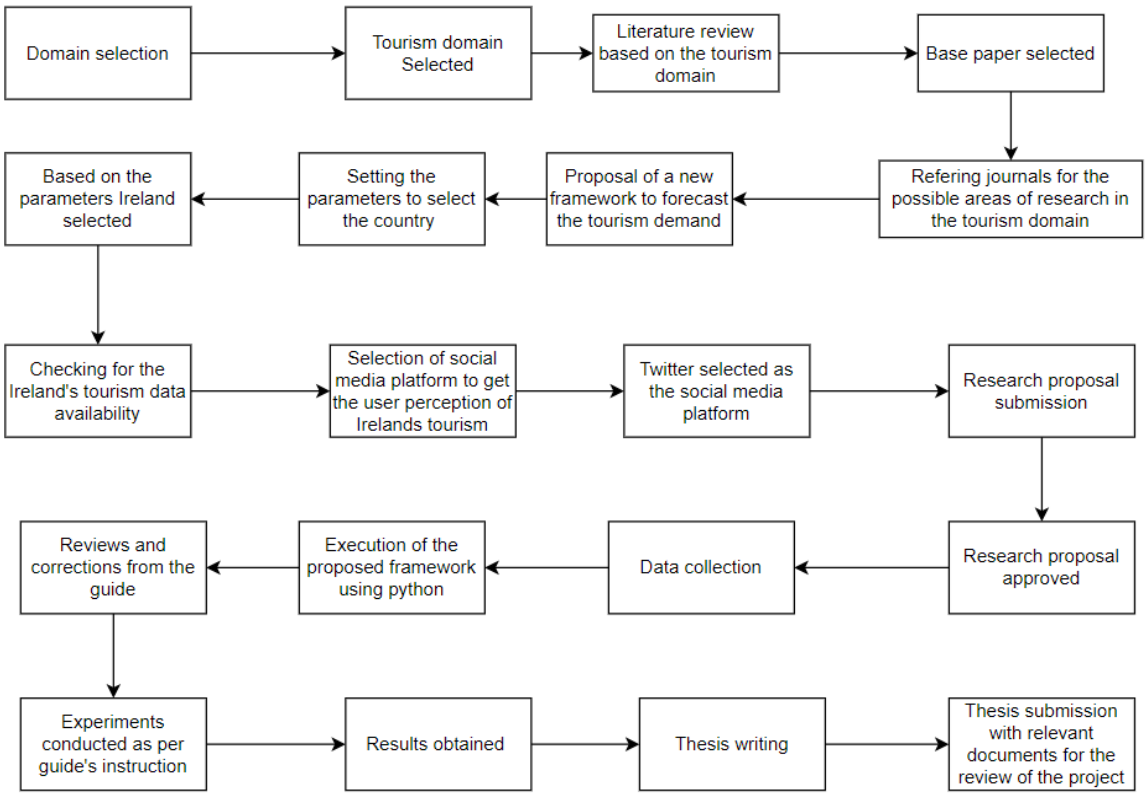

Fig.1. Research methodology process flow diagram

3.1. Sample data collection

Ireland is chosen as the destination for this research because it is one among the top 100 most visited English-speaking tourist locations in the world (Brilliant Maps, 2015). Twitter has been chosen as the primary data source to get the social media data because it has been observed that the most of their tourist share their views and experiences about the places visit in twitter as tweets (Alana K. Dillette). Most of the marketing companies across the globe are using twitter as their advertising tool to carry out their marketing campaigns to reach out to their customers (Stephanie Hays, 2012).

In this research tweets have been collected from the twitter using a open source python package called pythonscraper. Since there has been difficulty in collecting the historical tweets using twitter API which allows to collect tweets for the last seven days. Tweets for five years from 2012 to 2016 have been collected by using the keywords “Ireland” “ireland” the start date and the end date using twitterscraper. The tweets collected in English language as the NLP models that is used are designed for English language only. A total of 50000 tweets have been with 10000 tweets from each year are collected to listen to the tourist’s opinion in twitter through NLP models. Tourist arrival numbers are extracted from the Ireland tourism website.

3.2. Natural language processing

3.2.1. Text preprocessing

Our research specifically deals with tweets that are in textual form. The first task is to remove URLs from the tweets, followed by the removal of stop words. A critical initial step in text pre-processing is tokenization, which involves breaking down the text into individual words. Subsequently, we apply lemmatization to convert the pre-processed text into its dictionary form.

3.2.2. Topic extraction

In the scope of this study, the implementation of the Natural Language Processing technique, Latent Dirichlet Allocation (LDA), becomes integral for the extraction of topics from a expansive high volume of tweet data. The defining parameter, represented by 'k', corresponds to the desired quantity of topics. Determination of an optimal 'k' value involves the utilization of a coherence score.

The methodology employed within this model aligns with Bayesian frameworks and is underpinned by mixture model concepts. The fundamental assumption is that the constituent words of each text (tweet) originate independently from a combination of separate containers, with each container holding a distinct word set. The generative procedure for each tweet adheres to the following sequence (Blei, 2003):

The inaugural step encompasses the decision for the document's length, indicated by N, which is chosen from a Poisson distribution with the parameter ζ. The subsequent step necessitates determining the topic distribution per document from a Dirichlet distribution, guided by the parameter α. This α serves as a prior for the topic distributions.

In respect to each of the N words in the tweet:

a) A topic, symbolized by Z_n, is chosen from a multinomial distribution.
b) A word, symbolized by W_n, is picked from a multinomial probability distribution contingent on the chosen topic Z_n.

The said methodology facilitates the modelling of tweet generation by encapsulating the inherent topics that drive the presence of certain words in each tweet. The application of the Bayesian framework enables comprehensive understanding of the topic distribution across all tweets, thereby illuminating patterns within the textual data.

3.2.3. Emotional Evaluation of Topics

After the topic extraction, the process continues with the sentiment analysis of each topics from the LDA model. This is executed with the assistance of Text Blob, an open-source software

library in Python. Topics undergo categorization into three primary sentiment classifications: positive, negative, and neutral. Following sentiment designation, a mean sentiment score per topic is computed, offering an aggregated sentiment rating.

3.3. Model building

The study involves predicting the tourist numbers which is our target variable. Now the mean sentiment score of topics extracted from the LDA model along with year are the independent variables and the tourist arrival numbers taken from the Ireland tourism website. The data is pre-processed by removing all the missing values and outliers. The data types of the models have been changed to build the models on top of the dataset.

Model construction involved the application of four separate methodologies: Linear Regression, Support Vector Regression with a linear kernel (SVR-linear), Support Vector Regression with an RBF kernel (SVR-RBF), and XGBoost. The repeated implementation of Leave One Out (LOO) training and evaluation promoted an extensive performance examination. During the training phase, each methodology received tuning through the allocation of relevant parameters and instruction on the training set, sparing one datum for validation purposes. Selection of the final model focused on superior accuracy and resilience across all repetitions.

3.4. Performance evaluation

The models used are Support vector regression with gaussian kernel, Support vector regression with linear kernel, Random Forest regression, linear regression model. Mean absolute error (MAE) and mean absolute percentage error (MAPE) are the two-performance metrics which are used to compare the performance of the criterion models and the proposed model. To overcome the overfitting of the models k-fold cross validation technique is used. To use this technique, we should decide the number of subsets into which the datasets have to be divided. This paper gives the number of fold value as 10 since most of the tourism related literature have used the same 10 to overcome the overfitting issue.

4. Design Specification

4.1. Data collection

- **Source:** Twitter and Ireland tourism website.
- **Tools:** Python scraper for tweets, manual extraction for tourist arrival numbers.
- **Time frame:** 2012 to 2016
- **Volume:** 50,000 tweets, with 10,000 tweets from each year.

4.2. Data cleaning

- **Removal of URLs:** From the tweet.
- **Removal of Stop Words:** From the Textual data.
- **Tokenization:** Breaking down the text into individual words.
- **Lemmatization:** Converting text into dictionary form.

4.3. Feature extraction

- **Topic extraction:** Using Latent Dirichlet Allocation (LDA).
- **Sentiment analysis:** Using Text Blob for positive, negative, and neutral sentiment classifications.

4.4. Model building

- **Models implemented:** Linear regression, Random forest, SVR, XGBoost.
- **Training and Validation:** Using Leave One Out (LOO) training and evaluation.
- **Tuning:** Allocation of relevant parameters for each model.
- **Target variable:** Tourist arrivals number
- **Independent variables:** Mean sentiment score of topics, year.

4.5. Evaluation metrics

- **Metrics:** MAE and MAPE are the performance metrics.
- **Comparison and Visualisations:** Comparing and visualising the accuracy of all models.

4.6. Ethical considerations:

- **Privacy:** Ensuring that the data collection and analysis comply with ethical guidelines and user privacy.

4.7. Tools and Technologies

- **Programming Language:** Python.
- **Libraries:** Pythonscraper for data collection, Text Blob for sentiment analysis, relevant libraries for machine learning models (e.g., scikit-learn for Linear Regression, SVR; XGBoost for XGBoost model).

5. Implementation

This part talks about how we got our info. We got some from what people say online and some from numbers about tourists. We used a cool way to understand the words people use online. It's called NLP. Then we made the best guesses about how many tourists will come using the info we got. Figure 2 gives a broader picture how the forecasting models are built.

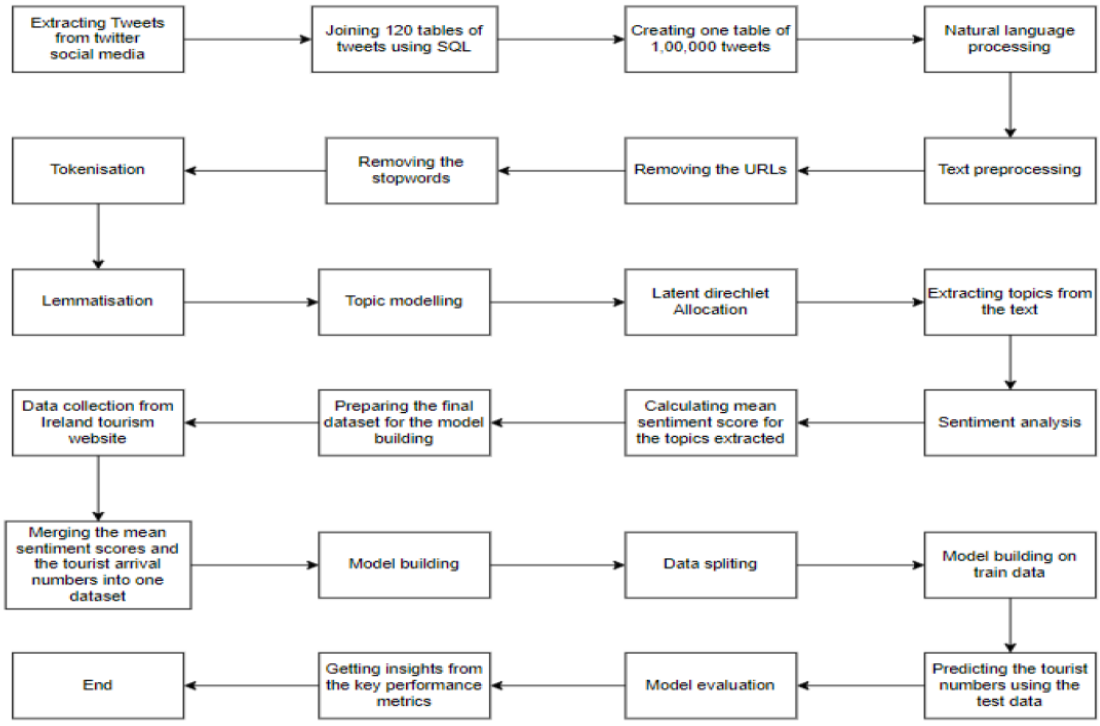

Fig.2 Process flow diagram for building the forecasting models

5.1. Loading required packages

Based on the user's comfort the Python notebooks are selected, jupyter notebook has been used for this research. There is a wide range of notebook options like Google Collab, Azure, etc. Firstly, the packages required for performing the Topic modeling using LDA have been loaded.

5.2. Data cleaning

The next step is to refine the data collected from Twitter which involves the process of removing the emojis, URLs followed by lemmatization and tokenization. The need for the removal of emojis and URLs arises because, at the end of topic modeling, models give no value to figuring out the topics from the bag of words.

5.3. Data Pre-processing

The first step of textual data preprocessing is the conversion of sentences into words followed by changing the form of words from their root. The next step is to remove the number of words that are common and also not relevant to our topic modeling purpose.

5.3.1. Tokenization

Tokenization is done to break the text into sentences and then the sentences into words. Punctuations are removed. Less number of characters in a word are removed including the removal of stop words. Lemmatized words undergo a transformation from third-person stemmed words to first person, both past and future tense verbs to present tense, then reduced back to their base form.

Before processing any text data, we always begin with tokenization. This involves breaking sentences into components like words, punctuation, symbols, etc., using rules tailored to each language. SpaCy is a pre-established natural language processing model that can discern the connections between words.

5.3.2. Lemmatization

Lemmatization is the process of breaking down words into their simplest versions. The total number of unique words in our vocabulary is reduced by this method. As a result, the document-word matrix has fewer columns and is more compact.

5.4. Topic modeling

A base model is initially developed to monitor progress during the hyper-parameter tuning phase. The LDA topic model algorithm necessitates a document word matrix and a dictionary as primary inputs. Initial steps include creating a dictionary, filtering out extremes, and generating a corpus object. The input to the LDA model is the generated corpus object.

Everything necessary to train the LDA model was prepared. Beyond the corpus and dictionary, the number of topics must also be specified. Five were selected for the base model. During the hyperparameter tuning phase, the optimal number of topics will be determined.

Displaying the topics generated by the LDA model as shown in the figure offers an approximate insight into the subject associated with each bag of words. Notably, the bags of words are arranged from the most pertinent to the least relevant for every topic.

Model perplexity and topic coherence serve as useful metrics to evaluate the effectiveness of a topic model. Historically, the topic coherence score has proven especially informative. A coherence score of 0.256 is notably low for most LDA models, yet it's worth noting this represents just the base model, which has the potential for significant improvement. The next step is to analyze the generated topics and their associated keywords after building the LDA model. The graphical chart as shown in the figure 4 in the pyLDAvis package stands out as a superb tool because it was created to work perfectly with Jupyter notebooks.

```
------ Topic 0 ------
ireland not northern m uk come live know s right

------ Topic 1 ------
ireland northern m amp s brexit not love irish state

------ Topic 2 ------
ireland job hire come good time jobfairy apply irish travel

------ Topic 3 ------
ireland year day good love new northern christmas c talk

------ Topic 4 ------
ireland new year go amp m today northern not day
```

Fig.3 Topics generated from the base model

Every bubble on the left-side plot of the figure signifies a topic. The size of the bubble indicates the prevalence of that topic. An effective topic model should display sizable, distinct bubbles distributed across the chart rather than gathered in one section. Models with an excess of topics often show overlapping, small bubbles concentrated in a specific area of the chart. Hovering over one of the bubbles updates the words and bars on the right side of the chart. The significant keywords represent those words that define the chosen topic. A visually appealing topic model has been constructed.

5.5. Hyperparameter tuning

5.5.1. Grid search

The primary tuning parameter for LDA models is components, which represent the number of topics. The search will also encompass learning decay, which dictates the learning rate. Other potential parameters for exploration include learning offset, which down-weights early iterations and should be greater than 1, and maximum iterations. Exploring these might be beneficial, given sufficient time and computational resources.

It's understood that the potential number of topics likely exceeds 10, yet the grid search determined that 10 topics outperform other quantities. It appears that Scikit-learn's grid search prioritizes perplexity over coherence value, and in this context, coherence value proves more effective. The subsequent stage will focus on determining the optimal number of topics.

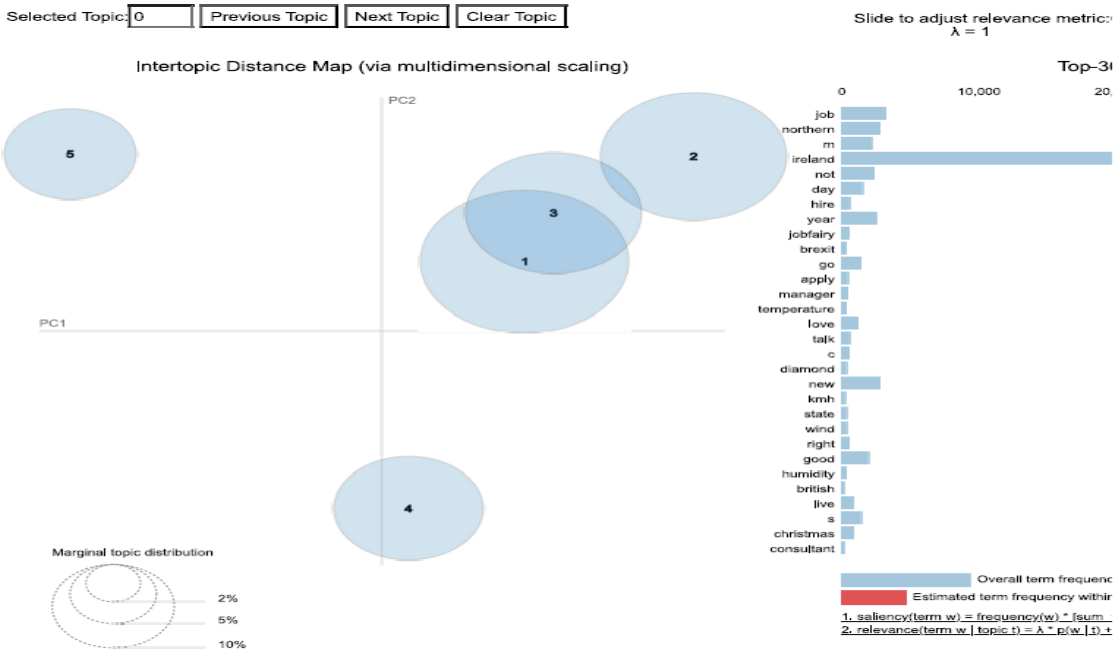

Fig.4 Topic distance Visualisation

5.5.2. Optimum number of topics

Selecting a topic count that coincides with the conclusion of swift coherence growth typically yields comprehensible and interpretable topics. Opting for a higher value might further delineate sub-topics. If identical keywords recur across various topics, it's likely an indicator that the chosen 'number of topics' is excessive. Adhering to these guidelines, 15 topics were determined to be the optimal count for this specific application as shown in figure 5.

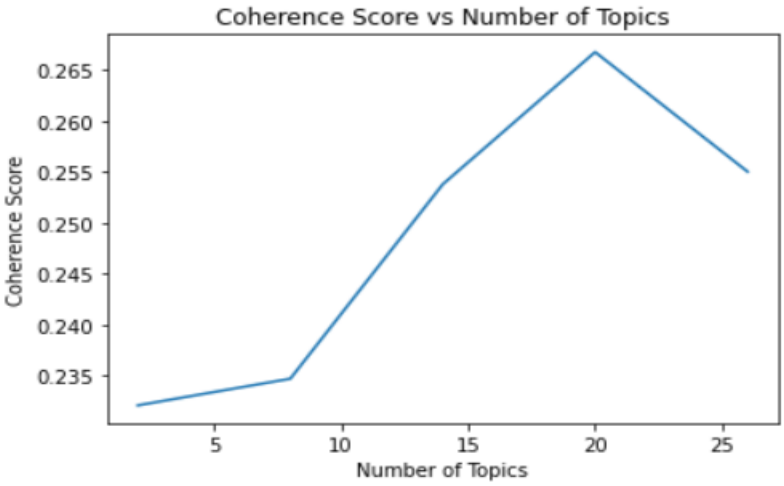

Fig.5 Coherence Score vs Number of Topics

5.5.2. Optimum number of passes

Passes, chunk size, and update every are parameters interconnected through an EM/Variational relationship. Delving into the intricacies of EM/Variational Bayes isn't the focus here, but those eager for more can consult a related Google forum post and its associated paper. In this context, tuning the number of passes is deemed sufficient, given that chunk size is not significant, and altering the update every isn't expected to considerably impact the final outcomes. Numerous pass counts were tested, and 20 appeared to yield the optimal outcome. However, the improvement was marginal, amounting to just a few decimal points as shown in figure 6.

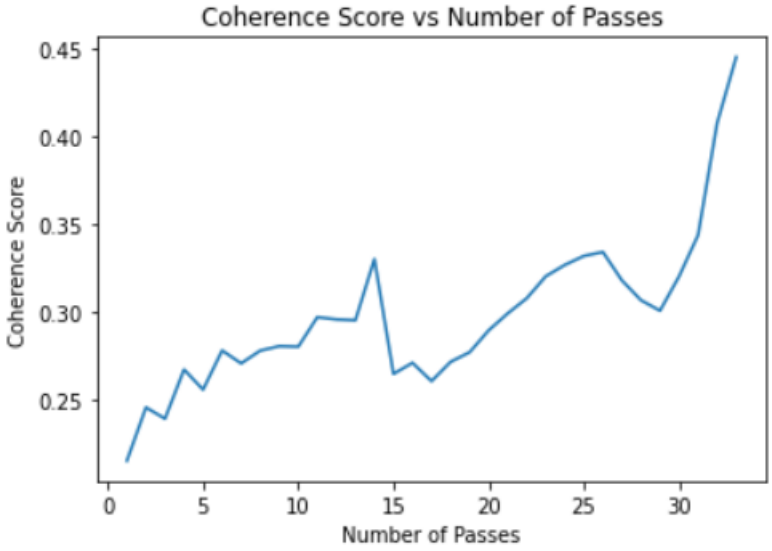

Fig.6 Coherence Score vs Number of Passes

5.5.2. Optimum number of Decay

In the final model decay value set is 0.5 in accordance with the best coherence score as shown in figure 7.

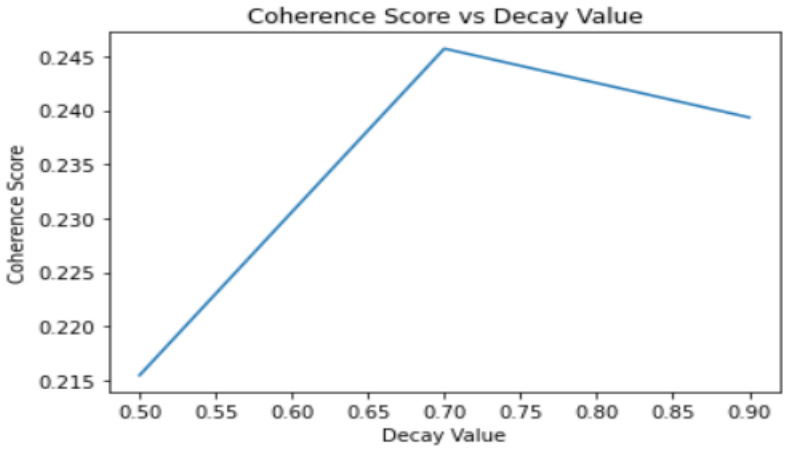

Fig.7 Coherence Score vs Decay Value

5.5.3. Minimum probability

Topics with probabilities below the specified threshold are disregarded by this hyperparameter as shown in figure 8.

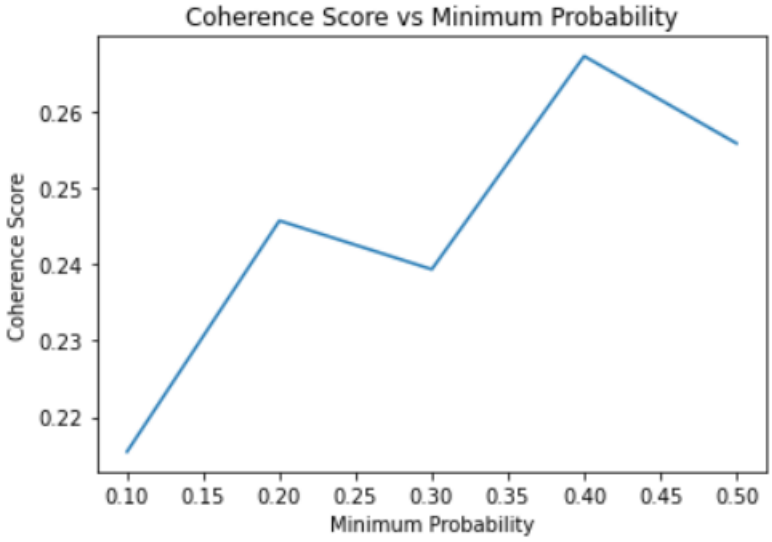

Fig.8 Coherence Score vs Minimum Probability

5.6. Final LDA model

Now a final LDA model is built with the above hyperparameter tuning done and the number of topics extracted is 15 as shown in figure 9 with all the hyperparameters.

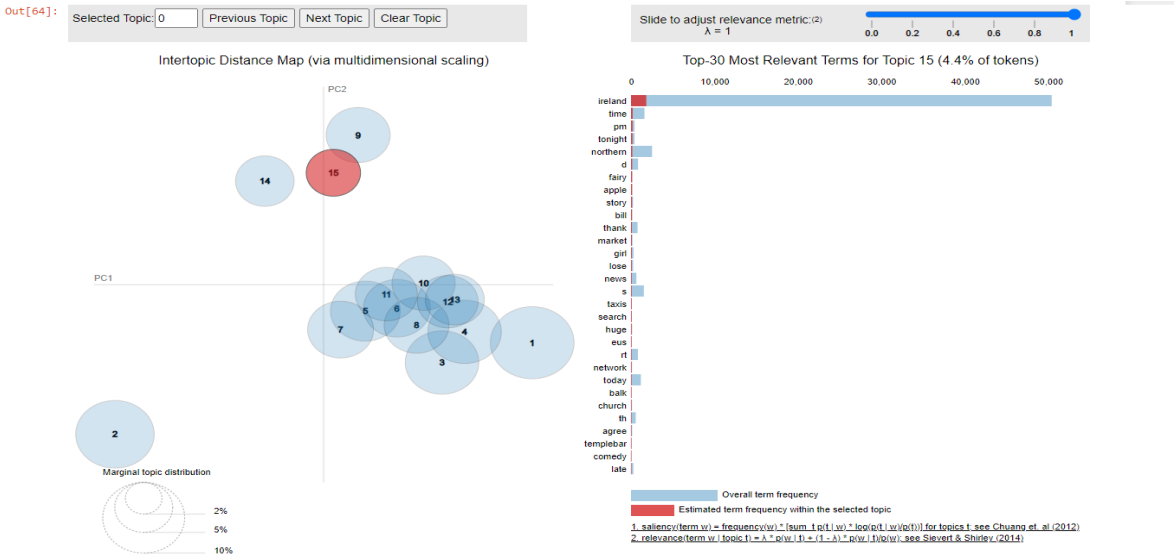

Fig.9 Final LDA model

5.6. Machine learning models for tourism demand forecasting

The study involves predicting the tourist numbers which is our target variable. Now the mean sentiment score of topics extracted from the LDA model along with year are the independent variables and the tourist arrival numbers taken from the Ireland tourism website. The data is pre-processed by removing all the missing values and outliers. The data types of the models have been changed to build the models on top of the dataset.

Model construction involved the application of four separate methodologies: Linear Regression, Support Vector Regression with a linear kernel (SVR-linear), Support Vector Regression with an RBF kernel (SVR-RBF), and XGBoost. The repeated implementation of Leave One Out (LOO) training and evaluation promoted an extensive performance examination.

During the training phase, each methodology received tuning through the allocation of relevant parameters and instruction on the training set, sparing one datum for validation purposes. Subsequently, methodologies received scrutiny via LOO validation, providing a thorough measure of their predictive abilities. Selection of the final model focused on superior accuracy and resilience across all repetitions. This process provided an understanding of the model's capacity to accurately forecast tourist arrival quantities based on specified topics, thus substantiating its potential for practical application.

6. Evaluation and Results

6.1. Experiments

In this research, experiments were conducted to find the optimal parameters for the LDA topic modeling. Also, experiments were conducted to determine the efficient machine learning model to forecast the tourism demand using the topics extracted and sentiment scores.

6.2. Experimentation with LDA Parameters

6.2.1. Number of topics

To determine the ideal number of topics for the LDA model, an initial range of topics from 5 to 20 was explored. For each specified K value, the LDA model underwent training, followed by a computation of the coherence score. The coherence score peaked at K=15 as shown in figure 5, suggesting that 15 topics are optimal for this dataset.

6.2.2. Number of Passes

The next experiment was conducted to fix the number of passes for the LDA model. The experiments were conducted with pass counts varying from 1 to 30. The coherence scores peaked at 20 and hence the number of passes fixed was 20 as shown in figure 6.

6.2.3. Minimum Probability

The next experiment was conducted to fix the number of passes for the LDA model. The experiments were conducted to find the topics with minimum probability. Those topics with minimum probability in crelation to the coherence scores were disregarded as shown in figure 8.

6.2.4. Decay Value

To identify the best decay value for the LDA model. In the experimentation process, decay values between 0. 1 and 1. 0 were evaluated. The findings indicated that a decay value of 0. 5 produced the maximum coherence score as shown in figure 7.

6.3. Experimentation with machine learning models

The aim was to pinpoint the most effective machine learning model for forecasting tourism demand. During the experimentation, four models, namely Linear Regression, SVR, Random Forest, and XGBoost, underwent training and validation. Performance metrics, including MAE and MAPE, were used to assess each model.

6.4. Topics Extracted from the final LDA model

The number of topics extracted by the LDA model is 15 which is fixed before building the LDA model by setting the highest coherence score. Following the LDA procedure with the

predetermined K value, Fig. 8 displays the outcomes. There are 15 circles on the left, each representing one of the topics identified by the LDA model. The circle sizes correspond to the frequency of each topic within the tweets. At the top-right, the Lambda (λ) value is adjustable, and a value of 0.6 was selected based on recommendations from Sievert and Shirley (2014).

The right section displays the top 15 salient terms for every topic. Topic names were derived by spotting a consistent theme among the most prevalent words within each topic. The pertinence of these topic names to tourism was then assessed. Topics with a discernible theme and a connection to tourism were retained, while the rest were disregarded. From this approach, 15 tourism-related topics emerged, as illustrated in Table 1.

Upon analyzing the average sentiment score for each topic annually, there was an examination of how attitudes towards specific topics correlated with shifts in tourist arrivals in Ireland. Table 2 gives the descriptive data for sentiment fluctuations across each topic over a decade and the findings indicate that sentiment scores for all topics. On average, sentiments were most positive about the topics of football and beach.

Table 1

Topics identified using LDA after filtering

Topics
Jobs
Weather
Travel
City
Irish whiskey
Christmas
Football
Business environment
Party
Accomodation
Guinness
Castle
Beach
New year
Marketing

6.5. Performance of the forecasting models

MAE and MAPE are the performance metrics used for selecting the best model that can forecast tourist numbers. The models employed are Linear regression, SVR, Random Forest, and XGBoost. The model with the least error will be considered the best model that can be employed in the tourism domain to forecast the tourism demand. The table and the graph as shown in figure 10 below gives the MAE and MAPE of the ML models.

Table 2
Performance of the forecasting models

Model	MAE (mean)	MAPE (mean)
Linear regression	59	0.6 %
Random forest	818	8.1 %
SVR	1554	15 %
XGboost	315	3.1 %

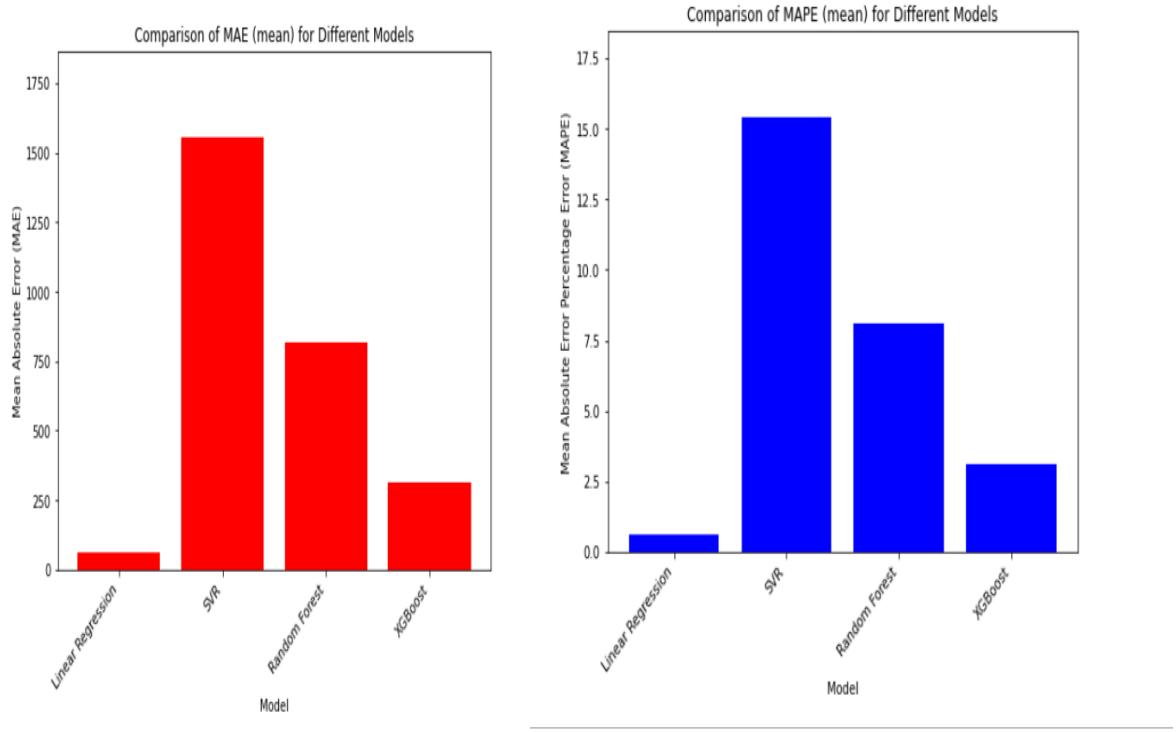

Fig.10 Comparison of MAE and MAPE of different models

Besides tweets about Christmas events, attitudes toward the business environment also significantly affect tourist arrivals in Ireland. This observation in fig 11 aligns with business studies that emphasize how the business setting in both the destination and origin points impacts tourist choices. For leisure tourism, an improved environment typically signifies enhanced tourism facilities, potentially drawing in tourists (Ghialy Yap, 2011). Furthermore, a consistent business climate appeals to business travelers. When companies have a positive perception of the destination's business environment, travel approvals become more frequent (forecasting, 2005).

6.6. Methodological implications

The methodological structure of this study gives us a completely different picture of analysing tourism demand. The framework used in this research involves the combination of topic modeling, sentiment analysis, and the XGBoost forecasting model to predict tourist arrivals and deep dive into factors that could influence tourism demand. As outlined in the literature review, determining the right independent variables for tourism demand modeling is complex due to the ever-changing preferences and perceptions of tourists.

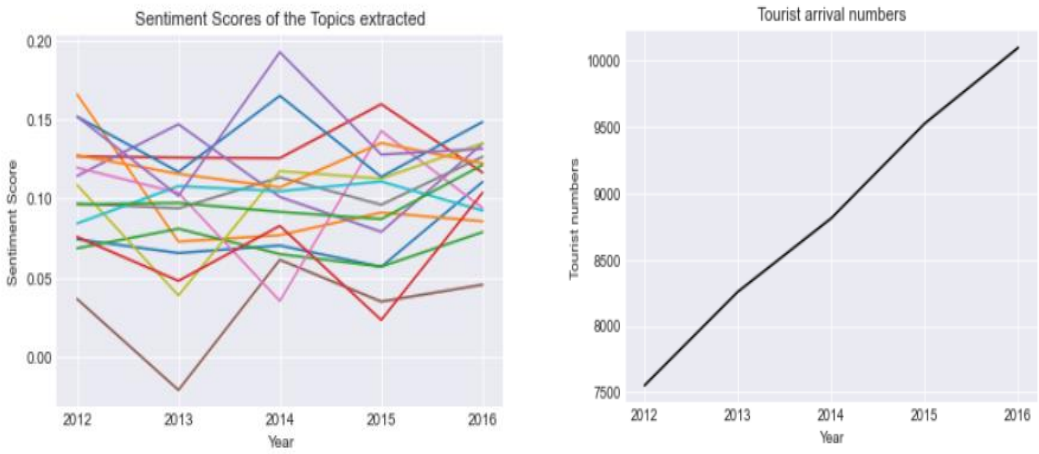

Fig. 11 Comparison of Tourist arrival numbers and sentiment scores

The role of NLP in capturing consumer thoughts in social media gives a whole new dimension to the tourist demand forecasting domain. This combination of topic modeling and sentiment analysis opens a new gateway to get the end user's thoughts and opinions effectively which can be used in many applications apart from the tourism industry.

7. Conclusion and future work

This study gives us a new framework by hearing the target audience's thoughts on social media. Linear regression is the model that can be adopted for this framework which is the major finding. This research is limited to 10 years of data which is limited so further research can be done with data from multiple platforms and also using some deep learning models in the future.

The trial of the forecasting method was just for one place, and it only looked at what people said on Twitter. So, the things found out might not be the same for other places or from info on other online sites. The study only used English words from Twitter, so it didn't see what people said in other languages. It might be good if other people try new ways, like BERTopic

or Multilingual BERT, to get info from words in many languages on different online sites. Maybe next time, people can see what's said on other social media platforms sites like Weibo, Ctrip, and WeChat. The info in this study was from each year for 10 years. But that's not a lot of data. So, the way the guessing was done might not be the best. It might be better if data is collected every week or day. That way, there's more data and the guessing might be better.

Also, it might be good to mix in other info when guessing how many tourists will come. Like, how much stuff is sold at tourist places, what customers say when you ask them things, updates from the tourism business, and news.

References

Alana K. Dillette, S. B. (n.d.). Tweeting the Black Travel Experience: Social Media Counternarrative Stories as Innovative Insight on TravelingWhileBlack.

Blei, D. M. (2003). Latent Dirichlet allocation.

Brilliant Maps . (n.d.). Retrieved from Brilliant Maps : https://brilliantmaps.com/top-100-tourist-destinations/

Brilliant Maps, 2. (2015). Retrieved from Brilliant Maps: https://brilliantmaps.com/top-100-tourist-destinations/

Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting.

Cho, V. (2010). A study of the non-economic determinants in tourism demand. *International journal of tourism research*.

De Vita, G. K. (2013). Role of the exchange rate in tourism demand. Ann. *Tourism Res*.

forecasting, J. o. (2005). A leading indicator approach to predicting short-term shifts in demand for business travel by air to and from the UK. *Journal of Forecasting*.

Ghialy Yap, D. a. (2011). Investigating other leading indicators influencing Australian domestic tourism demand. *Science Direct*.

Haiyan song, R. .. (2019). A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting. *Science direct*.

Jean Max Tavares, N. c. (2016). The determinants of international tourism demand for Brazil. *Sage journals*.

Liu, A. P. (2017). Tourism's vulnerability and resilience to terrorism . *Science direct*.

Muhammad shafiullah, U. k. (2018). Determinants of international tourism demand: Evidence from Australian states and territories. *Sage journals*.

Stephanie Hays, S. j. (2012). Social media as a destination marketing tool: its use by national tourism organisations.

Wai Hong Kan Tsui, F. B. (2016). International arrivals forecasting for Australian airports and the impact of tourism marketing expenditure. *Sage journals*.

Yuan-Yuan Liu, F.-M. T.-H. (2018). Big Data analytics for forecasting tourism destination arrivals with the applied Vector Autoregression model. *Science direct*.

Mehrbaksh Nilashi , F.B. (2017) A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques. *Science direct.*

Yulie li, Zhibin lin, Sarah Xiao, F.B. (2022) Using social media big data for tourist demand forecasting: A new machine learning analytical approach. *Science direct.*