



PREDICTION SOLUTIONS 2024

International Conference on Prediction Solutions for Technical and Societal
Systems

ISBN: 978-1-68558-212-8

November 3rd - 7th, 2024

Nice, France

PREDICTION SOLUTIONS 2024 Editors

Petre Dini, IARIA, USA/EU

PREDICTION SOLUTIONS 2024

Forward

The International Conference on Prediction Solutions for Technical and Societal Systems (PREDICTION SOLUTIONS 2024) event series is concerned with prediction facets in various technical, societal, and sense-based observations by gathering information, identify patterns, and defining (future) system requirements. The output of prediction is a set of critical factors used to evaluate systems' characteristics qualitatively and quantitatively, as well as forecast the service demand directions. Prediction provides input on scalability, resilience, resource allocation, etc. and provides guidelines for future decisions. The conference was held in Nice, France, November 3 - 7, 2024..

Organization predictions, market predictions, weather predictions, automation predictions, or ultimately, Garner curve technical predictions, constitute mechanisms and fundamental driving (future evaluation) methodologies for society's industrial, investment, and customer behavioral expectations.

Generally, the prediction is concerned with estimating the outcomes for unseen data (rather than document observations); a sub-discipline considering time series data is the forecast, where the temporal dimension is shorter. Predictive processes are based on historical data combined with statistical modeling, data mining techniques, and machine learning. Prediction is leveraging data and patterns that are continuously gathered by forming datasets, using simulated samples, and statistics techniques.

The process starts with sensing data, identifying situations, and changes of different situational parameters. Via correlation functions and hypothesis, a variation of an input metric under observation triggers a set of metrics along with quantitative/qualitative characterizations (what? how much? when, for how long? etc.).

Most of the technical decisions in transportation systems based on long terms traffic patterns, QoS delivery and SLA satisfaction in wireless and ubiquitous systems (based on QoE) or progresses in adopting 5G/6G and Industry 4.0/5.0) are based on an ensemble of successful technical factors, successful service delivery, and lessons learned. However, further development is linked to collecting data, applying mechanisms, predicting, testing, and confirming the perceived tendency.

Complex domains, dynamics changes, and entangled metrics relations make some events difficult to predict, with a large range of impreciseness. Generally, predictions are based on observable (sensed) metrics of a phenomenon in a given context, without considering internal variables/parameters of the phenomenon.

For forecasts (shorter terms predictions), the precision is critical (eventually concluding with an estimated probability). For hurricanes, the prediction of landing location is critical. Therefore, internal parameters are considered. For very short terms, forecasting is only partially possible, and dynamics that change suddenly leave no time to process unpredictable details (tornadoes).

Recursive predictions might be triggered by repetitive patterns, which are observed occurrences of a series of predictions with high score, eventually being endorsed as (pseudo-)laws. The most famous is the so-called "Murphy's Law".

We take this opportunity to thank all the members of the PREDICTION SOLUTIONS 2024 Technical Program Committee as well as the numerous reviewers. The creation of such a broad and high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to the

PREDICTION SOLUTIONS 2024. We truly believe that, thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the PREDICTION SOLUTIONS 2024 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope the PREDICTION SOLUTIONS 2024 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress with respect to prediction technologies. We also hope that Nice provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city

PREDICTION SOLUTIONS 2024 Chairs

PREDICTION SOLUTIONS 2023 Steering Committee Chair

Sandjai Bhulai, Vrije Universiteit Amsterdam, The Netherlands

PREDICTION SOLUTIONS 2024 Steering Committee

Corné de Ruijt, Windesheim University of Applied Sciences, The Netherlands

Robin van Ruitenbeek, Lensor, the Netherlands

Ieva Meidutė-Kavaliauskienė, Vilnius Gediminas Technical University & General Jonas Žemaitis Military Academy of Lithuania, Lithuania

PREDICTION SOLUTIONS 2024 Publicity Chair

Lorena Parra Boronat, Universitat Politècnica de Valencia, Spain

Sandra Viciano Tudela, Universitat Politècnica de Valencia, Spain

PREDICTION SOLUTIONS 2024

Committee

PREDICTION SOLUTIONS 2024 Steering Committee Chair

Sandjai Bhulai, Vrije Universiteit Amsterdam, The Netherlands

PREDICTION SOLUTIONS 2024 Steering Committee

Corné de Ruijt, Windesheim University of Applied Sciences, The Netherlands

Robin van Ruitenbeek, Lensor, the Netherlands

Ieva Meidutė-Kavaliauskienė, Vilnius Gediminas Technical University & General Jonas Žemaitis Military Academy of Lithuania, Lithuania

PREDICTION SOLUTIONS 2024 Publicity Chair

Lorena Parra Boronat, Universitat Politècnica de Valencia, Spain

Sandra Viciano Tudela, Universitat Politècnica de Valencia, Spain

PREDICTION SOLUTIONS 2024 Technical Program Committee

Abdelkader Adla, Oran1 University, Algeria

Liaqat Ali, University of Science and Technology of Fujairah, UAE

Sandjai Bhulai, Vrije Universiteit Amsterdam, The Netherlands

Corné de Ruijt, Windesheim University of Applied Sciences, The Netherlands

Abdessamad Didi, University Mohamed I, Morocco

Deepti Gupta, Huston-Tillotson University, USA

Leila Hamdad, Ecole Nationale Supérieure d'Informatique, Algeria

Santosh Joshi, Florida International University, USA

Yasuko Kawahata, Rikkyo University, Japan

Henry Leung, University of Calgary, Canada

Oleksandr Letychevskiy, Glushkov Institute of Cybernetics of National Academy of Science, Ukraine

Maria Elisabete Neves, Polytechnic of Coimbra | Coimbra Business School Research Centre | ISCAC & CETRAD | UTAD, Portugal

Alexander Makarenko, Institute of Applied Nonlinear Analysis at National Technical Institute of Ukraine (Igor Sikorsky Kyiv Polytechnic Institute), Ukraine

Ieva Meidutė-Kavaliauskienė, Vilnius Gediminas Technical University, Lithuania

Rasha Osman, University of Khartoum, Sudan

Bochra Rabbouch, Sousse University, Tunisia

Nabil Searcher, Ecole supérieure en informatique - Sidi Bel Abbes 8 mai 1945, Algeria

Patrick Siarry, Université Paris-Est Créteil, France

Angelo Sifaleras, University of Macedonia, Greece

Robin van Ruitenbeek, Lensor, the Netherlands
Shaomin Wu, Kent Business School, UK

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission to reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.


I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Sentiment Analysis of Social Media: A Case Study on Big Tech Layoffs <i>Mehdi Mekni and Ahmed Muntasir Hossain</i>	1
Tourist Mobility Forecasting with Region-based Flows and Regular Trips <i>Fernando Terroso-Saenz, Juan Morales-Garcia, Miguel Puig, Ginesa Martinez-del Vas, and Andres Munoz</i>	10

Sentiment Analysis of Social Media: A Case Study on Big Tech Layoffs

Mehdi Mekni , Ahmed Muntasir Hossain

Department of Electrical and Computer Engineering and Computer Science
Connecticut Institute of Technology

University of New Haven

West Haven, United States

e-mail: {mmeckni | ahoss1}@newhaven.edu

Abstract—Digital reputation management systems are essential for maintaining and improving online reputations. However, current systems like Online Social Network Interactions face critical issues, such as limited effectiveness, high costs, and inaccuracy. Sentiment analysis, a natural language processing technique, can enhance digital reputation management by extracting opinions, emotions, and attitudes from textual data. We propose developing Sentiment Analysis of Social Media, an open-source, multi-channel, multi-engine sentiment analysis software. SASM collects data from Twitter, Reddit, and Tumblr, filtering and analyzing trends using Microsoft Text Analytics, IBM Watson Natural Language Understanding, and Google Cloud Natural Language API. A case study on Google, Amazon, and Microsoft will validate the system and evaluate the performance of the three engines. SASM offers a unique approach by providing reliable sentiment analysis, leveraging multiple engines, and sourcing diverse social media content, enabling companies to manage their digital reputation effectively and affordably.

Keywords—Sentiment Analysis, Natural Language Processing, Social Media Platforms.

I. INTRODUCTION

Social media platforms serve as communication channels between companies and customers. Companies promote their products, and customers post reviews and ask questions. Analyzing customer-posted content is crucial, and sentiment analysis is effective due to the large data volume. Sentiment analysis uses Natural Language Processing (NLP) and text analysis tools to determine the sentiment of a text, categorized as positive, negative, or neutral [1]. Major applications include brand monitoring, campaign monitoring, and competitive analysis [2]. This helps companies evaluate their community presence and make informed decisions. However, companies struggle to collect and filter reviews effectively.

Sentiment analysis software analyzes direct reviews (website, emails, surveys), limiting exposure to social media data. Recently, Samara et al. [3] developed Online Social Network Interactions (OSNI) Analytics to collect and analyze Tweets about companies using Microsoft Text Analytics. The engine categorizes Tweets as negative, mixed, neutral, or positive and returns the results to the company. This research aims to integrate multiple social media channels (Twitter, Reddit, Tumblr) into a Sentiment Analysis of Social Media (SASM), enhancing its ability to analyze social media content. Additionally, the project plans to expand SASM's sentiment analysis capabilities by including Microsoft Text Analytics (MTA), IBM Watson Natural Language Understanding (IWNLU), and Google Cloud Natural Language API (GCNLA). The research

question associated with our project is the following "R1: How do MTA, IWNLU, and GCNLA differ in their sentiment analysis of social media content, and which engine provides the most accurate and nuanced sentiment classification when applied to data from various social media channels?"

The goal of this project is to investigate and confirm the technical feasibility of integrating multiple sentiment analysis engines with social media channels, while also comparing their performance using a real-world case study. Data will be collected about Google, Amazon, and Microsoft, which are major IT contributors and recently experienced mass layoffs, sparking social media discussions. The project will use the Software Development Life Cycle (SDLC). Each social media platform offers a free connector for academia and research, allowing data retrieval. The data will be processed, categorized, and evaluated using sentiment analysis engines. Future work includes combining results from multiple engines to develop a hybrid model for more reliable and accurate results.

The rest of the paper is organized as follows: Section II presents sentiment analysis techniques and applications. Section III outlines the methodology, the studied use case and the associated data management plans. Section IV provides a comprehensive description of SASM, including software requirements engineering, design, and architecture. Section V presents case study results and demonstrates SASM's contributions. Sections VI and VII discuss technical choices, methodologies, and future work.

II. RELATED WORK

In this section, relevant studies of scholars, such as Haruechaiyasak et al. [4], By et al. [5], Baron and Mekni [6], who analyzed social media content and its impact on corporate digital reputation and branding were reviewed. Moreover, their approaches for multi-platform data collection were analyzed to determine an effective method for the proposed study. The rest of the literature review is organized as follows: Impact of Social Media on Corporate Digital Reputation and Branding, Sentiment Analysis of Content on Multiple Social Media Channels, Advantages of Proprietary Text Analytics Engines, and Evaluation of Multiple Sentiment Analysis Models.

A. Impact of Social Media on Corporate Digital Reputation and Branding

Haruechaiyasak et al. [4] developed a software framework, S-Sense, to analyze Thai content. They collect data from

Twitter and Pantip, a Thai language forum, about mobile services. S-Sense evaluates analysis modules and components and studies the impact of using different lexicon sets for model training. The authors highlight recent advancements in software frameworks for brand monitoring, campaign monitoring, and competitive analysis, emphasizing the significance of social media on corporate digital reputation and branding.

Similarly, By et al. [5] investigated over 1000 Facebook posts to gauge public sentiment towards Rai, an Italian public broadcasting service, compared to the private company La7. They use the Sentiment and Knowledge Mining System, iSyn Semantic Center, for data collection and sentiment analysis. The study underscores the role of social media in brand imaging and marketing, as user reviews and ratings on social platforms significantly influence purchase decisions. The research supports analyzing social media content by showing that Facebook posts indicate a positive sentiment towards La7, aligning with observations from Osservatorio di Pavia and Auditel. These institutions' data is used to validate the sentiment analysis results from iSyn Semantic Center.

These scholarly works relate directly to the application of the SASM software model. To improve upon previous research, the proposed study incorporates different text analytics engines in the SASM model and evaluates their performance. It also integrates multiple social media channels (Twitter, Reddit, and Tumblr) to enhance its ability to analyze social media content, resulting in a more statistically significant dataset. Lastly, Haruechaiyasak et al. [4] suggest expanding the S-Sense domain, which aligns with the proposed enhancements for the SASM model.

B. Sentiment Analysis of Content on Multiple Social Media Channels

In several previous studies, data is collected from a single platform, primarily Twitter and Facebook [5], [7]. This limits companies from obtaining a holistic analysis of their products. To address this, Ali et al.[8] collect data from four platforms: Twitter, Reddit, Instagram, and news forums. The research aims to identify disease outbreak locations using social media sentiment. During emergencies, people share information across platforms, and analyzing sentiment and spatiotemporal data reveals people's behavior and geographic locations [8]. Each platform has a unique community and information-sharing method. Collecting data from multiple platforms provides an unbiased overview of disease outbreaks. This aligns with the proposed study, which aims to identify sentiment towards three major technology companies: Google, Amazon, and Microsoft. Their research underscores the importance of multi-platform data collection. Thus, this paper suggests integrating several media channels into the SASM software model.

C. Evaluation of Multiple Sentiment Analysis Models

Praciano et al.'s [9] research analyzes spatiotemporal trends in the Brazilian election using NLP toolkits, TextBlob and OpLexicon combined with Sentilex, for sentiment analysis.

They apply machine learning algorithms—Support Vector Machine (SVM), Naïve Bayes, Decision trees, and logistic regression—to classify text sentiment. They compare the performance of these algorithms using metrics like accuracy, precision, recall, and F1 score, validating results with 2014 Brazilian presidential election data from the Superior Electoral Court database. Their model effectively predicts election results, with SVM achieving the highest accuracy of about 90%.

This section reviewed relevant studies on the impact of social media on corporate digital reputation and branding. Existing works developed tools and frameworks for data collection and sentiment analysis across various platforms, emphasizing the importance of multi-platform analysis in understanding corporate branding. However, these studies often rely on single platforms, limiting the breadth of insights into brand sentiment. To address these limitations, the proposed study aims to incorporate multiple sentiment analysis engines and analyze data from Twitter, Reddit, and Tumblr, thus providing a more comprehensive dataset. Our project builds on these findings to enhance the SASM software model for more effective sentiment analysis.

III. METHODOLOGY AND DATA

In this study, we created an open-source, multi-channel, multi-engine sentiment analysis software for social media and digital reputation management purposes: SASM. The application collects data/posts from three different social media channels, Twitter, Reddit, and Tumblr. Initially, developer accounts would have to be set up, and API keys for the respective platforms would have to be requested. This would provide the model access to the social media platforms. These platforms' API keys would then be added to the code for the SASM model. Moreover, the program would need to be configured to ensure appropriate data collection as each platform has its own constraints regarding the length of their texts. After configuration, the program will be able to search all these platforms' databases for the particular keyword inputted and return relevant posts. The data will be collected from each of these social media channels simultaneously to compare and observe the trends in the opinions of people across a defined period. The software model then filters, aggregates, and analyzes trends in the sentiment of content posted on social media while leveraging the three sentiment analysis engines. The results of the analysis are then displayed on a dashboard. The performance of each engine on each of the social media platforms would be compared and their relative performance would be determined using pie-charts.

The proposed case study intends to collect data about layoffs from Google, Amazon, and Microsoft. The posts would be collected from the three social media channels. The three proprietary sentiment analysis engines would be used to determine the sentiment of the data collected. The relative agreement of these engines would be evaluated to determine the ideal combination.

A. Case Study

The technology industry has been hit hard by layoffs in 2022 and 2023, with some of the biggest names in the field, such as Amazon, Microsoft, and Google, all experiencing workforce reductions. Through the first week of December 2022, 219,959 people were affected by 1,405 rounds of layoffs at tech companies worldwide, according to TrueUp's tech layoff tracker [10]. Several factors contributed to the layoffs, including the COVID-19 pandemic, inflation, and increasing interest rates [11]. With the current state of the economy, we will likely continue to see these types of workforce reductions in the future [12]. Due to the relevance of the layoffs and the vast data surrounding the conversation, we collected data on Google, Amazon, and Microsoft layoffs for our case study. We used SASM to collect, analyze, and assess several hundred posts from Twitter, Reddit, and Tumblr related to the above-mentioned layoffs. In the following subsections, we will briefly introduce each company we studied as well as the social media platforms (Twitter, Reddit, and Tumblr) we adopted.

1) *Google*: In January 2023, Google announced a plan to lay off approximately 12000 employees. The layoffs were part of a larger restructuring effort to ensure that their product areas and roles align with their highest priorities as a company [13]. To support people during this difficult decision, according to a company statement, Google promised to pay United States (U.S.) employees during the full notification period (minimum 60 days), offer a severance package starting at 16 weeks salary plus two weeks for every additional year at Google, and accelerate at least 16 weeks of GSU vesting. Additionally, the company stated that they would pay all 2022 bonuses & remaining vacation time, and offer 6 months of healthcare, job placement services, and immigration support for those affected. Outside the U.S., employees would be supported in line with local practices and regulations [14].

2) *Amazon*: In November 2022, Amazon laid off approximately 10,000 employees. The impacted employees were working on Alexa and Amazon's Luna cloud gaming service. Additionally, Amazon's hardware and services, human resources, and retail teams were affected during the layoff [15]. According to a company memo, the job cuts were intended "to lower their cost to serve so that they could continue investing in the wide selection, low prices, and fast shipping that Amazon customers love" [16]. Furthermore, the memo stated that U.S. workers will be getting a "60-day non-working transitional period with full pay and benefits, plus additional several weeks of severance depending on the length of time with the company, a separation payment, transitional benefits, and external job placement support" [16].

3) *Microsoft*: In January 2023, Microsoft announced to layoff of 10,000 employees to address the turbulent economic times and rising interest rates. Less than 5% of the company's entire workforce was affected by the job cuts, which ended in the third fiscal quarter of that year, March 2023 [17]. The affected employees were working on HoloLens and Microsoft Edge; moreover, two major game studios under Microsoft, 343 Industries and Bethesda, were significantly impacted by the

process [18]. Microsoft stated that they intend to be thoughtful and transparent throughout the whole process and provide 60 days' notice before termination, above-market severance pay, six months of healthcare coverage, ongoing stock vesting, and career transition support to impacted employees [19].

B. Social Media Data Source

Data was collected from three primary social media channels: Twitter, Reddit, and Tumblr.

1) *Twitter*: According to [1], Twitter is considered one of the best social media platforms through which the opinions/sentiments of a large group of people towards a particular topic can be obtained. Daily, Twitter sends out approximately 500 million Tweets causing it to become one of the largest social media platforms [20]. Due to its massive user base and high volume of real-time, publicly available data; Twitter is an ideal social media platform for sentiment analysis. Furthermore, Twitter users range from all age groups, socio-economic status, and demographics allowing the sentiment extracted from their posts to be an accurate representation of society [21]. The Twitter API V2 provides a recent search endpoint that returns Tweets from the last seven days that match a search query. In this study, the application provides the specified search keyword in the query and pulls all recent Tweets containing the keyword.

2) *Reddit*: Reddit is a popular social media platform, where the emphasis is on community rather than a single user [22]. In 2022, there was an average of 50 million daily active users. Recent Reddit statistics in April 2021 show that it was the tenth most-used social networking site in the United States [23] and the 15th most-used social platform worldwide [24]. It comprises a large number of communities known as subreddits. Each subreddit has a discussion board for a certain topic. Depending on their interests, users can create or subscribe to a variety of subreddits [22].

Social media platforms, such as Twitter, Instagram, and Meta have API restrictions that prevent applications from extracting data. However, using the Pushshift Reddit Dataset, social media researchers can easily query and analyze billions of submissions and comments on Reddit. "Pushshift is a social media data collection, analysis, and archiving platform that since 2015 has collected Reddit data and made it available to researchers". The Pushshift Reddit dataset is updated in real-time and provides an API to search, aggregate and perform exploratory analysis on the entirety of the dataset [25]. In this study, user comments are pulled from all subreddits within 30 days of the API call containing a specified search keyword.

3) *Tumblr*: Tumblr is a microblogging and social networking website founded by David Karp in 2007 that is currently owned by Automattic. The service allows users to post multimedia content to form short-form blogs known as tumblogs [26]. Tumblogging has not been used in current research and would be beneficial in obtaining a holistic understanding of the sentiment of people towards a particular topic.

The official Tumblr API provides a `/tagged` endpoint that allows the application to retrieve posts with a specific tag

[27]. Tags make it easier for readers to find posts about a specific topic on a user’s blog [28]. In this study, Tumblr posts containing tags with the specified search keyword were retrieved.

IV. SENTIMENT ANALYSIS OF SOCIAL MEDIA (SASM)

SASM is an open-source, cloud-based digital reputation management software solution. The application collects user posts from social media, performs sentiment analysis, and provides meaningful information to the end-user to aid them in evaluating and assessing their corporation’s, product’s, or service’s online reputation.

A. Requirements Engineering

The SASM system has been designed to meet specific requirements that aim to support marketing, reputation and branding management activities [29]. Figure 1 details the system use case diagram actors and main processes. Two key actors have been identified; (1) User and (2) Social-Media API. While the first actor interacts with SASM to specify the search term, the second actor interacts with SASM providing data feeds and retrieval functionalities. The goal behind considering the Social-Media API is to allow SASM to integrate independently and concurrently different social media channels.

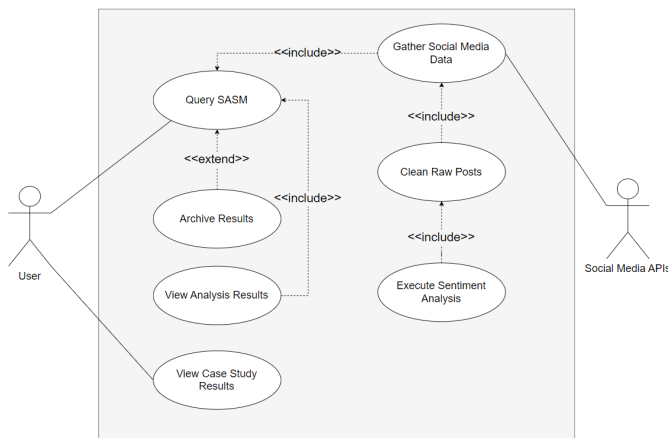


Figure 1: Sentiment Analysis of Social Media (SASM) Use Case Diagram

According to the software development life cycle and best practices, Requirements describe the characteristics that a system must have to meet the needs of the stakeholders. These requirements are typically divided into functional and non-functional requirements. Functional Requirements [FR] describe how software must behave and what are its features and functions [30]. Non-Functional Requirements [NFR] describe the general characteristics of a system [31] They are also known as quality attributes.

The following is a selection of functional requirements:

- **[FR1]** The system shall allow the user to input a search keyword (maximum 512 characters) into the search bar

and click on submit. The system will then display relevant graphs comparing and contrasting the public sentiment associated with the keyword on the aforementioned social media platforms;

- **[FR2]** The system shall allow the user to view the results of the searched keyword and configure the charts to extract the necessary information;
- **[FR3]** The system shall allow the user to view the results of the case study and configure the charts to extract the necessary information;
- **[FR4]** The system shall allow the user to view the top 10 posts for each social media platform in a tabular format.

The above-listed functional requirements have been analyzed and validated with stakeholders and the following set of quality attributes (non-functional requirements) has been derived:

- **[NFR1]** Availability: The system shall be available 24/7/365;
- **[NFR2]** Operability: the system shall be capable of communicating and retrieving data from common and well-established social-media platforms: Twitter [32], Reddit [33] and Tumblr [34].
- **[NFR3]** Storage: The system shall store the results of each unique searched keyword in separate data stores and therefore create a collection of archived data;
- **[NFR4]** Accessibility: The system shall be integrated into the Laboratory for Applied Software Engineering Research (LASER) website, and it should support all browsers. Additionally, all material presented by the system must meet the Web Content Accessibility Guidelines.

1) *Scenarios:* Figure 1 illustrates the use case diagram that outlines the necessary actions to fulfil the system requirements. Each user can take multiple paths within this use case. A scenario represents a specific path that a user takes while interacting with the system. It portrays a practical example of how the system is used by one or more users, outlining the steps, events, and actions that occur during the interaction. Usage scenarios can range from highly detailed, describing precisely how the user interacts with the interface, to moderately high-level, outlining the essential actions without specifying how they are performed. This section provides a detailed description of the usage scenarios depicted in the SASM system’s use case diagram.

Figure 2 detail the process of inputting a search term and triggering the Listener. Figures 3 and 4 illustrate the process of gathering social media posts and performing data cleaning to prepare for the sentiment analysis process, detailed in Figure 5. Figure 6 depicts a scenario where the user inputs a previously used search term, resulting in an archive data store containing both the current and past searches. Lastly, Figure 7 outlines the visualization of the sentiment analysis results for the search term, and Figure 8 specifies the visualization of the sentiment analysis results for the case study.

Use case Name: Query SASM
Preconditions: The user must navigate to the Home page
Main Sequence:
 1) User must input a search term/phrase
 2) User must click on the submit button
Outcome:
 1) Listener is triggered to Gather Social Media data

Figure 2: Query SASM Scenario

Use case Name: Gather Social Media Data
Preconditions: The user must query SASM
Main Sequence:
 1) Listener retrieves the search term
 2) Aggregator searches for posts related to the search term on Twitter, Reddit, and Tumblr
 3) Aggregator retrieves posts and forwards them to the Cleaning service
Outcome:
 1) Data for the search term is gathered.

Figure 3: Gather Social Media Data Scenario

Use case Name: Clean Raw Posts
Preconditions: A list of raw posts from Twitter, Reddit, and Tumblr, related to the search term, is available for processing
Main Sequence:
 1) Cleaner receives a list of raw posts and removes unneeded stopwords
 2) Cleaned posts is sent to Microsoft Text Analytics service, Google Cloud Natural Language service, and IBM Watson Natural Language Understanding service to obtain the sentiment value
Outcome:
 1) Gathered Social Media Data is cleaned

Figure 4: Clean Raw Posts Scenario

Use case Name: Execute Sentiment Analysis
Preconditions: List of clean posts has been sent to the sentiment analysis engines
Main Sequence:
 1) Engines produces a sentiment value for each of the posts
 2) Sentiment values of each post are appended to their corresponding documents in a MongoDB collection
Outcome:
 1) Each document in the MongoDB collection consists of the full post, cleaned post, social media platform, and the respective sentiment value from each engine.

Figure 5: Execute Sentiment Analysis Scenario

Use case Name: Archive Results
Preconditions: User specified a search term which was previously inputted
Main Sequence:
 1) Results from the current search are stored in the pre-existing collection creating an archived data store
 2) Results displayed to the user consist of the entire collection
Outcome:
 1) An archived data collection is created for the specified search term

Figure 6: Archive Results Scenario

Use case Name: View Analysis Results
Preconditions: The user has navigated to the Home page, inputted a search term and clicked on the submit button
Main Sequence:
 1) User is shown all analysis result reports that pertain to their search term
 2) User filters their search for specific information, i.e. social media platforms, sentiment analysis engines, and more
 3) Analysis results from the filter are displayed
Outcome:
 1) User has viewed the gathered analysis results

Figure 7: View Analysis Results Scenario

Use case Name: View Case Study Results
Preconditions: User navigated to the Results page
Main Sequence:
 1) User is shown all analysis result reports that pertain to the case study
 2) User filters their search for specific information, i.e. information technology companies, social media platforms, sentiment analysis engines, and more
 3) Analysis results from the filter are displayed
Outcome:
 1) User has viewed the case study analysis results

Figure 8: View Case Study Results Scenario

B. Software Architecture and Design

SASM is implemented using a client-server architecture. The client and server communicate using a RESTful API. The interactions between the client, server, and all the modules contained in the server are depicted in Figure 9. The following subsections detail their functionalities:

1) *Dashboard:* A user navigates to the Home Page in the SASM Dashboard, provides a search keyword in the input field, and clicks on the submit button. The search keyword is sent to the Listener Module via a HTTP POST request.

2) *Listener Module:* The Listener Module is bounded to an Azure Functions HTTP Trigger. The HTTP Trigger invokes the Listener Module when the HTTP POST request containing the search keyword is received. The search keyword is then passed off to the Aggregator Module.

3) *Aggregator Module:* The Aggregator Module is responsible for querying and retrieving data containing the search keyword from the three social media platforms, Twitter, Reddit, and Tumblr using their respective APIs. The collected data is published to an Azure Service Bus Topic: *General Cleaner*.

4) *Cleaner Module:* The Cleaner Module subscribes to the *General Cleaner* topic and fetches data from the Service Bus. The data is then cleaned as follows:

- *Data Preprocessing:* Process of removing usernames, hashtags, URLs, newlines, punctuation, numbers, and stop words [35]. Stop words are ubiquitous terms in writing such as 'the', 'and', 'I', and so on that do not give insights into the specific topic of a document. During data preprocessing, these stop words are removed from the text to identify words that are more infrequent and potentially more pertinent to the context [36]. Additionally, HTML links (URLs), usernames, hashtags, and embedding are removed as they do not add value to the data [35]. Data preprocessing is an integral step in ensuring the text is

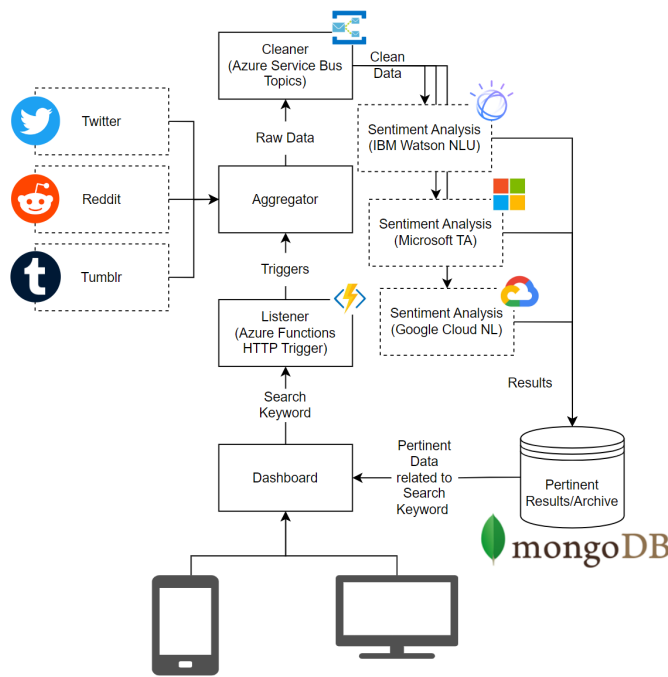


Figure 9: SASM Software Architectural Overview

cleaned and relevant data is extracted and formatted for analysis.

- **Tokenization:** Process of splitting text to individual words [35].
- **Lemmatization:** Process of collecting the inflected elements of a word so that they may be identified as a single unit, called the word’s lemma or its vocabulary form. The algorithm converts words to their basic form or root [37]. In Python programming, these tasks can be carried out by NLTK built-in library functions [38].

5) **Sentiment Analysis Module:** The Sentiment Analysis module receives a list of clean posts. Each post is sent to the individual sentiment analysis engines, MTA, GCNLA, and IWNLU, using their respective APIs. The engines then return the associated sentiment value which is converted to a number between -1 to +1:

- Values near negative 1 have a negative sentiment.
- Values near 0 have a neutral sentiment.
- Values near positive 1 have a positive sentiment.

The results are stored in a cloud-based MongoDB Atlas database [39], and the SASM dashboard retrieves the data and displays it using several infographics. The results from the dashboard can be seen in Figures 10, 11, and 12.

In Figure 10, a scatter plot displays the sentiment values for the most recent posts that were retrieved from Twitter containing the search keyword, *iPhone*. In the SASM dashboard, similar scatter plots for Reddit and Tumblr are generated as well to provide an individual and unique understanding of each social media platform. The plots provides insight into the overall sentiment of people on a particular social media platform towards the search keyword. Moreover, for a specific

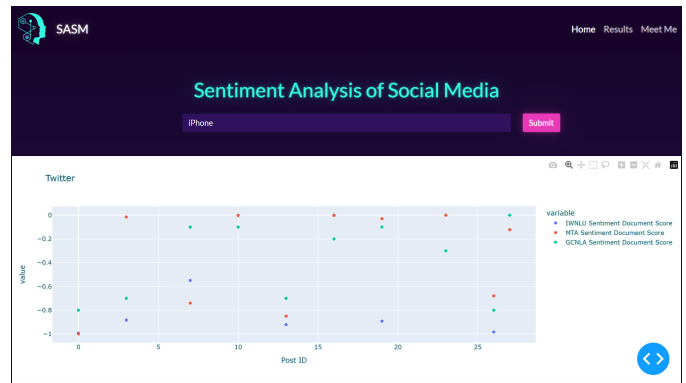


Figure 10: Home Page: Sentiment Analysis of Posts on Twitter

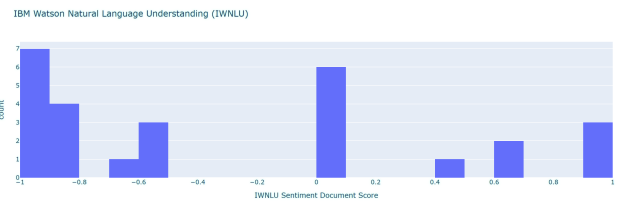


Figure 11: Frequency Distribution of the Sentiment Value for Posts analyzed by IWNLU

post, it allows the user to compare the sentiment value detected by each sentiment analysis engine: IWNLU, GCNLA, and MTA.

Figure 11 displays the frequency distribution of the sentiment values generated by the IWNLU engine. In the SASM dashboard, frequency distribution plots for GCNLA and MTA are also available for users’ to view. The count of the posts is cumulative and across all social media platforms.

Lastly, Figure 12 outlines the data collected in a tabular form. It consists of the post ID, full text of the post, social media platform, and the corresponding sentiment values from each engine. The user may filter the table based on a specific social media platform (Twitter, Reddit, and Tumblr) using the dropdown and display those particular posts.

Post ID	Post	Social Media Platform	IWNLU Sentiment Score	GCNLA Sentiment Score	MTA Sentiment Score
0	"I don't like my iPhone but my stupid iPhone changed it. This is extremely upsetting. I hate the AI inside my phone"	"Twitter"	-0.99336	-0.800000011920929	-1
1	"Don't get me started lol I had my first online master degree for VP just to be able and of course she got a private code. We were on the phone together the day of people while in the waiting room before the queues were long. She was"	"Reddit"	0.640958	-0.800000011920929	-0.97

Figure 12: Results for All Social Media Platforms in Tabular Form

V. RESULTS

To compare, contrast, and evaluate the performance of the three sentiment analysis engines, GCNLA, MTA, and IWNLU, we used a relevant case study. Our study focused on the mass layoffs in 2022-2023 from three major information technology companies - Google, Amazon, and Microsoft. Using a PowerShell Task Scheduler, posts were collected for 3 days, starting on January 31st 2023 to February 2nd, 2023. We used SASM to query and retrieve posts containing the following keywords: *google layoffs*, *amazon layoffs*, and *microsoft layoffs*. A total of 1607 posts were collected for the above mentioned keywords, but due to API limitations, more data could not be collected. 653 posts containing the keywords "google" and "layoffs"; 501 posts containing the keywords "amazon" and "layoffs"; and 453 posts containing the keywords "microsoft" and "layoffs" were collected from the three social media platforms. The collected data was then visualized using pie charts.

Each chart displays the percentage of posts that have sentiment values in agreement with each other (+- 0.1) for specific combinations of sentiment analysis engines. For instance, Figure 13 consists of 4 pie charts for the Google Layoffs collection. The first pie chart shows the percentage of posts collected from all social media platforms:

- 38.3% of the posts had sentiment values that were not within +- 0.1 of each other for all combinations of engines: MTA & GCNLA, MTA & IWNLU, IWNLU & GCNLA;
- 33.7% of the posts had sentiment values that were within +- 0.1 of each other for only MTA and IWNLU engines;
- 12.4% of the posts had sentiment values that were within +- 0.1 of each other for only MTA and GCNLA engines;
- 10.6% of the posts had sentiment values that were within +- 0.1 of each other for only GCNLA and IWNLU engines;
- 5.03% of the posts had sentiment values that were within +- 0.1 of each other for all combinations of engines: MTA & GCNLA, MTA & IWNLU, IWNLU & GCNLA.

The process was repeated for each social media platform, Twitter, Reddit, and Tumblr. Their results are displayed in the second, third, and fourth pie charts in Figure 13, respectively. This was done to evaluate and analyze trends across different social media platforms and it was observed that generally for pairs of sentiment analysis engines, MTA and IWNLU had the greatest percentage of posts where the sentiment values were in agreement (+-0.1 of each other). This was followed by MTA and GCNLA, and lastly, IWNLU and GCNLA.

It was also interesting to observe that the trend was apparent across all the search keywords. Figure 14 and Figure 15 display the corresponding pie charts for Amazon and Microsoft layoffs. It can be seen that the decreasing order of sentiment analysis engine pairs for the number of posts with similar sentiment values continue to be MTA & IWNLU, MTA & GCNLA, and IWNLU & GCNLA.

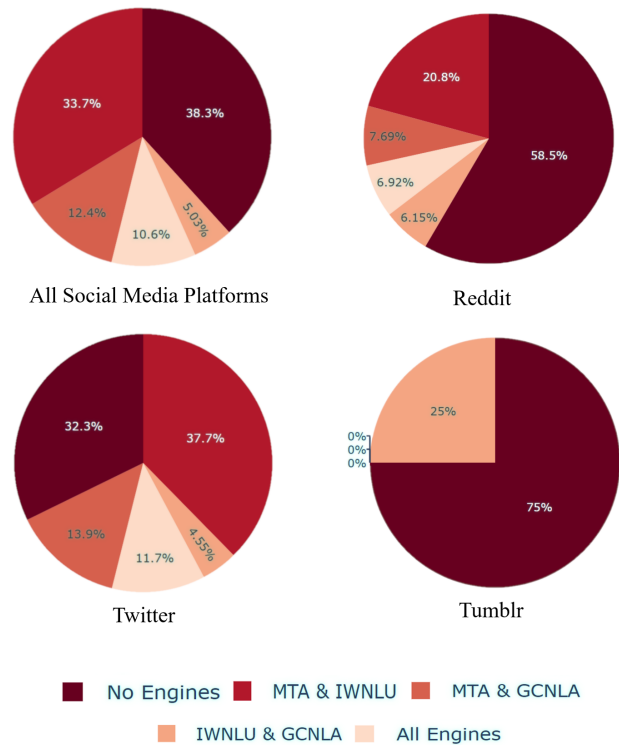


Figure 13: Google Layoffs

VI. DISCUSSION

The research project faced several limitations in the data collection and analysis process. One major limitation was that the MTA engine could not analyze more than 10 posts at a time limiting the total number of posts that could be analyzed in each iteration. Additionally, Tumblr, which was included as a data source, was found to be an unreliable source of information due to it being image-based and the limitation in the number of posts that were available related to technology and layoffs. Furthermore, we were unable to verify the sentiment analysis engine’s results and had to trust proprietary tools for analysis. We were only able to visually verify the results by observing the proximity of the dots in the scatter plots presented in the Home Page.

Another limitation was the inability to include spatiotemporal analysis in the scope of the project. This was due to the fact that while Twitter returned spatiotemporal data; Tumblr and Reddit did not. Since data was being collected from three different channels, extracting spatiotemporal data for all three was not feasible. Moreover, the application had a limitation in search accuracy, as it sometimes returned posts that were not related to the search query.

The project also encountered limitations in collecting data from other languages due to the occurrence of Unicode errors, which resulted in the application ignoring those posts. This highlights the need for a better application that takes multi-language into account for effective data collection.

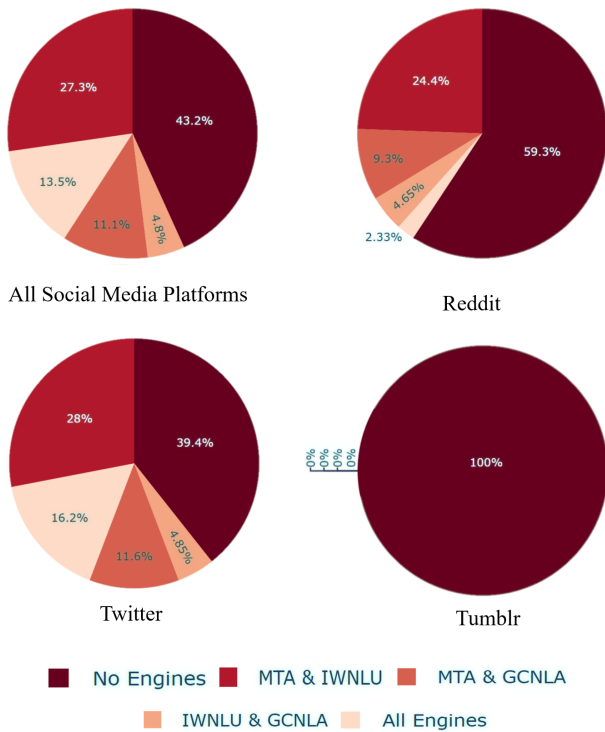


Figure 14: Amazon Layoffs

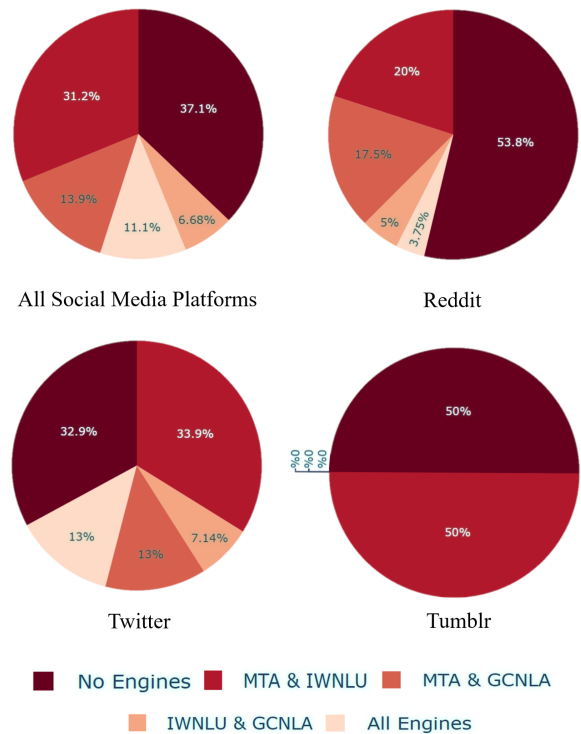


Figure 15: Microsoft Layoffs

Moreover, when the data is limited, these platforms might underperform because their models are typically trained on large, varied corpora and require diverse inputs to generalize well. A small data set may lack the nuance or range of expressions needed to test the models' abilities to handle different tones, contexts, or subtleties in sentiment.

In addition, sentiment analysis engines also have their limitations in analyzing data from different languages, with some engines supporting more languages than others. The 512-character search keyword limitation also poses a challenge in effectively searching for relevant data, as it may not be possible to include all relevant keywords within the character limit.

Moreover, the amount of data that can be pulled from the social media platform is limited by the API calls, which can also affect the quality of the data collected. There is also a limitation in the amount of data that can be passed to the sentiment analysis engines, which can affect the accuracy of the analysis.

VII. CONCLUSION AND FUTURE WORK

In this paper, we introduced Sentiment Analysis of Social Media, an open-source, multi-channel, multi-engine sentiment analysis software for social media and digital reputation management purposes. SASM collects user posts from three social media platforms, performs sentiment analysis using three analytic engines, and provides meaningful information to the end-user to aid them in evaluating and assessing their

corporation's, product's, or service's online reputation. To verify and validate the effectiveness of SASM, a case study was conducted to collect data about layoffs from Google, Amazon, and Microsoft from Twitter, Reddit, and Tumblr. The goal was to study the feasibility of developing a multi-channel & multi-engine platform and potentially comparing, contrasting, and evaluating the performance of the three sentiment analysis engines: GCNLA, IWNLU, and MTA.

Future research should focus on developing better applications that can effectively collect data from multiple languages. Efforts should be made to overcome the limitations of API calls to enhance the quality and quantity of data collected for sentiment analysis. Moreover, additional efforts will focus on expanding data collection to ensure a more comprehensive dataset. Larger datasets will support more fine-grained comparisons between different sentiment analysis engines will be essential to evaluate their performance across various contexts and use cases, allowing for more accurate and nuanced insights. Furthermore, performing a spatiotemporal analysis of the collected data would be interesting as it would allow companies to evaluate their digital presence in different locations and dedicate resources accordingly. Lastly, using a different and diverse social media platform such as Meta should be considered for data collection.

REFERENCES

[1] A. J. Nair, G. Veena, and A. Vinayak, "Comparative study of twitter sentiment on covid-19 tweets," in *2021 5th Interna-*

- tional conference on computing methodologies and communication (ICCMC), IEEE, 2021, pp. 1773–1778.
- [2] R. Panikar, R. Bhavsar, and B. Pawar, “Sentiment analysis: A cognitive perspective,” in *2022 8th International Conference on advanced computing and communication systems (ICACCS)*, IEEE, vol. 1, 2022, pp. 1258–1262.
 - [3] F. Samara, S. Ondieki, A. M. Hossain, and M. Mekni, “Online social network interactions (osni): A novel online reputation management solution,” in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, IEEE, 2021, pp. 1–6.
 - [4] C. Haruechaiyasak, A. Kongthon, P. Palingoon, and K. Trakultaweekoon, “S-sense: A sentiment analysis framework for social media sensing,” in *Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, 2013, pp. 6–13.
 - [5] F. Neri, C. Aliprandi, F. Capeci, and M. Cuadros, “Sentiment analysis on social media,” in *2012 IEEE/ACM international conference on advances in social networks analysis and mining*, IEEE, 2012, pp. 919–926.
 - [6] C. R. Barone and M. Mekni, “Advancing continuous authentication using smart real-time user activity fingerprinting,” in *2023 IEEE Third International Conference on Signal, Control and Communication (SCC)*, IEEE, 2023, pp. 01–06.
 - [7] M. Trupthi, S. Pabboju, and G. Narasimha, “Sentiment analysis on twitter using streaming api,” in *2017 IEEE 7th international advance computing conference (IACC)*, IEEE, 2017, pp. 915–919.
 - [8] K. Ali, H. Dong, A. Bouguettaya, A. Erradi, and R. Hadjidj, “Sentiment analysis as a service: A social media based sentiment analysis framework,” in *2017 IEEE international conference on web services (ICWS)*, IEEE, 2017, pp. 660–667.
 - [9] B. J. G. Praciano *et al.*, “Spatio-temporal trend analysis of the brazilian elections based on twitter data,” in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE, 2018, pp. 1355–1360.
 - [10] J. J. Watkins, “‘i’m especially sorry to those impacted’: Conventional visions of leadership in the 2022/2023 tech layoffs,” 2024.
 - [11] Z. Saba, “Layoffs and corporate performance: Evidence based on the us tech industry,” *Journal of Economics and Finance*, pp. 1–24, 2024.
 - [12] K. Vedantam, “Tech layoffs: Us companies with job cuts in 2022 and 2023,” [Online], Crunchbase News, Apr. 2023, [Online]. Available: <https://news.crunchbase.com/startups/tech-layoffs/>.
 - [13] A. Satariano and N. Grant, “Google parent alphabet to cut 12,000 jobs,” *The New York Times*, 2023, Online; accessed 20-Jan-2023.
 - [14] S. Pichai, *A difficult decision to set us up for the future*, Google, [Online]. Available: <https://blog.google/inside-google/message-ceo/january-update/>, Jan. 2023.
 - [15] J. Peters, “Amazon confirms cuts to hardware and services teams,” *The Verge*, Nov. 2022, [Online]. Available: <https://www.theverge.com/2022/11/16/23462439/amazon-layoffs-cuts-hardware-services-teams>.
 - [16] M. Clark, “Amazon begins another round of job cuts as it lays off more than 18,000 people,” *The Verge*, Jan. 2023, [Online].
 - [17] C. Thorbecke and H. Ziady, “Microsoft is laying off 10,000 employees,” *CNN*, Jan. 2023, [Online].
 - [18] T. Warren, “Microsoft announces big layoffs that will affect 10,000 employees,” *The Verge*, Jan. 2023, [Online]. Available: <https://www.theverge.com/2023/1/18/23560315/microsoft-job-cuts-layoffs-2023-tech>.
 - [19] Q.ai, “What microsoft’s recent layoffs mean for the company and investors,” *Forbes*, Jan. 2023, [Online].
 - [20] T. V. S. Krishna *et al.*, “A novel ensemble approach for twitter sentiment classification with ml and lstm algorithms for real-time tweets analysis,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 3, pp. 1904–1914, 2024.
 - [21] Y. Lheureux, “Predictive insights: Leveraging twitter sentiments and machine learning for environmental, social and governance controversy prediction,” *Journal of Computational Social Science*, vol. 7, no. 1, pp. 23–44, 2024.
 - [22] S. Nyawa, D. Tchuente, and S. Fosso-Wamba, “Covid-19 vaccine hesitancy: A social media analysis using deep learning,” *Annals of operations research*, vol. 339, no. 1, pp. 477–515, 2024.
 - [23] J. Gottfried, “Americans’ social media use,” *Pew Research Center*, 2024.
 - [24] E. A. Radday, M. Mekni, L. Page, A. Sula, and L. Brown, “Evolving cybersecurity education: An analysis of the gen-cyber teacher academy’s progression from 2022 to 2023 and beyond,” *Journal of Computing Sciences in Colleges*, vol. 39, no. 8, pp. 113–127, 2024.
 - [25] S. Rani, K. Ahmed, and S. Subramani, “From posts to knowledge: Annotating a pandemic-era reddit dataset to navigate mental health narratives,” *Applied Sciences*, vol. 14, no. 4, p. 1547, 2024.
 - [26] Q. Shi, W. Xu, and Z. Miao, “Image-text multimodal classification via cross-attention contextual transformer with modality-collaborative learning,” *Journal of Electronic Imaging*, vol. 33, no. 4, pp. 043 042–043 042, 2024.
 - [27] Tumblr, *Tumblr api*, <https://www.tumblr.com/docs/en/api/v2>, [Online; accessed 22-Oct-2024].
 - [28] *Tagging your posts – help center*, <https://help.tumblr.com/hc/en-us/articles/226161387-Tagging-your-posts>, [Online], n.d.
 - [29] W. D. Schindel, “Requirements statements are transfer functions: An insight from model-based systems engineering,” *INSIGHT*, vol. 27, no. 5, pp. 27–34, 2024.
 - [30] J. J. Shah and M. T. Rogers, “Functional requirements and conceptual design of the feature-based modelling system,” *Computer-Aided Engineering Journal*, vol. 5, no. 1, pp. 9–15, 1988.
 - [31] M. Glinz, “On non-functional requirements,” in *15th IEEE International Requirements Engineering Conference (RE 2007)*, IEEE, 2007, pp. 21–26. DOI: 10.1109/RE.2007.45.
 - [32] Twitter, *Twitter*, <https://twitter.com/>, [Online; accessed 22-Oct-2024], 2023.
 - [33] Reddit Inc, *Reddit inc*, <https://www.reddit.com/>, [Online; accessed October 2023], 2023.
 - [34] Tumblr, *Tumblr*, <https://tumblr.com/>, [Online], 2023.
 - [35] M. Mekni *et al.*, “Smart detection for heart health,” in *Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence*, 2024, pp. 117–122.
 - [36] M. Mekni and D. Haynes, “Smart community health: A comprehensive community resource recommendation platform,” in *HEALTHINF 2020 - 13th International Conference on Health Informatics, Proceedings; Part of 13th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2020*, SciTePress, vol. 11, 2020, pp. 614–624.
 - [37] D. Khyani and B. S. Siddhartha, “An interpretation of lemmatization and stemming in natural language processing,” *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology*, vol. 22, pp. 350–357, 2021.
 - [38] NLTK, *Nltk :: Natural language toolkit*, <https://www.nltk.org/index.html>, Accessed: 2024-10-22.
 - [39] MongoDB Inc, *Mongodb atlas*, <https://www.mongodb.com/>, [Online; accessed 22-Oct-2024], 2023.

Tourist Mobility Forecasting with Region-based Flows and Regular Trips

Fernando Terroso-Saenz 

Technical University of Cartagena
Cartagena, Spain

e-mail: fernando.terroso@upct.es

Juan Morales-García 

Universidad Católica de Murcia (UCAM)
Murcia, Spain

e-mail: jmorales8@ucam.edu

Miguel Puig 

Universidad Católica de Murcia (UCAM)
Murcia, Spain

e-mail: mpuig@ucam.edu

Ginesa Martínez-del Vas 

Universidad Católica de Murcia (UCAM)
Murcia, Spain

e-mail: gmvas@ucam.edu

Andres Muñoz 

University of Cadiz
Cadiz, Spain

e-mail: andres.munoz@uca.es

Abstract—One of the most prominent courses of action in the tourist sector is the development of predictors to anticipate the flow of incoming and outgoing tourists of a region. To do so, most of the existing approaches usually take tourist-related flows as the only primary input to perform the prediction. The present work assesses the suitability of composing a deep-learning predictor that fuses touristic displacements with data extracted from a general-purpose human-mobility dataset. The proposal has been tested in the Region of Murcia, a Spanish administrative area with a lively tourist sector. Results show that our approach achieves up to 46% Root Mean Square Error (RMSE) reduction with respect to a baseline only relying on tourist data.

Keywords—tourist mobility ; deep neural networks , human mobility flows ; time series forecasting

I. INTRODUCTION

The tourism sector has undergone extensive research aimed at devising intelligent solutions to enhance both business processes and customer experiences in multiple platforms, such as online social networks [1]. This integration has facilitated the analysis of tourist mobility behavior. Consequently, forecasting tourists' flows holds significant implications in areas such as tourism marketing or services, empowering tourism institutions and stakeholders to more efficiently manage their resources [2][3].

However, the advancement of prediction algorithms to forecast tourist flows (e.g., the volume of incoming or outgoing tourist trips within a geographical area such as a city) typically depends on a *univariate* approach, where the target flow serves as the primary input for the predictor [4][5]. Yet, the utilization of alternative forms of human mobility data as *exogenous* variables to enhance prediction accuracy has not been thoroughly investigated.

The primary objective of this study is to evaluate the viability of enhancing a tourist-flow prediction model with human movement data obtained from sources that document regular and daily movements within a specific geographical area. This concept is rooted in the notion that daily human movements towards a region could offer an alternative yet supplementary perspective on its tourist flows. For instance, a significant increase in inbound tourists to a city attributed to a social event might be preceded by a decrease in commuter journeys towards that region several hours before the event begins. Anticipating

such a decline could be leveraged by a predictive model to enhance the accuracy of forecasting future tourist visitation patterns.

In order to evaluate our approach, we have used several instances of a model comprising a stack of convolutional and recurrent neural network layers for time series forecasting in order to anticipate different types of touristic flows towards the Region of Murcia (RM), a Spanish Administrative area in Spain with an active tourist industry. In that sense, the predictor is feed with different subflows extracted from an open nationwide human-mobility dataset to anticipate the overall number of incoming tourists towards RM several weeks ahead.

The salient contribution of this work is the fusion of different datasets that allows the development of a touristic mobility predictor that merges the regular and tourist movement of a geographical region so as to forecast its incoming touristic flow. The key benefit of this *multi-flow* approach is that it allows a much more accurate estimation of the tourists arriving to a region than an approach solely relying on tourist-related flows for several time horizons.

The remainder of the paper is structured as follows. Section 2 summarizes the most relevant current approaches for human mobility prediction in the touristic sector. Then, Section 3 describes the use-case setting of our study. In Section 4, the most important results of the deployed palette of predictors are described and evaluated. Lastly, Section 5 summarizes the main conclusions and potential future research lines motivated by this work.

II. RELATED WORK

In recent years, there has been a large interest in harnessing methodologies to fuse data from heterogeneous sources so as to forecast tourist movements, thus enhancing tourism planning and management through the utilization of multivariate datasets. One approach has been based on the usage of tourist flows from one region to improve the prediction accuracy in another area. For instance, Zhu et al. [6] examined tourist flows from six countries to forecast tourist arrivals in Singapore, proposing a pairwise modeling approach to account for interdependence among countries within the same geographic region. Analyzing

flows from 1995 to 2013 and predicting up to 20 quarters ahead, they demonstrated improvements in Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) by incorporating pairwise flows compared to models treating flows independently, particularly evident in annual predictions. A similar approach was followed by Yang et al. [7] investigated the combination of spatial and temporal tourist flow datasets in China, contrasting univariate models like ARIMA with multivariate space-time autoregressive moving average (STARMA) models. Their study, spanning tourist mobility data from 29 Chinese regions between 1987 and 2016, showcased enhanced accuracy of STARMA models, especially in neighboring regions with strong spatial correlations.

Another line of research has focused on the integration of datasources not directly related to human mobility to enhance the prediction performance. As a matter of fact, Zhang et al. [8] integrated tourism flow volumes with various Search Intensity Indicators (SII) from Google Trends to refine the accuracy of Machine Learning and Deep Learning models in forecasting tourist arrivals in Hong Kong over different time horizons. A similar approach was adopted for predicting tourist demand in Macau, China, by Law et al. [9], who evaluated diverse Deep Learning models, particularly those employing attention mechanisms, surpassing conventional ML techniques like Support Vector Regression. Besides, De-Jesus et al. [10] made use of data reporting the evolution of the COVID-19 pandemic in the Philippines to enhance the prediction of inbound tourist to that country confirming that integrating such type of data improved the model accuracy.

The novel aspect of our current work lies in the nature of data used to enhance tourist mobility prediction. Unlike previous approaches relying on web-based indicators, or COVID-19 data as exogenous variables, we leverage direct human movement data extracted from an open and general-purpose feed specific to the geographical area of interest.

III. USE-CASE SETTING

The focal point of our investigation has been the Region of Murcia (RM), an autonomous community in Spain situated in the southeast of the country (see Figure 1). This area boasts a population of approximately 1.5 million people and covers an expanse of 11,313 km². In terms of tourism, this region welcomed over 1,300,000 visitors in 2022, marking a 45% increase compared to 2021 [11].

A. Datasets

We utilized two distinct mobility datasets, the first encompasses tourist mobility within the Region of Murcia (RM), while the second encompasses total human mobility within the same region. This way, we had two different views of how people moved around the target region.

1) *Tourist Mobility Dataset (TMD)*: The flow of tourists in RM was gathered through the Tourist Mobility Dataset (TMD) provided by the Tourism Institute of the Region of Murcia [12] as part of its Smart Region project. This dataset captures the inbound and outbound movement of tourists in

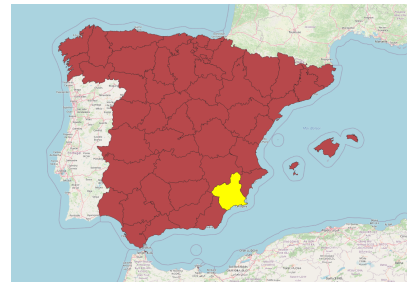


Figure 1. Location of the Region of Murcia (in yellow) with respect to the rest of autonomous communities in Spain depicted in red.

RM over a 16-month period spanning from January 1st, 2022, to April 30th, 2023. These flows are derived from the network events generated by mobile phones connected to the Telefonica network, one of the leading carriers in Spain [13]. The data undergoes anonymization, extrapolation, and aggregation stages to compute the final tourist flows included in the dataset.

An important aspect of this dataset is its distinction between the incoming flow of national (residing in Spain) and international (arriving from other countries) tourists (\mathcal{NT} and \mathcal{IT} , respectively), as well as excursionists (\mathcal{NE} and \mathcal{IE} , respectively). The former category comprises individuals who spend at least one night in the region (e.g., staying in a hotel, camping, or tourist accommodation), while the latter encompasses day trippers who visit RM for the day but do not spend the night away from their primary residence.

The format of the dataset defines the flows on a weekly basis. In this manner, the mobility of the w -th week of the year y for a city c is defined as a single tuple,

$$\langle y, w, c, f_{work}^m, f_{work}^a, f_{work}^n, f_{end}^m, f_{end}^a, f_{end}^n \rangle$$

where f_{work}^m , f_{work}^a and f_{work}^n are the *week slices* comprising the overall incoming flows towards c during the morning (m), afternoon (a) and night (n), respectively, considering all the working days of the w -th week. Similarly, f_{end}^m , f_{end}^a and f_{end}^n provide the same time-sliced flows for the weekend days (Saturday and Sunday in Spain). Thus, for each combination of year, week and city (y, w, c) the dataset comprises 4 different tuples, one for each type of touristic flow, 1) national excursionists \mathcal{NE} , 2) national tourists \mathcal{NT} , 3) international excursionists \mathcal{IE} and 4) international tourists \mathcal{IT} . For the sake of clarity, Figure 2 shows the number of incoming tourists and excursionists (regardless its origin) for the 70 weeks covered by the dataset.

Given the aforementioned flows, we computed 3 aggregated ones comprising the overall number of tourists \mathcal{T} ($= \mathcal{NT} + \mathcal{IT}$), the overall number of excursionists \mathcal{E} ($= \mathcal{NE} + \mathcal{IE}$) and the overall number of visitors \mathcal{A} ($= \mathcal{T} + \mathcal{E}$).

Next, we composed a timeseries for each flow $\mathcal{F} \in \langle \mathcal{NE}, \mathcal{NT}, \mathcal{IE}, \mathcal{IT}, \mathcal{T}, \mathcal{E}, \mathcal{A} \rangle$ covering the 70 weeks under study with the following format $\mathcal{F}_{TM} = \langle f_{work}^{m,1} \rightarrow f_{work}^{a,1} \rightarrow f_{work}^{n,1} \rightarrow f_{end}^{m,1} \rightarrow f_{end}^{a,1} \rightarrow f_{end}^{n,1} \rightarrow f_{work}^{m,2} \rightarrow f_{work}^{a,2} \rightarrow f_{work}^{n,2} \rightarrow f_{end}^{m,2} \rightarrow f_{end}^{a,2} \rightarrow f_{end}^{n,2} \rightarrow \dots \rightarrow f_{work}^{m,70} \rightarrow f_{work}^{a,70} \rightarrow f_{work}^{n,70} \rightarrow f_{end}^{m,70} \rightarrow f_{end}^{a,70} \rightarrow f_{end}^{n,70} \rangle$

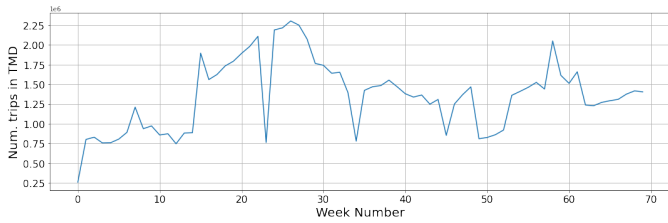


Figure 2. General flow of incoming tourists and excursionists to the Region of Murcia during the period of study considering the tourist mobility dataset.

$f_{work}^{m,70} \rightarrow f_{end}^{m,70} \rightarrow f_{end}^{a,70} \rightarrow f_{end}^{n,70}$) where, for example, $f_{work}^{m,i}$ is the record comprising the overall value of the \mathcal{F} flow during the working days' mornings of the i -th week.

2) *General Human Mobility Dataset (GMD)*: This dataset was obtained from the nationwide human mobility report published by the Spanish Ministry of Transportation (SMT) in January 2022 [14]. It provides information on the number of trips per hour between 2,735 cities across Spain, covering both the mainland and insular regions. This dataset can be viewed as a collection of tuples, each taking the form,

$$\langle date, hour, m_{origin}, m_{dest}, n_{trp}, dist \rangle$$

reporting that there was n_{trp} human trips from the city m_{origin} to the city m_{dest} and whose distance was $dist$ km during the indicated $date$ and $hour$.

As per official reports [15], these mobility data were derived from Call Detail Records (CDRs) of 13 million users from an undisclosed mobile carrier. After anonymization, the dataset was utilized to extrapolate comprehensive mobility patterns representative of the Spanish population at a national scale, subsequently released as open data. It is important to note that this dataset encompasses the movements of individuals irrespective of their mode of transportation.

Utilizing this dataset, we filtered its flows by retaining records that satisfy the following two criteria: (1) their destination m_{dest} is one of the cities in RM, and (2) their distance $dist$ exceeds a certain threshold δ . The first criterion captures the inbound flows to RM, while the second criterion refines these inbound flows based on the distance traveled by visitors to reach RM. Since the threshold δ defines the minimum distance, we avoid the fact that some of the resulting flows include almost all the records of the initial dataset.

Considering the peninsular area of Spain has an approximate radius of 540km and commuters' average travel distances range from 19 to 34 km [16], we employ three distinct δ values: 100, 400, and 800 km. Using these thresholds, we initially derived a subset of short-distance trips with $\delta = 100km$ (\mathcal{F}_{GM}^{100}), aiming to capture regular and non-touristic trips alongside various types of tourist flows to RM. Subsequently, we constructed a second subset representing medium-distance trips with $\delta = 400km$ (\mathcal{F}_{GM}^{400}) and a third subset encompassing long-distance travelers with $\delta = 800km$ (\mathcal{F}_{GM}^{800}). This approach allowed us to progressively filter out the proportion of regular and non-touristic trips in each subset by increasing the value of δ . In that sense, each subflow is defined as a timestamped sequence $\mathcal{F}_{GM}^{\delta} = \langle f_{gm,\delta,work}^{m,1} \rightarrow f_{gm,\delta,work}^{a,1} \rightarrow f_{gm,\delta,work}^{n,1} \rightarrow$

$f_{gm,\delta,end}^{m,1} \rightarrow f_{gm,\delta,end}^{a,1} \rightarrow f_{gm,\delta,end}^{n,1} \rightarrow f_{gm,\delta,work}^{m,2} \rightarrow f_{gm,\delta,work}^{a,2} \rightarrow f_{gm,\delta,work}^{n,2} \rightarrow f_{gm,\delta,end}^{m,2} \rightarrow f_{gm,\delta,end}^{a,2} \rightarrow f_{gm,\delta,end}^{n,2} \rightarrow \dots \rightarrow f_{gm,\delta,work}^{m,70} \rightarrow f_{gm,\delta,work}^{a,70} \rightarrow f_{gm,\delta,work}^{n,70} \rightarrow f_{gm,\delta,work}^{m,70} \rightarrow f_{gm,\delta,work}^{a,70} \rightarrow f_{gm,\delta,work}^{n,70} \rangle$ where, for example, $f_{gm,\delta,work}^{m,i}$ is the record comprising the overall value of the trips towards RM, covering a distance of at least δ km, during the morning of the working days of the i -th week according to the GMD.

Figure 3 shows the time series of the aforementioned subflows in RM during the same time period considered for the touristic dataset (2022/01/01-2023/04/30). As observed, the 3 time series have a very different order of magnitude. Furthermore, \mathcal{F}_{GM}^{100} exhibits a quite *flat* pattern throughout the whole period of study capturing a mostly-stationary travel behaviour. This is consistent with the fact this GMD flow is the one that captures more regular and commuting trips from the three ones. On the contrary, flows \mathcal{F}_{GM}^{400} and \mathcal{F}_{GM}^{800} comprised sharper peaks during the different holiday seasons covered in the study. For example, Figure 3 shows a large increment of incoming trips in both flows during the two Easter holidays under consideration. This reveals that these two timeseries captured more *seasonal* travel patterns and, thus, more compatible with tourist-related displacements.

As we can see, each GMD subflow provides different point of view of the human mobility in the Region of Murcia so that they can be used to study whether the usage of general human mobility might improve the prediction of a particular flow of visitors. It is important to remark that the GMD and SMT represent different mobility flows. However, some redundancy might occur between both datasets as the GMD might comprise a certain number of touristic trips.

IV. DESCRIPTION OF THE PREDICTOR

The focus of this paper lies on addressing the tourist mobility prediction challenge, which can be framed as a regression problem:

Given the weekly time slice w , the number of incoming tourists and/or excursionists over the past w_{prev} time slices according to the TMD $\mathcal{F}_{TM}^w = \langle f^w, f^{w-1}, \dots, f^{w-w_{prev}}, \mathcal{S} \rangle$, and the number of incoming trips based on the GMD within a distance-threshold δ for the same time lags $\mathcal{F}_{GM}^{\delta,w} = \langle f_{gm,\delta}^w, f_{gm,\delta}^{w-1}, \dots, f_{gm,\delta}^{w-w_{prev}} \rangle$, **Determine** a mapping function \mathcal{P} :

$$\mathcal{P}(\mathcal{F}_{TM}^w, \mathcal{F}_{GM}^{\delta,w}) \rightarrow \mathcal{F}_{TM}^{w+T}$$

where \mathcal{F}_{TM}^{w+T} represents the estimated number of tourists and/or excursionists arriving in RM at the $(w+T)$ th week slice as per the TMD study, with T denoting the prediction time horizon ($T \geq 1$). Notably, the novelty in this predictive model lies in its integration of GMD-based trips, supplementing the information derived solely from the TMD source.

To implement the predictor model, we have used a combination of a Convolutional Neural Network (CNN) with a Long short-term model (LSTM), resulting in a CNNLSTM model. As Figure 4 shows, this model firstly compresses

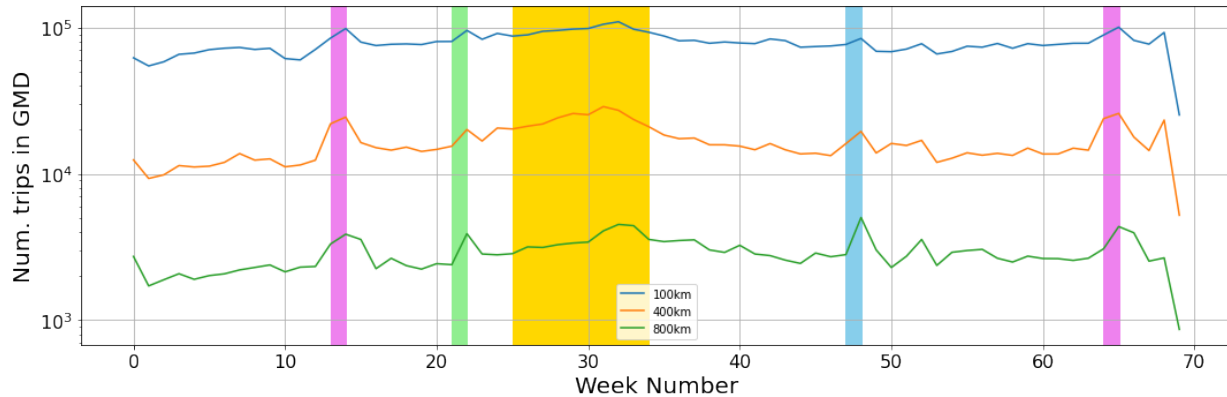


Figure 3. General flow of incoming tourists and excursionists to the Region of Murcia, considering different distance-based filtering (δ) during the period of study considering the general human mobility dataset. The violet areas represent the Easter holidays in 2022 and 2023 respectively, the green and blue ones nation bank holidays and the yellow area the summer period (July and August) in 2022.

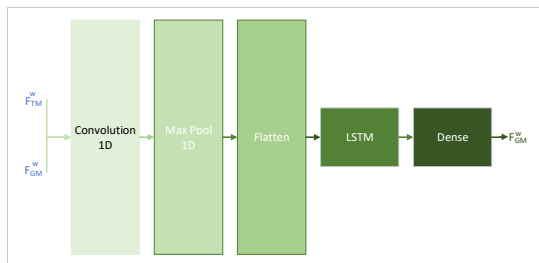


Figure 4. Layer architecture of the CNNLSTM applied in the study.

and extracts the relevant features of the incoming bi-variate timeseries comprising \mathcal{F}_{TM}^w , $\mathcal{F}_{GM}^{\delta,w}$ flows by means of a one-dimensional convolutional and a max-pool layer. Then, the resulting sequence is *flattened* to a 1D vector in order to be processed by the downstream LSTM and dense layers and generate the estimated \mathcal{F}_{TM}^{w+T} flow. In that sense, we opted to use a CNN instead of applying feature selection prior to an LSTM because CNNs are capable of efficiently capturing spatial and local patterns in time series data, which helps identify complex relationships between the data before being processed by the LSTM. This approach enables better extraction of relevant features directly from the time series, eliminating the need for prior manual feature selection.

V. EVALUATION OF THE PREDICTOR

In this section we evaluate the accuracy of the CNNLSTM predictor described in the previous section.

A. Metrics

In terms of evaluating the CNNLSTM model, the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) [17] stand out as two widely utilized metrics for assessing accuracy in predicting continuous variables. These metrics are well-suited for comparing models as they quantify the average prediction error of the model in the units of the variable of interest. Their definitions are as follows,

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

where, for our experiment, y_i is the real number of touristic trips, \hat{y}_i is the predicted number of trips and n is the number of observations. Furthermore, we complement these metrics with the Mean Average Prediction Error (MAPE) metric that is calculated as follows:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100.$$

B. Model Hyper-parameters

Table I comprises the configuration of the hyper-parameters applied in the CNNLSTM model.

C. Evaluation Results

Table II shows the metric values of the CNNLSTM model for different combination of inputs. In order to properly evaluate the impact of our approach, the table also shows the results of a baseline model that is only fed with touristic mobility data (\mathcal{F}_{TM}). Bearing in mind the description of the predictor stated in Section IV, this baseline can be defined as a univariate function $\mathcal{P}(\mathcal{F}_{TM}^w) \rightarrow \mathcal{F}_{TM}^{w+T}$.

As observed in Table II, at least one of the models enriched with general mobility data outperformed the univariate alternative for almost all the target flows. For example, in order to predict the overall flow of visitors to RM (\mathcal{A}), the predictor fed with the GMD flow with a 800km distance threshold (\mathcal{A} , \mathcal{F}_{GM}^{800}) reduced the MAE from 30,678.195 to 29,311.016 (-5%). A higher improvement is observed in the case of the national excursionists (\mathcal{NE}), in which the RMSE dropped from 15,732.939 in the univariate version to 8,544.438 (-46%) in the model using the \mathcal{F}_{GM}^{800} flow. Concerning the \mathcal{F}_{GM}^{400} flow, it

TABLE I
HYPER-PARAMETERS OF THE CNNLSTM MODEL

Hyperparameter	Description	Value
Filters	Features detector	64
Kernel size	Filters matrix used to extract the features from the dataset	2
Strides	Number of timesteps shifts over the input sequence	4
Activation function	Function that decide if a neuron should be (or not) activated	Tanh
Batch size	Size of batch used for training/forecasting	32
Epochs (+ <i>EarlyStopping</i>)	Number of epochs used in training	15000
Optimizer	Function that optimises the learning of a artificial intelligence model	Adam
Loss function	Function used for evaluate the error of the model in each epoch	MSE
Learning rate (+ <i>ReduceLRonPlateau</i>)	Percentage change with which weights are updated at each iteration	0.003
Train-test split	Rate of the dataset used to train and evaluate the models	90% (train), 10% (test)

allowed to improve the accuracy of the prediction of the \mathcal{NE} flow (MAPE from 35.145 to 16.443, -54%) and the \mathcal{T} flow (MAE from 16,909.334 to 15,519.060, -9%).

An important finding of this evaluation is that the most suitable distance threshold δ to compose the GMD flow varies depending on the target tourist flow. This makes sense as the nature of each target flow is quite different. For example, a 400km- δ provided a higher accuracy for the national flows, \mathcal{NE} and \mathcal{NT} , whereas the 800km provided the best results for the international-tourist flow (\mathcal{IT}). Moreover, the 400km threshold was the best configuration to predict the overall touristic flow (\mathcal{T}) but the 800km one allowed the best accuracy for the overall excursionist (\mathcal{E}). This dichotomy of distances for tourists and excursionists explain why the most accurate model to predict the global flow \mathcal{A} depends on the evaluation metric under consideration as we can see in the first 4 rows of Table II.

It is important to remark that the predictor actually improved its accuracy when it incorporated the regular flows filtered with δ equals to 400 or 800km. However, the \mathcal{F}_{GM}^{100} did not provide a clear improvement for any of target flows. In that sense, \mathcal{F}_{GM}^{400} and \mathcal{F}_{GM}^{800} were the two flows exhibiting a higher seasonality with peaks at certain holiday periods revealing that the *weight* of the touristic displacements was quite high in such flows (Section III-A2). This suggests that the actual improvement of the predictor occurs when it is enriched with an exogenous flow comprising *latent* touristic displacements in a quite strong manner. That is, when it provides an *alternative* view of the touristic flows of RM discarding the regular displacements at high degree.

Furthermore, Figure 5 shows the RMSE of each model for each target flow and different values of time horizons T , namely 6, 12, 18, 24, 32 and 48 week slices. Since the TMD defines 6 slices per week (Section III-A1), such time horizons can be also regarded as 1, 2, 3, 4, 5.3 and 8 weeks. As observed, the major improvement of the CNNLSTM with GMD flows usually occurred for large time horizons above 24 slices (4 weeks). This is specially noticeable in the \mathcal{A} (Figure 5a), \mathcal{E} (Figure 5b) and \mathcal{NE} (Figure 5d) flows. It is also worth mentioning the fact that the model enriched with $\mathcal{F}_{GM,800}$ clearly outperformed the other models for all the time horizons in order to predict the \mathcal{IT} flow. This is consistent with the fact that predictors enriched with GMD data could learn the latent mobility patterns

TABLE II
ERROR METRICS OF EVALUATED MODELS. THE VALUES IN BOLD INDICATE THE LOWEST ERROR FOR EACH (TARGET FLOW, METRIC) PAIR.

Target flow	Model's input	MAE	RMSE	MAPE
\mathcal{A}	\mathcal{A}	30678.195	39895.783	13.764
\mathcal{A}	$\mathcal{A}, \mathcal{F}_{GM,100}$	44274.756	58651.799	20.762
\mathcal{A}	$\mathcal{A}, \mathcal{F}_{GM,400}$	29402.056	34364.585	12.206
\mathcal{A}	$\mathcal{A}, \mathcal{F}_{GM,800}$	29311.016	35820.294	11.998
\mathcal{IE}	\mathcal{IE}	3816.539	4983.614	24.589
\mathcal{IE}	$\mathcal{IE}, \mathcal{F}_{GM,100}$	3939.599	5628.872	20.193
\mathcal{IE}	$\mathcal{IE}, \mathcal{F}_{GM,400}$	5980.346	7317.516	37.581
\mathcal{IE}	$\mathcal{IE}, \mathcal{F}_{GM,800}$	4370.739	5344.350	26.865
\mathcal{NE}	\mathcal{NE}	12739.819	15732.936	35.154
\mathcal{NE}	$\mathcal{NE}, \mathcal{F}_{GM,100}$	9769.504	11509.648	25.019
\mathcal{NE}	$\mathcal{NE}, \mathcal{F}_{GM,400}$	6871.259	8544.438	16.443
\mathcal{NE}	$\mathcal{NE}, \mathcal{F}_{GM,800}$	7461.158	9421.957	16.706
\mathcal{E}	\mathcal{E}	15197.112	20773.676	29.911
\mathcal{E}	$\mathcal{E}, \mathcal{F}_{GM,100}$	15496.672	20741.639	33.164
\mathcal{E}	$\mathcal{E}, \mathcal{F}_{GM,400}$	14815.044	17828.300	28.326
\mathcal{E}	$\mathcal{E}, \mathcal{F}_{GM,800}$	11095.700	15434.727	20.736
\mathcal{IT}	\mathcal{IT}	8403.148	10333.727	12.709
\mathcal{IT}	$\mathcal{IT}, \mathcal{F}_{GM,100}$	10044.294	11134.676	15.101
\mathcal{IT}	$\mathcal{IT}, \mathcal{F}_{GM,400}$	10709.579	12664.216	15.613
\mathcal{IT}	$\mathcal{IT}, \mathcal{F}_{GM,800}$	7694.914	8850.736	11.922
\mathcal{NT}	\mathcal{NT}	16799.757	20668.949	14.330
\mathcal{NT}	$\mathcal{NT}, \mathcal{F}_{GM,100}$	21503.516	24611.753	18.729
\mathcal{NT}	$\mathcal{NT}, \mathcal{F}_{GM,400}$	13661.015	18604.120	10.718
\mathcal{NT}	$\mathcal{NT}, \mathcal{F}_{GM,800}$	18360.056	23826.051	14.729
\mathcal{T}	\mathcal{T}	16909.334	22205.017	9.350
\mathcal{T}	$\mathcal{T}, \mathcal{F}_{GM,100}$	27295.051	32113.587	15.807
\mathcal{T}	$\mathcal{T}, \mathcal{F}_{GM,400}$	15519.060	20499.031	8.806
\mathcal{T}	$\mathcal{T}, \mathcal{F}_{GM,800}$	17648.112	24587.760	9.052

from several points of view and, thus, anticipate their long-term behaviour in a more accurate manner.

Finally, the aforementioned evaluation shows that the usage of alternative human trips actually improved the prediction of most of the touristic flows under consideration. However, this improvement actually occurred when the GMD flow, $\mathcal{F}_{GM}^{\delta}$, comprised seasonal mobility that was compatible with the touristic activity rather than reporting regular and commuting trips. Besides, this improvement was most noticeable as long as the target time horizon increased.

VI. CONCLUSIONS AND FUTURE WORK

The utilization of human mobility data is revolutionizing the tourism industry, enabling the development of predictive models to optimize resource allocation for hotel companies.

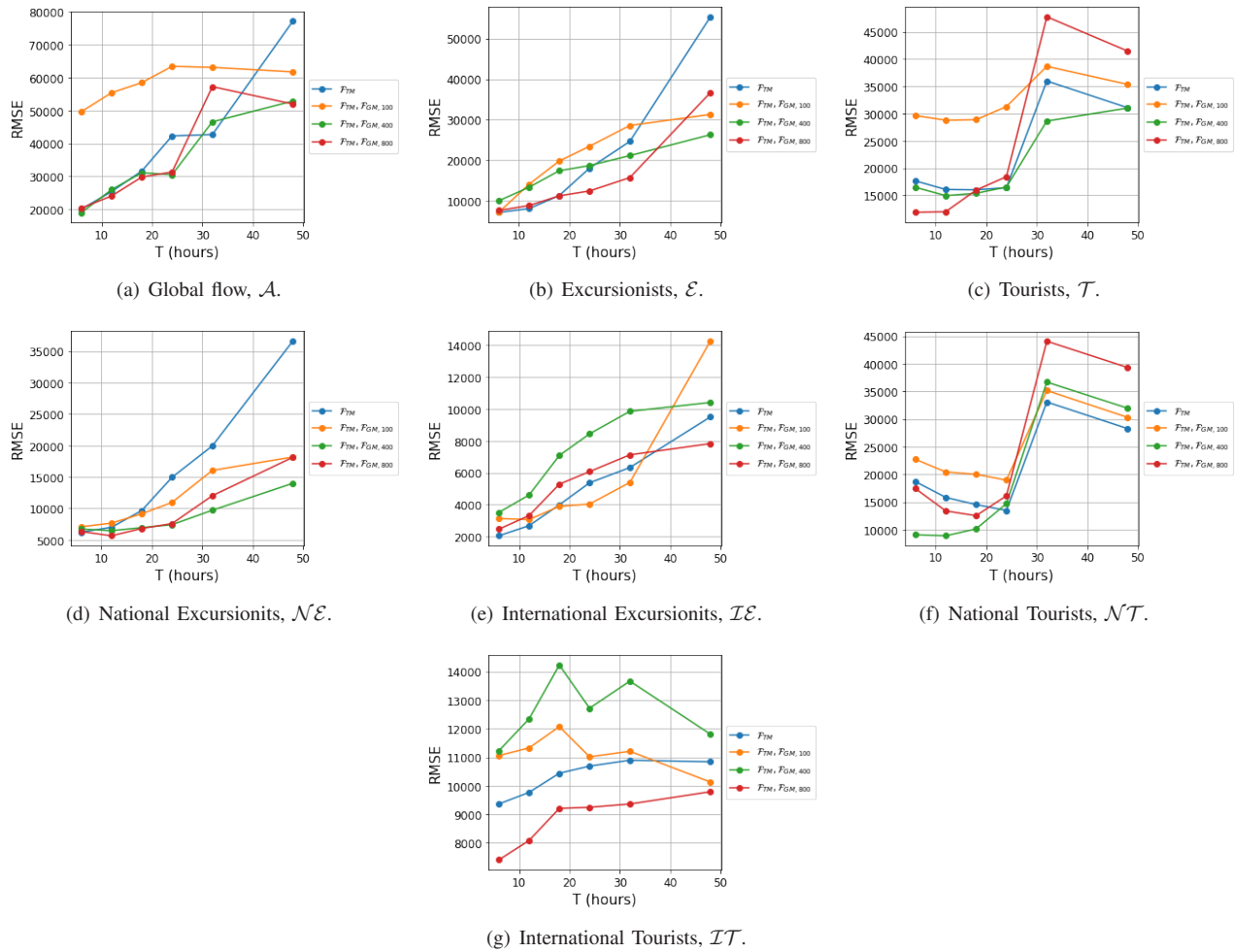


Figure 5. RMSE per time horizon for each target flow.

However, many existing models overlook general mobility patterns beyond tourist flows. In this study, we propose integrating general mobility data to enhance the accuracy of tourism flow predictions. Our innovative approach combines Convolutional Neural Network and Long-Short Term Memory models, enabling us to forecast tourist flows up to 8 weeks ahead with increased precision.

Testing our methodology on data collected from the Region of Murcia (Spain) over a 16-month period demonstrates significant improvements in accuracy, with error reductions exceeding 50%. This underscores the potential of integrating general mobility data into existing predictive models to better anticipate tourist behaviors.

Two avenues of research can be explored further in subsequent stages of this study. Firstly, the integration of additional contextual data, such as weather conditions, events, and holidays could further enhance the predictive accuracy of the tourist flow model. Secondly, expanding the analysis to encompass multiple regions would offer a broader understanding of tourist flows at a national level. This might reveal intricate patterns and dynamics between different areas, thereby enriching the

comprehensiveness of tourist mobility forecasting.

ACKNOWLEDGEMENTS

Financial support for this research has been provided by the Instituto de Turismo de la Región de Murcia through project CAT/TU/47-22 "Cátedra Internacional de Inteligencia Turística Región de Murcia UCAM-ITREM".

REFERENCES

- [1] B. Armutcu, A. Tan, M. Amponsah, S. Parida, and H. Ramkissoon, "Tourist behaviour: The role of digital marketing and social media", *Acta psychologica*, vol. 240, p. 104 025, 2023.
- [2] G. S. Atsalakis, I. G. Atsalaki, and C. Zopounidis, "Forecasting the success of a new tourism service by a neuro-fuzzy technique", *European Journal of Operational Research*, vol. 268, no. 2, pp. 716–727, 2018.
- [3] H. Albuquerque, C. Costa, and F. Martins, "The use of geographical information systems for tourism marketing purposes in aveiro region (portugal)", *Tourism management perspectives*, vol. 26, pp. 172–178, 2018.
- [4] C. Li, P. Ge, Z. Liu, and W. Zheng, "Forecasting tourist arrivals using denoising and potential factors", *Annals of Tourism Research*, vol. 83, p. 102 943, 2020. DOI: 10.1016/j.annals.2020.102943.

- [5] W. Wang *et al.*, “A multi-graph convolutional network framework for tourist flow prediction”, *ACM Transactions on Internet Technology (TOIT)*, vol. 21, no. 4, pp. 1–13, 2021. DOI: <https://doi.org/10.1145/3424220>.
- [6] L. Zhu, C. Lim, W. Xie, and Y. Wu, “Modelling tourist flow association for tourism demand forecasting”, *Current Issues in Tourism*, vol. 21, no. 8, pp. 902–916, 2018.
- [7] Y. Yang and H. Zhang, “Spatial-temporal forecasting of tourism demand”, *Annals of Tourism Research*, vol. 75, pp. 106–119, 2019.
- [8] Y. Zhang, G. Li, B. Muskat, and R. Law, “Tourism demand forecasting: A decomposed deep learning approach”, *Journal of Travel Research*, vol. 60, no. 5, pp. 981–997, 2021.
- [9] R. Law, G. Li, D. K. C. Fong, and X. Han, “Tourism demand forecasting: A deep learning approach”, *Annals of tourism research*, vol. 75, pp. 410–423, 2019.
- [10] N. M. De Jesus and B. R. Samonte, “AI in tourism: Leveraging machine learning in predicting tourist arrivals in philippines using artificial neural network”, *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, pp. 1–8, 2023.
- [11] Government of the Region of Murcia, *Tourism Statistics - Region of Murcia*, https://www.turismoregiondemurcia.es/es/estadisticas_de_turismo/, Accessed: October 8, 2024, 2024.
- [12] Government of the Region of Murcia, *Smart Tourism Destination (DTI) - ITREM*, <https://www.itrem.es/nexo/dti/>, Accessed: October 8, 2024, 2024.
- [13] Telefónica S.A., *Telefónica - Telecommunications Company*, <https://www.telefonica.com/es/>, Accessed: October 8, 2024, 2024.
- [14] Ministry of Transport, Mobility and Urban Agenda, *Daily basic mobility study using Big Data*, <https://www.mitma.es/ministerio/proyectos-singulares/estudios-de-movilidad-con-big-data/estudio-basico-diario>, Accessed: October 8, 2024, 2024.
- [15] M. Secretariat of State for Transport and U. Agenda, “Analysis of Mobility in Spain Using Big Data Technology During the State of Alarm for Managing the COVID-19 Crisis”, Spanish Ministry of Transport, Mobility and Urban Agenda, Tech. Rep., 2020.
- [16] G. Pasaoglu *et al.*, “Travel patterns and the potential use of electric cars—results from a direct survey in six european countries”, *Technological Forecasting and Social Change*, vol. 87, pp. 51–59, 2014.
- [17] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”, *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005, cited By (since 1996)149.