An investigation into using Twitter data to support the tourism industry in Goa

A Dissertation for
Course code and Title: IMC601 & Project Dissertation
Credits: 6
Submitted in partial fulfilment of Bachelor's Degree in M.Sc Integrated Data Science

by

VED BHALCHANDRA REDKAR 2103 368950884661 202100225
ANDRE EDUARDO NORONHA 2104 190573999788 202100147
DHRUV SOMNATH PATIL 2102 800679192585 202100216

Under the supervision of

SHRI. RAMDAS KARMALI

Goa Business School

Goa University

April 2024

DECLARATION BY STUDENT

We hereby declare that the data presented in this Dissertation report entitled, "An investigation into using Twitter data to support the tourism industry in Goa" is based on the results of investigations carried out by us in the M.Sc Integrated (Data Science) at the Goa Business School, Goa University/College under the Supervision of Shri. Ramdas Karmali and the same has not been submitted elsewhere for the award of a degree or diploma by me. Further, we understand that Goa University or its authorities will be not be responsible for the correctness of observations / experimental or other findings
given the dissertation.

I hereby authorize the University authorities to upload this dissertation on the dissertation repository or anywhere else as the UGC regulations demand and make it available to any one as needed.

Dhruv Somnath Patil
2102

Andre Eduardo Noronha
2104

Ved Bhalchandra Redkar
2103

Date: -
Place: Goa University, Goa
Business School

COMPLETION CERTIFICATE

This is to certify that the dissertation report "An investigation into using Twitter data to support the tourism industry in Goa" is a bonafide work carried out by Mr_____

under my supervision in partial fulfilment of the requirements for the award of the degree of M.Sc Integrated (Data Science) in the Discipline Computer Science and Technology at the Goa Business School, Goa University.

Shri. Ramdas Karmali

Date:
Place: Goa University, Goa
Business School

Stamp

CONTENTS

CONTENTS

CONTENTS

Preface

Goa, known for its pristine beaches, vibrant culture, and rich heritage, has long been a coveted destination for domestic and international tourists alike. However, in recent years, Goan tourism has been experiencing a decline, posing significant challenges to the state's economy and livelihoods dependent on the Tourism Sector. This project aims to address this pressing issue by harnessing the power of social media, particularly Twitter, to provide valuable insights and suggestions to the government for revitalising Goan tourism.

The project will entail comprehensive data collection and analysis of Goan tourism, focusing on demographics, attractions, seasonality, and factors influencing visitor numbers. Leveraging natural language processing, sentiment analysis of Goan tourism tweets will identify prevalent themes and suggestions. Twitter data and other sources will be scrutinized to pinpoint critical challenges and opportunities. Engaging stakeholders through Twitter and other platforms will facilitate input gathering for enhancing tourism strategies. From these insights, actionable policy recommendations will be crafted, spanning infrastructure, marketing, sustainability, and community engagement. Ongoing support for policy implementation and continuous evaluation will ensure adaptability and effectiveness in addressing Goan tourism issues.

By harnessing the real-time insights and collective intelligence available on Twitter, this project seeks to empower the government of Goa with timely and data-driven recommendations for revitalizing the Tourism Sector. Through collaborative efforts and innovative solutions, we aspire to reverse the declining trend in Goan tourism and ensure a sustainable and prosperous future for this iconic destination.

Acknowledgement

We are immensely grateful for the contributions and support received from our colleagues throughout the duration of this project, which was vital in bringing this research to fruition. We would like to extend our heartfelt thanks to Mr. Ved Redkar, whose extensive involvement in the literature review and research significantly shaped our foundational understanding and approach. Mr. Redkar's diligent efforts in survey execution, data cleaning, and model generation were crucial in ensuring the integrity and robustness of our analysis.

We also owe a great deal of gratitude to Mr. Andre Noronha for his pivotal role in data cleaning and transformation. His expertise was instrumental in refining our datasets, which enabled the effective running of our predictive models. Mr. Noronha's contributions to the literature review further enriched our research framework, providing a deeper insight into the current trends and methodologies.

Furthermore, we would like to acknowledge Mr. Dhruv Patil for his dedicated work with the modeling aspects of our project. His technical proficiency brought significant advancements to our analytical capabilities. Mr. Patil's involvement in literature review and data cleaning also helped maintain the continuity and quality of our research process.

Special thanks are also due to Mr. Ramdas Karmali, who has been a guiding force throughout this project. His insights and guidance were instrumental in steering the direction of our research and ensuring that we remained aligned with our objectives and scientific rigor.

The collaborative effort of these individuals have been invaluable, and it is with great appreciation that we recognize their contributions to this project. Their expertise and commitment were key to achieving the insights and outcomes presented in this research.

Tables and figures

Tables

Tables and figures

Tables

Tables and figures

Tables

Tables and figures

Figures

Tables and figures

Figures

Tables and figures

Figures

Tables and figures

Figures

Tables and figures

Figures

Tables and figures

Figures

Tables and figures

Figures

Tables and figures

Figures

Tables and figures

Figures

Tables and figures

Figures

Abbreviations used

A1 - Initial dataset (155 keywords)
A2 - Beach related dataset (36 keywords)
A3 - Beach related but more focused dataset (only certain beaches) (17 keywords)
A4 - Final dataset (Sample taken of A3)

V1 - First cleaned variation of A1
V2 - Second cleaned variation of A1
V3 - Third cleaned variation of A1
V4 - Fourth cleaned variation of A1
V5 - Fifth cleaned variation of A1

D1 - First cleaned variation of A4
D2 - Second cleaned variation of A4
D3 - Third cleaned variation of A4
D4 - Fourth cleaned variation of A4
D5 - Fifth cleaned variation of A4

Twitter - X

Abstract

This study was initiated to explore the potential of Twitter data in enhancing policy-making for the tourism sector in Goa by leveraging the insights obtained through sentiment analysis. Recognizing the transformative impact of social media on public discourse, we employed advanced Natural Language Processing (NLP) techniques and deep learning models to analyze tweets related to tourism in Goa.

The methodology involved collecting and segmenting Twitter data based on keywords that reflect key aspects of tourism, followed by a sentiment analysis to gauge public opinion and identify prevailing trends and concerns. Our findings revealed significant patterns in public sentiment that are crucial for informed decision-making in tourism management. The analysis indicated varying degrees of positive and negative perceptions that tourists hold regarding different facets of the Goa tourism experience. Importantly, the data pointed to specific areas where policy interventions could significantly enhance tourist satisfaction and sustainably grow the tourism sector.

These observations suggest that Twitter, as a rich source of real-time public opinion, can provide actionable insights that help shape effective tourism policies. By tapping into the collective sentiment of tourists, stakeholders in Goa's tourism industry can implement strategic changes that enhance visitor experiences and ensure the long-term viability of this vital economic sector. This study underscores the importance of integrating social media analytics into the decision-making processes of tourism management to maintain and enhance Goa's appeal as a premier travel destination.

Keywords: *Tourism, Policy, Twitter, Sentiment, BERT, RoBERTa*

1. INTRODUCTION

1.1 Background

Tourism is a significant economic pillar for Goa, a state renowned globally for its picturesque landscapes, rich history, and vibrant culture. The sector not only supports local livelihoods but also contributes substantially to the state's economy. However, like many popular tourist destinations, Goa faces the dual challenge of maximizing the economic benefits of tourism while managing its social and environmental impacts. Addressing these challenges requires a nuanced understanding of tourist behaviors, preferences, and their perceptions of the destination.

In recent years, the advent of digital technology and social media has transformed how information is shared and consumed, particularly in the tourism industry. Platforms like Twitter have become powerful tools for expression and communication, offering real-time insights into public opinion and sentiment. These platforms are not only used by tourists to share their experiences but also by local businesses and governmental bodies to engage with visitors and promote tourism.

Given the richness of data available from social media, there is a tremendous opportunity to harness these digital footprints to gain deeper insights into tourist sentiments and trends. This approach is particularly relevant in the context of Goa, where traditional data collection methods may not fully capture the diverse and dynamic nature of tourist interactions and experiences.

This research paper seeks to explore the potential of social media analytics, specifically through Twitter data, to enhance the understanding of tourism dynamics in Goa. By employing Natural Language Processing (NLP) and deep learning techniques, this study aims to analyze the sentiments expressed in tweets related to Goa’s tourism. The objective is to uncover the underlying patterns and themes that characterise public perception, thereby providing a basis for informed decision-making and strategic planning.

Through this background, the research sets the stage for a detailed examination of how innovative data analysis techniques can be leveraged to improve tourism management and policy-making in Goa. By focusing on actual visitor feedback gathered through social media, the study contributes to the broader efforts to enhance the sustainability and resilience of the tourism sector in the region.

1.2 Aim and Objectives

Aim of the Study

The aim of this study is to empower the Goa Tourism Sector by utilizing data-driven insights and sentiment analysis to enhance decision-making processes, improve collaboration among stakeholders, and enrich the overall tourist experience in Goa.

Objectives of the Study:

1.*Data Analysis for Insight Generation:* To collect and analyze Twitter data related to tourism in Goa using advanced natural language processing (NLP) techniques to derive meaningful insights that can inform strategic decisions.

2.*Stakeholder Collaboration:* To foster stronger collaboration among tourism stakeholders in Goa by providing them with actionable data insights that facilitate coordinated efforts and shared goals.

3.*Revitalisation of Tourism Experiences:* To use the insights gained from sentiment analysis to revitalize and enhance the tourism offerings in Goa, ensuring that visitors have memorable and engaging experiences.

1.3 Research Question

1.3.1 Hypothesis

Can Twitter data act as a resource to assist the tourism sector in Goa?

1.3.2 Problem Statement

Despite Goa's historical significance and natural beauty, the Tourism Sector is experiencing a concerning decline. This decline threatens the livelihoods of local businesses, the economic stability of the region, and the overall appeal of Goa as a tourist destination. However, the root causes of this decline remain ambiguous, and there is a lack of actionable insights to guide strategic interventions. Inaccurate perceptions, negative reviews, and unaddressed challenges may be some of the factors that contribute to the diminishing appeal of Goa among tourists and locals alike. Without a comprehensive understanding of these issues and a targeted approach to address them, the revitalisation of Goa's Tourism Sector remains an elusive goal. Therefore, there is an urgent need for a data-driven solution that can analyse sentiments, and provide recommendations to the Tourism Sector for steady growth.

1.4 Scope

The scope of this research encompasses a comprehensive analysis of Twitter data to extract valuable insights into public sentiment regarding tourism in Goa. Utilizing advanced Natural Language Processing (NLP) and deep learning methodologies, the project aims to systematically categorize and evaluate the sentiments and thematic concerns expressed by tourists and stakeholders on social media. This includes the identification of prevailing positive and negative sentiments, as well as the extraction of recurring themes related to the tourism experience. The insights derived from this analysis will inform strategic decision-making and policy formulation aimed at enhancing the tourism sector's performance and sustainability. Moreover, the project seeks to contribute to the academic and practical understanding of the application of social media analytics in tourism management, providing a model that could be replicated or adapted for other regions or sectors. Through this endeavor, the project will directly support the ongoing initiatives to maintain and improve Goa's status as a premier global tourism destination, ensuring it remains attractive and viable for future generations of travelers.

2. LITERATURE REVIEW

2.1 Introduction

In December 2023, typically the peak of Goa's tourism season, there has been a noticeable decline in the number of high-quality tourists—those who stay longer and spend more—according to local tourism stakeholders. Instead, Goa has seen an increase in low-budget tourists, which has raised concerns among local business owners. Stakeholders expressed frustration over government policies, highlighting the need for strategies that attract more affluent, typically foreign, tourists. They also emphasized the importance of implementing a single window system for obtaining necessary permissions to streamline processes and support local businesses.

Several issues were pointed out as contributing to the decline in high-quality tourism. These include bureaucratic delays, such as the late issuance of shack licenses, which led to cancellations and financial losses for hoteliers. The absence of a competitive and welcoming visa policy was also mentioned as a barrier, making Goa less attractive compared to other destinations that offer easier access and better deals.

Local tourism leaders called for immediate action from the government to improve the situation. They advocated for better promotion in international markets, more competitive pricing, and enhancements in safety and general tourism infrastructure. The consensus was clear: for Goa to remain competitive on an international level and attract the desired caliber of tourists, significant changes in policy and management are essential.

In the following sections of this paper, we will explore the factors contributing to the decline in tourism, define what constitutes quality tourists, examine the growth trends in the tourism sector, and assess how social media can aid in making informed decisions. Additionally, we will provide recommendations for government policy formulation.

2.2 Tourism developments in Goa

2.2.1 Growth in Domestic and Foreign Tourist Arrivals

Tourist visits to Goa have consistently increased since 1961, with domestic tourists primarily comprising the initial influx. The number of tourists grew from 50,000 in 1964 to 200,000 by 1974-75. During the 1970s, the region also started attracting foreign tourists, particularly hippies, to the coastal areas of Baga and Anjuna. Tourist numbers reached 384,000 in 1980, and by 1985, this number had nearly doubled to 775,000. The year 1985 was pivotal for Goa's tourism industry, marking the beginning of high-end tourism with the introduction of charter flights from abroad, enhancing foreign exchange earnings. The 1990s saw further acceleration in tourism growth, with numbers reaching 1.059 million. By 2004, tourist visits had surged to approximately 2.1 million. In 2015, total tourist arrivals crossed the five million threshold, including 541,000 foreign visitors, despite economic downturns in key European and Russian markets. This significant rise of 30 percent from the previous year underscores the growing appeal of Goa, [2] particularly as a destination for high-end tourism, facilitated by the advent of international charter flights.

2.2.2 High-end Tourism

The paper reference [2] taken talks about onset of Goa as a High-end tourism which refers to a segment of the travel industry that caters to tourists seeking luxury experiences. These tourists typically expect high levels of service, exclusivity, and comfort. High-end tourism often involves stays at premium hotels or resorts and access to unique, personalized experiences that are not widely available to the general public. This type of tourism is also characterized by higher spending per tourist compared to other types of travelers. The focus is on quality, privacy, and providing tailored services that meet the specific desires and needs of affluent travelers.

2.2.3 Constituents of Foreign tourists

Period	Domestic Tourists	Foreign Tourists	Total Tourists
1970-2014	7.31***	10.83***	7.71***

Table 1: Average Annual Growth rates of Tourist Arrivals in Goa (in percent)

Table 1 shows that the total number of tourist arrivals in Goa has increased at an average annual rate of 7.71% between 1970 and 2014. The growth rate of foreign tourists, at 10.83%, surpasses that of domestic tourists, which stands at 7.31%. Notably, the proportion of charter tourists among total foreign visitors has surged from just 4% in 1985 to 53% by 2015. When analyzing foreign tourists by their country of origin, notable shifts are evident. Initially, in the early 1970s, the majority of foreign tourists came from the USA, the UK, and Germany, with the USA leading until 1974-75. Subsequently, the UK has consistently sent the largest number of tourists, except during a brief period from 1979 to 1982 when Germany was the top source. From 1994 to 2004, the UK remained the predominant source of tourists, mainly charter visitors, averaging 114,000 arrivals per year, followed by Germany with 24,400. While tourist arrivals from Portugal have declined over time, recent years have seen a rise in tourists from Russia, Finland, and Germany, although the UK continues to be the leading source of foreign tourists annually.

Additionally, it's important to note that there has been a consistent growth in Russian tourists visiting Goa from 2014 to 2018, contributing to the overall increase in foreign tourist arrivals.

Countries	1994-95	2004	2014
U.K.	**74.00**	**41.00**	**28**
Finland	**-**	**9.66**	**5**
Germany	**14.90**	**7.29**	**10.89**
Russia	-	**7.05**	**29.14**
U.S.A	1.50	1.68	1.94
Sweden	-	2.66	3.67
France.	1.88	2.22	6.83
Switzerland.	3.64	2.60	2.65
Portugal	-	1.05	0.65
Others	4.08	24.79	11.23

Table 2: Foreign Tourist Arrivals in Goa by Country of Origin (in percent)

Table 2 highlights changes in the origins of foreign tourists visiting Goa over several years. In 1994-95, the majority of foreign tourists were from the U.K., accounting for 74% of the total, followed by Germany with 15%. By 2004-05, however, the share of tourists from both the U.K. and Germany had declined to 41% and half of the previous percentage, respectively. This shift was due to an increase in tourists from France and Russia. By 2014, Russian tourists constituted 29.14% of foreign arrivals, closely followed by the U.K. at 28%, and Germany at 10.89%.

The data suggests that while the U.K. remains a significant source of tourists, Goa is attracting more diverse international visitors, establishing itself as a globally recognized tourist destination. The state's tourism department noted that despite economic downturns in Europe, tourist arrivals from various countries, including the U.K., Russia, France, and the UAE, have increased(2014). This growth is partly attributed to the introduction of the e-tourist visa by the Indian government in November 2014, which significantly boosted arrivals.

2.2.4 Growth of Hotels and Restaurants in Goa

The growth in tourist inflows to Goa has significantly spurred the expansion of its hotel industry. From 1964 to 1975-76, the number of hotels grew from 49 to 138, representing an average annual growth rate of 16.51%. During this period, the number of starred hotels increased slightly from four to seven, while the total bed capacity rose from 1,048 to 3,671.

By 1980, the number of hotels had escalated to 196, including 12 star hotels and 184 non-star hotels, providing a total of 6,587 beds. By 1988, the number of beds had reached 11,140, spread across 281 hotels and lodges. In 1993, the bed count further increased to 15,100 with approximately 7,500 rooms, about 22.48% of which were in star category hotels.

Data from the Tourism Master Plan of 2011 highlighted that in 1994 there were 400 hotels with 17,500 beds, and by 1996, the number had risen to 436 hotels with 18,391 beds. Tiswadi, Salcete, and Bardez talukas had the highest number of hotels. Notably, 77% of the hotels were located along the coastal belt, accommodating 77% of domestic tourists and 95% of foreign tourists who preferred these areas.

By 2005, the total number of hotels in Goa had surged to 2,156, a nearly tenfold increase since statehood. These establishments provided 19,312 rooms and 36,618 beds. In this year, 78 hotels were in the star category, offering 5,836 rooms and 11,441 beds.

A decade later, in 2015, the number of hotels reached 3,358, with 31,767 rooms and 56,595 beds, including 63 star category hotels with 5,362 rooms and 10,001 beds. This growth highlights not only the expansion of star and quality category hotels but also a significant rise in informal sector establishments such as small hotels, joints, and restaurants.

2.2.5 Tourism's Role in Goa's Employment Structure (1991-2011)

Between 1991 and 2011, Goa's economy saw a significant shift in employment from the primary (agriculture) and secondary (manufacturing) sectors to the service sector, which includes tourism. Data from the Census of India and the National Sample Survey Organization (NSSO) reveal changes in how people work in Goa.

In 1991, 30% of workers were in the primary sector, 20.43% in the secondary sector, and 42.49% in the service sector. By the year 2000, based on the NSSO's 55th round of data, the primary sector's share dropped to 19.63%, the secondary sector rose slightly to 22.75%, and the service sector increased to 57.62%, driven by growth in community services and the hospitality industry.

By 2011-12, the trend continued with the primary sector further declining to 14.28%, and the secondary sector jobs in manufacturing and construction at 15.87% and 10% respectively. Meanwhile, service sector employment climbed to 60%, with the tourism sector alone accounting for 24.25% of jobs, underscoring its growing importance in Goa’s economy.

Tourism services have become a crucial factor in driving economic growth in Goa, significantly influencing both output and employment. Over time, Goa's economy has undergone a structural transformation, with a growing proportion of the workforce shifting into the service sector. This change highlights the service sector's role as a major employer and underscores the critical importance of tourism services within the broader employment landscape of the state.

2.2.6 Costs and Challenges of Tourism in Goa

Tourism in Goa impacts almost everyone in the region economically, both positively and negatively. While it brings substantial economic benefits, tourism also entails several costs. These include direct expenses faced by tourism businesses and government expenditures on infrastructure improvements to accommodate tourists. Additionally, the local community often bears the cost of congestion and other related issues.
[2] These economic costs are significant factors to consider in the development of state tourism strategies. Without careful planning and control, tourism can lead to unwanted consequences. For instance, unmanaged tourism growth could diminish Goa's appeal as a serene and safe destination. Various economic and socio-cultural challenges arising from mass tourism in the region will be explored in the following sections.

2.2.7 Deteriorating Condition of Beaches in North Goa

Study [2] also talks about the condition of the beaches, The once pristine white beaches of North Goa have become overcrowded, cluttered with an array of shacks, hotels, and motels. This congestion is not only a violation of Coastal Zonal Regulations but has also spurred multiple pollution issues. The rapid and disorganised expansion of these beachside accommodations contributes significantly to environmental degradation through improper disposal of large volumes of solid and liquid waste.

The tourism infrastructure at popular spots like Calangute and Vagator beaches is inadequate and fails to meet standard quality expectations. During peak tourist seasons, Calangute beach is plagued by garbage heaps that emit foul odors, and it lacks essential facilities such as changing rooms, toilets, and proper waste management systems. There is also a notable absence of maintenance for amenities like benches, inadequate police patrolling, and poor lighting.

Vagator beach presents an even grimmer scenario with no urinals, drinking water facilities, changing rooms, adequate dustbins, or security measures. Even the tourist police booth is in disrepair, underscoring a profound neglect of essential services. These issues highlight a critical lack of planning, coordination, and timely execution of policies essential for maintaining the natural beauty and visitor-friendliness of Goa's beaches.

2.2.8 Suggestions to Improve the Standard and Quality of Tourism in Goa

The paper [2] has provided some ways to enhance the standard and quality of tourism in Goa, several strategic measures are recommended for immediate implementation by the authorities:

1.*Diversification of Tourism Products:* To alleviate the pressure on beaches, the focus should shift towards:

Eco-tourism: Promoting activities in Goa's Western Ghats, wildlife sanctuaries, and natural forests.

Adventure Tourism: Encouraging sports like parasailing, surfing, and trekking in the hinterland regions of Sanguem, Satteri, and Canacona.

Business Tourism: Enhancing infrastructure for conventions and creating facilities like golf courses and world-class shopping centers.

Health Tourism: Developing health resorts, botanical gardens, and Ayurvedic clinics with incentives for private investment.

2.*Beach Cleanliness and Maintenance:* Initiatives such as forming a beach cleanliness task force, banning hawkers and plastic bags, beautifying beaches, and enforcing regulations against illegal structures. Implementing sewage treatment systems for coastal properties is also crucial.

3.*Transportation and Infrastructure Improvement:* Upgrading roads, expanding airport facilities, enhancing local bus services, introducing a metering system for taxis, and setting up comprehensive tourist information centers.

These actions aim not only to attract more foreign tourists but also to sustain Goa’s image as a clean and green destination. Given Goa's competitive tourism landscape, both domestically and internationally, it's essential to continuously innovate and improve the tourism experience.

But these suggestions come out be very general and we need to make it more specific when it comes to suggesting the Government in policy designs. Hence, engaging with the community and considering public opinion in policy-making will be vital in addressing the challenges at beaches and other tourist points effectively. By involving local stakeholders and the opinion of tourists in these discussions, Goa can craft policies that are well-suited to maintaining its unique charm and appeal as a top tourist destination.

2.2.9 Conclusion

The research paper [2] explicitly addresses the emergence of high-end tourism (2.2.1), defined as the segment of the travel industry catering to tourists seeking luxury experiences. This highlights that the perception of Goa becoming an expensive destination is not linked to the current downturn in tourism numbers. Instead, it indicates that the shift towards more upscale offerings is a distinct evolution of the market rather than a cause of the decline.

2.3 Decline in Tourism of Goa

2.3.1 What is the decline about?

The notion of a decline in tourism frequently serves as the subject of numerous articles(2022-23); however, this is often a misinterpretation or an ambiguously used term that could lead to various misunderstandings. For instance, one might ask whether it implies a reduction in the number of visitors. In fact, the overall visitor count in the state has remained relatively stable, primarily due to the consistent numbers of domestic tourists. Yet, this stability masks a significant shift in the composition of the tourist population, particularly a noticeable decrease in high-spending tourists, who are predominantly foreign visitors but also include domestic tourists who spend generously. This group's diminished presence is critical because their spending substantially contributes to the local economy.

2.3.2 Trend in Domestic tourists

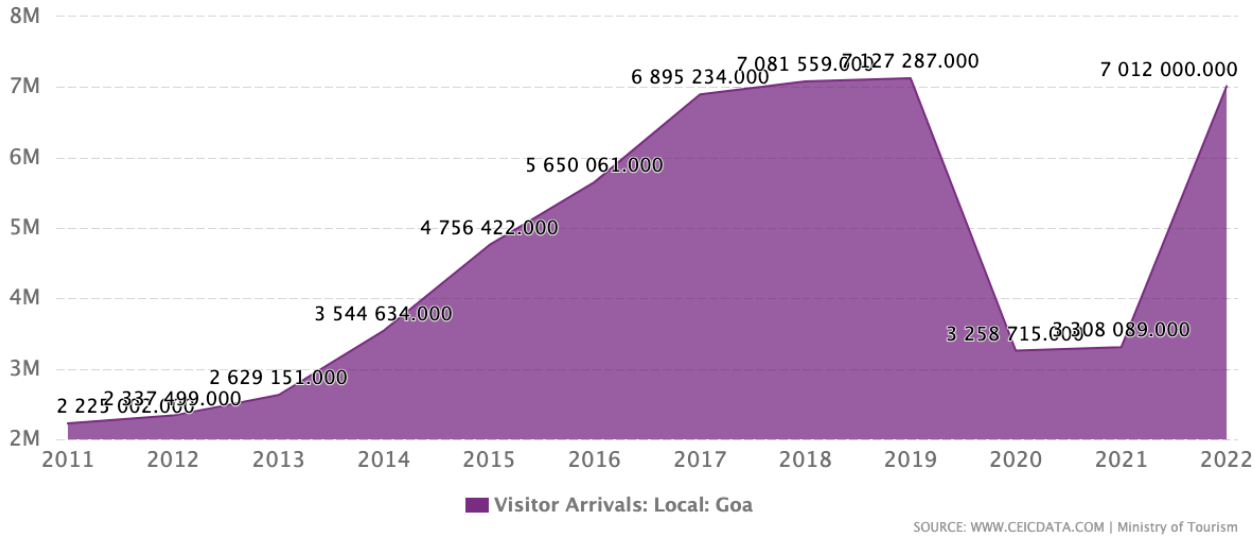


Figure 1: Number of domestic tourists that visited Goa (2011 - 2022)

The graph shows the number of visitors to Goa from 2011 to 2022 according to the Ministry of Tourism [CEICdata.com]. There was a sharp drop in visitors in 2020, which likely reflects the impact of the COVID-19 pandemic on tourism.

- In 2011, there were approximately 2.2 million visitors.
- By 2019, the number of visitors had risen to over 7 million.
- There was a drop to approximately 3.3 million visitors in 2021.
- In 2022, the number of visitors recovered to over 7 million.

2.3.3 Trend in foreign tourists

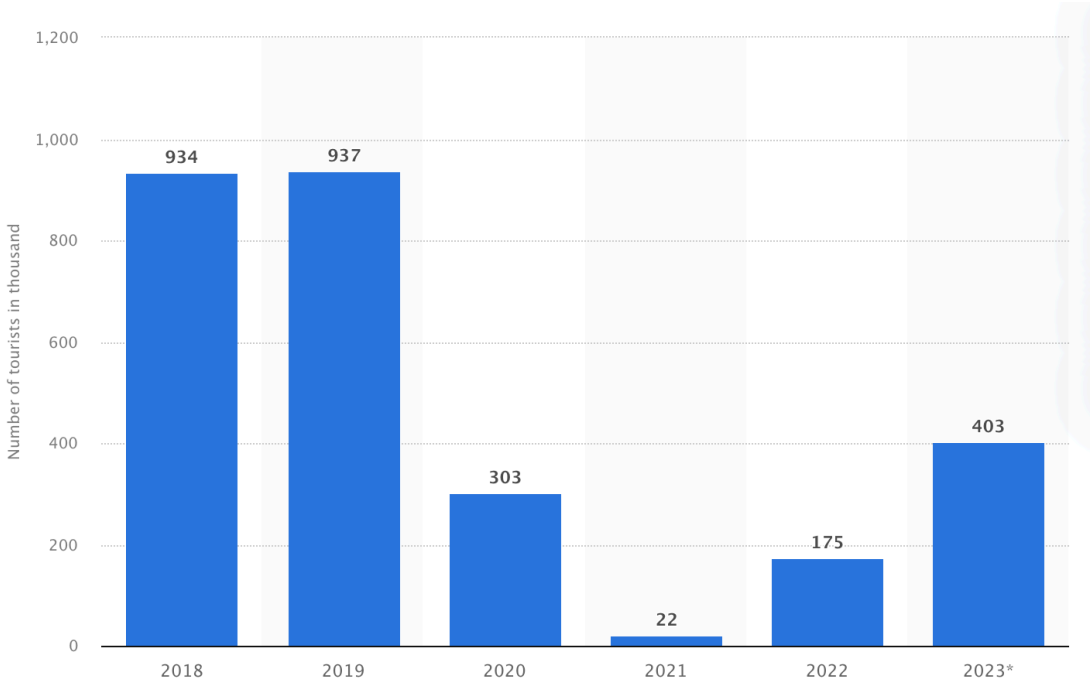


Figure 2: Number of foreign tourists that visited Goa (2018 - 2023)

The number of foreign tourists visiting Goa fluctuated between 2018 and 2023. There were 934,000 foreign tourists in 2018 and 937,000 in 2019. The number of visitors then dropped significantly in 2020, likely due to the COVID-19 pandemic, to just 22,000. The number of foreign tourists increased slightly in 2021 to 175,000 and then again in 2022 to 303,000. In 2023 (marked with an asterisk), there were an estimated 403,000 foreign tourists visit Goa. While the number of foreign tourists has rebounded somewhat since the pandemic, it has not yet reached pre-pandemic levels.

2.3.4 Clarification on decline

This observation serves as a key indicator of the underlying reasons for the perceived decline in tourism in Goa. It is important to clarify that this decline specifically pertains to a decrease in the number of foreign tourists visiting the region, rather than a general decrease in overall tourism. Despite the apparent downturn, it's noteworthy that the footfall from domestic tourists has remained quite stable over time, underscoring a shift rather than a total reduction in tourism.

This distinction is crucial for understanding the dynamics affecting Goa's tourism industry. The stability in domestic tourism suggests that the appeal of Goa as a local destination remains strong, while the drop in foreign tourists could be attributed to a variety of factors. These may include economic conditions in source countries, competitive destinations offering more attractive packages, or perhaps changes in visa policies and air connectivity, which could deter foreign visitors.

Tourist arrivals in Goa are rebounding from the pandemic slump, with a noticeable shift from international charters to domestic tourism. According to the Union tourism ministry, the last five years have seen a steep decline in international tourists, while domestic visits have remained relatively stable. In 2018, Goa welcomed 70.8 lakh domestic tourists compared to 9.3 lakh foreign tourists. By last year, domestic figures slightly dipped to 70.1 lakh, but foreign arrivals plummeted to just 1.7 lakh.
[2]

President of the Travel and Tourism Association of Goa, notes that domestic tourism is performing well, almost reaching last year’s numbers. However, international tourism is recovering slowly, hindered by global uncertainties like wars.
While the domestic tourist sector in Goa thrives, smaller family-run hotels that traditionally rely on foreign charters are facing challenges.

In this context, it becomes essential for stakeholders to analyse these trends thoroughly and develop targeted strategies to attract foreign tourists back to Goa. This might involve enhancing marketing efforts in key foreign markets, improving infrastructure, or offering unique experiences that are tailored to international tastes and preferences. Additionally, maintaining the robustness of domestic tourism while rebuilding the international visitor base will be pivotal in ensuring the overall health and sustainability of Goa's tourism sector.

2.3.5 Foreign tourist footfall in India

In 2018 and 2019, India welcomed 10.56 million and 10.93 million international visitors, respectively. However, these numbers plummeted to 2.74 million in 2020 and further to 1.52 million in 2021, primarily due to the lockdowns triggered by the COVID-19 pandemic. In 2023, India's inbound tourism sector has witnessed a robust recovery, with foreign tourist arrivals increasing by 64 percent compared to 2022. Specifically, from January to December 2023, the country recorded 9,236,108 international visitors, up from 6,437,467 during the same period the previous year.

This suggests that the decline in the number of foreign tourists is confined to Goa, while the overall national figures remain stable.

2.3.6 Possible reasons for the decline

2.3.6(a) e-Visa delay

A major reason being pointed at is e-visa but if e-Visa delays were truly a major barrier, we would expect to see a drop in foreign tourists across all of India's popular destinations, not just Goa. The fact that the decline is mostly seen in Goa suggests that other local factors might be at play. These could include competition from other tourist spots, changes in what tourists are looking for, or local economic conditions.
This indicates that while e-Visa issues might pose some challenges for attracting foreign tourists to Goa, they are probably not the main reason for the decline. There are likely several reasons contributing to this trend, and a deeper look into all these factors is needed to fully understand and tackle the decrease in foreign tourists in Goa.

2.3.6(b) Russia - Ukraine War

In the aftermath of the war, wages in Russia have seen a significant rise, with the lowest earners experiencing the fastest wage growth at around 20% annually. Despite the challenges posed by the war, such as a high number of casualties and continued inflation issues, the Russian economy has shown resilience. Salaries have been increasing notably, driven by substantial government expenditure and a labor shortage exacerbated by military recruitment.

As a result, real disposable income has seen growth not witnessed in many years, with the average monthly wage now 30% higher than two years prior. This economic boost has provided opportunities for social and economic mobility that were previously unattainable for many Russians, allowing some to make significant life improvements such as purchasing homes or starting new businesses. Despite the ongoing challenges, this period has marked a notable increase in people's disposable incomes following the war.[3]

In 2018, Russians continued to dominate the foreign tourist arrivals in Goa, with British tourists coming in second, as revealed by statistics during the Goa Assembly's budget session. The Tourism Minister, Manohar Ajgaonkar, reported that from January to October 2018, approximately 3.11 lakh Russian tourists visited Goa, followed by 1.48 lakh British tourists. Fast forward to January 2023, Russian tourists also significantly marked their presence in Dubai, registering as the second-largest group of inbound travelers with a notable 63 percent increase from the previous year, totaling around 115,000 visitors. [4]

Comparatively, in 2017, Russians made up 58 percent of Goa's international tourist arrivals, which significantly dropped to 7.41 percent (12,626 tourists) of the 1.69 lakh foreign arrivals in the state for the year 2022, indicating a shift in travel patterns among Russian tourists over the years.

2.3.6(c) Rise of Spirituality

A discernible shift is taking place within Indian tourism. Beyond the allure of beaches and historical landmarks, a growing number of foreign visitors are seeking experiences that nourish the soul. This burgeoning trend highlights the significance of spiritual tourism in India, a nation with a longstanding legacy as a cradle of diverse religions and philosophies. The magnetism is undeniable. Ancient cities like Varanasi, a vibrant center of Hindu pilgrimage on the Ganges, offer a window into timeless rituals and profound faith. Rishikesh, nestled in the Himalayas and aptly christened the "Yoga Capital of the World," beckons with the promise of inner peace through its yoga retreats and meditation centers, all set against a backdrop of breathtaking mountain vistas. However, India's spiritual smorgasbord extends far beyond these iconic destinations. Dharamshala, a heaven for Tibetan Buddhism, welcomes visitors seeking the wisdom of the Dalai Lama. Amritsar, the holiest city in Sikhism, resonates with a spirit of community and service. And Auroville, an experimental spiritual township, embodies the pursuit of human unity. This surge in spiritual tourism reflects a growing global yearning for meaning and inner well-being, a hunger that India, with its ancient wisdom and vibrant practices, is uniquely positioned to satiate.

S.No	State/UT	FTVs in Million	Percentage Share
1	Gujarat	1.78	20.70
2	Maharashtra *	1.51	17.60
3	West Bengal	1.04	12.08
4	Delhi *	0.82	9.50
5	Uttar Pradesh	0.65	7.56
6	Tamil Nadu	0.41	4.74
7	Rajasthan	0.40	4.62
8	Kerala	0.35	4.02
9	Punjab	0.33	3.84
10	Madhya Pradesh	0.20	2.38
	Total of Top 10	7.47	87.03
	Others	1.11	12.97
	Total	**8.59**	**100.00**

Table 3: Share of Top 10 States/UTs of India in Number of Foreign Tourist Visits in 2022

***: Data for 2022 is estimated by applying all India growth rate for 2022/19 on 2019 data**

In 2022, Gujarat emerged as the leading state in India for attracting foreign tourists, according to data released by the Tourism Ministry. This notable achievement highlights Gujarat's growing appeal as a prime destination on the global tourism map.

2.4 Social Media Analytics

2.4.1 Overview

Social media has transformed communication, becoming a mainstream tool for brands of all sizes. These entities utilize social media to understand customer needs and increase profits through ongoing public engagement. In recent years, governments and policymakers have also recognized social media's value, using it to foster greater public participation and influence over governmental initiatives. For instance, citizens now use platforms like Twitter to highlight issues such as potholes, directly engaging with authorities by showcasing the problems or potential dangers, often with a tone of mockery or serious concern.

2.4.2 Twitter as a tool

A study[6] was performed to employ Twitter analytics to examine two significant public engagement cases to demonstrate how insights derived can inform policy. Social media analytics is increasingly recognised as crucial for business and governance, enabling a better understanding of public perceptions and the potential to revise policies based on these insights.

Information dissemination through social media is cost-effective and broad-reaching, allowing for bidirectional communication where citizens can directly respond to government posts with queries, feedback, or complaints. This interaction not only facilitates domestic information exchange but can also extend internationally, enhancing global cooperation on various issues.

2.4.3 Study to demonstrate the effect of Twitter in Policy Making

GST, or Goods and Services Tax, aims to simplify the tax system in India by consolidating multiple taxes into a single framework, promoting a unified market.
A study analyzed public reactions on Twitter to both Demonetisation and GST from November 2016 to February 2017.

They observed a significant difference in public engagement: Demonetisation tweets soared to 159,433 due to its immediate impact on everyday transactions, while GST, which interests mainly economists and businesses, garnered 33,570 tweets.

Tweet analysis showed that reactions to Demonetisation were swift and intense, peaking quickly after its announcement and then sharply declining. In contrast, discussions about GST were more consistent over time, reflecting ongoing speculation about its future implications rather than immediate impact.

This comparative study highlights how different economic reforms resonate with the public and how these dynamics are captured through social media interactions.

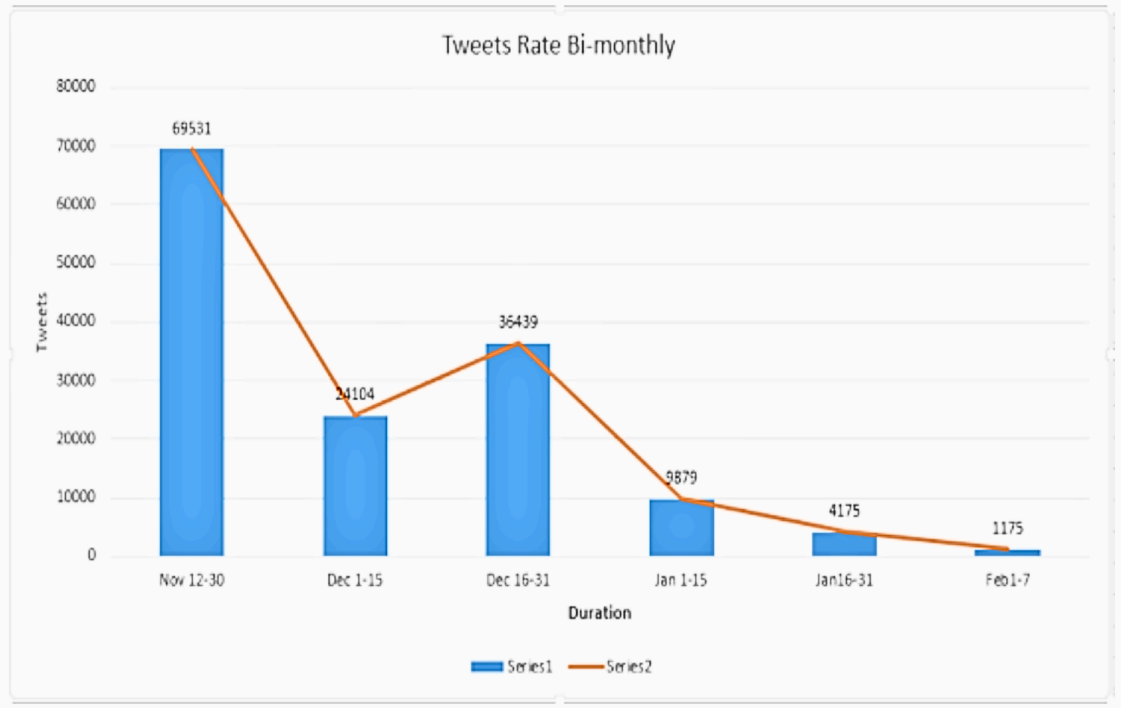


Figure 3: Tweet rate for demonetisation

The two figures clearly show that the response to demonetization was more immediate and intense, with tweet volumes sharply dropping from 69,531 in the first 15 days following the announcement to just 1,175 by the beginning of February 2017.

In contrast, the discussion around GST remained steady throughout the data collection period. These trends provide insight into public engagement levels and the breadth of discussion during these periods. Initially, demonetization sparked a lot of conversation, but interest waned over time. Meanwhile, the conversation around GST has been more consistent, possibly because it was not yet been implemented, leading people to discuss it more speculatively based on their understanding and expectations of its potential impact.

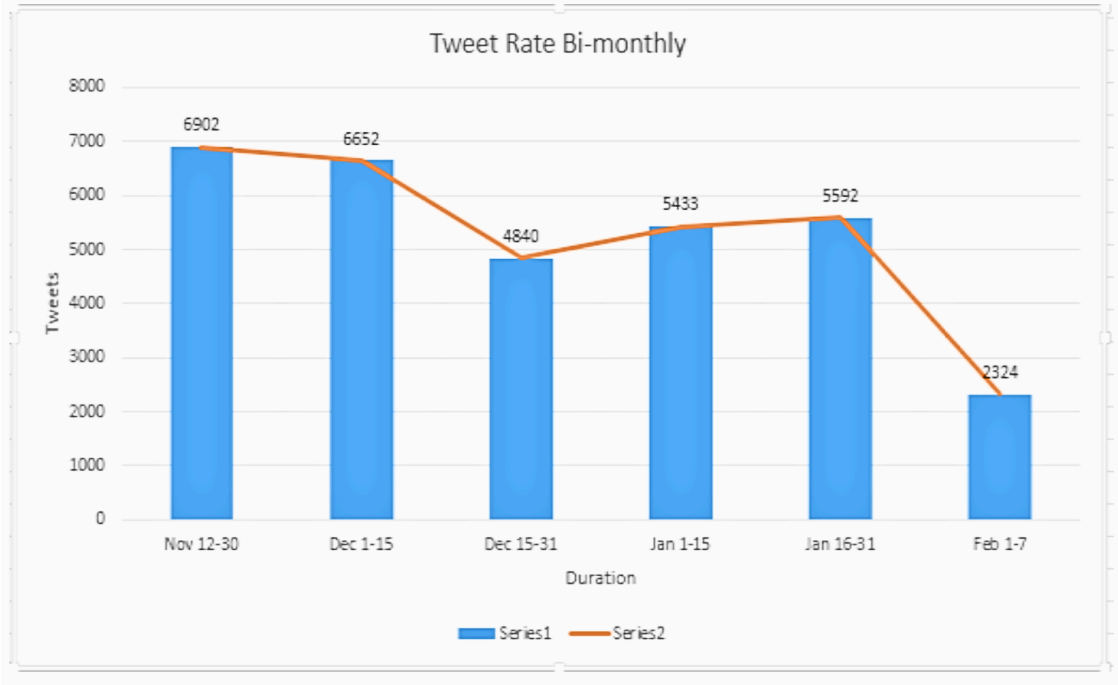


Figure 4: Tweet rate for GST

2.4.4 Descriptive Analytics: Understanding Tweet Statistics and User Engagement

2.4.4(a) Tweet analytics overview

The analysis of tweet statistics provides a quantitative measure of public engagement. For the selected period, tweets about Demonetisation totaled 159,433, whereas those about GST amounted to 33,570. Each tweet regarding Demonetization included hashtags like #Demonetization or #Demonetisation, often accompanied by other related hashtags. Similarly, GST-related tweets included the #GST hashtag and typically featured additional related hashtags. certain metrics such as tweets per user, unique tweets per user, and retweets per user provide an idea of how engaged people are with an issue. For instance, unique tweets per user were higher for demonetization than for GST, indicating broader public awareness and discussion around demonetization. However, the retweet percentages suggest that while demonetization discussions were widely shared, conversations about GST were more limited, possibly indicating a narrower reach or interest primarily among specific user groups.

2.4.4(b) Inference for Policymakers

These statistics are invaluable for policymakers for several reasons:

Scope and Scale of Engagement
The significantly higher number of tweets about Demonetisation indicates a broader and more intense public engagement compared to GST. This suggests that Demonetization affected a larger segment of the population and elicited a more substantial reaction, reflecting its widespread impact across various demographics.

Prioritization and Resource Allocation
Understanding which topics have higher levels of public interest and engagement can help in prioritizing policy issues. More engagement, as seen with Demonetization, often means a greater diversity of opinions and potentially more misinformation or 'noise.' Policymakers can use this insight to allocate resources more effectively, focusing efforts on managing and directing the conversation to ensure accurate information dissemination.

Managing Public Discourse
With limited resources, the government needs to strategically manage its focus. Data showing higher engagement on issues like Demonetization can lead to targeted interventions to shape and guide public discourse, ensuring that constructive discussions prevail over misinformation.

2.4.5 User Statistics

User statistics delve deeper into the demographics of the individuals participating in these discussions. Analysis includes:

Number of Unique Users
Identifies how many distinct individuals are discussing the topic, which can indicate the reach of the issue.

User Activity
Measures how active users are in terms of tweeting, retweeting, and engaging with content.

Influence Characteristics
Assesses which users have more influence in the discussion, based on factors like follower count, engagement rate, and content virality.

	#Demonetization	#GST
Users	61698	12722
Tweets/User	2.584087	2.638736

Table 4: Tweet rate for GST

2.4.5(a) User Profiling and Influencers

Policymakers can gain deeper insights by examining the demographics of the users involved in these discussions, such as their gender, influence level, and location. Users can be categorized into different groups based on these traits. For example[6], some users are recognized as influencers due to a high number of retweets. These influencers can be positive, spreading supportive messages, or negative, focusing on criticism. Policymakers can leverage positive influencers to enhance message reach and manage negative influencers to mitigate misinformation and address public concerns effectively.

2.4.5(b) Most Visible Users

Apart from influencers, there are users who frequently participate in discussions but may not have a wide following. These "most visible users" are often more active in conversations and can be crucial in spreading information due to their consistent presence. Analyzing the content of their tweets can help determine if their influence is positive, negative, or neutral. For example[6], in discussions on demonetization, one user might highlight benefits like reduced funding for illegal activities, while another criticizes the policy.

2.4.6 Content Analytics

Understanding the flow of conversation and the topics that capture the public's attention is crucial. This understanding is largely facilitated through the strategic analysis of hashtags, which are often used by social media users to categorize content and highlight key themes. Hashtags not only enhance the discoverability of posts but also link individual messages to larger, often global, conversations.

# Demonetization	
#Demonetization	45000
#Demonitisation	15256
#BlackMoney	1790
#India	1108
#Modi	1032
#cashless	564
#IAmWithModi	510
#BJP	449
#CashlessEconomy	441
#RBI	384

# GST	
#GST	18726
#GSTbill	1543
#DeMonetisation	1272
#GSTCouncil	857
#Budget	720
#tax	499
#GSTNews	423
#GSTIndiaExpert	401
#ArunJaitley	397
#RunSimple	391

Table 5: Tweet rate for GST

The analysis of the most frequently used hashtags enables policymakers to gauge public trends and sentiments about specific issues or policies. By identifying the key terms commonly associated with a particular topic, policymakers can more effectively tailor their communication to reach and resonate with their intended audience. For instance, when a new policy amendment is introduced, the government can enhance its communication strategy by incorporating popular hashtags related to the topic. This approach ensures wider dissemination and greater impact of the information.

A notable observation is that "Demonetisation" frequently appears as a key term alongside discussions about other significant topics, such as the GST. Several tweets highlight public concerns that demonetization might adversely affect the timely implementation of the GST, indicating a perception of economic instability. Examples of such tweets include:

- "@ShekharGupta @ndtv Would be real shame if GST gets delayed because of demonetisation. Too much pain and very little gain so far #GST"
- "#Demonetisation, #GST to transform India's business.."
- "#GST is good economics, #demonetisation is not."

These tweets reveal widespread public concern linking demonetization to challenges in implementing GST. The government needs to address these concerns with clear communication and thorough analysis to clarify the connection between demonetization and GST. Transparent responses can reduce misunderstandings and enhance informed public discussions.

2.4.7 Word Cloud Analytics

Word clouds were created to visually represent the most frequently used words related to Demonetization and GST, providing valuable insights for policymakers. These visuals can help identify key terms that resonate with the public, useful for targeted social media campaigns. For instance, the frequent appearance of "Paytm" in the Demonetization word cloud suggests significant public discussion about digital payment platforms, indicating increased digital penetration and highlighting areas for service improvement.

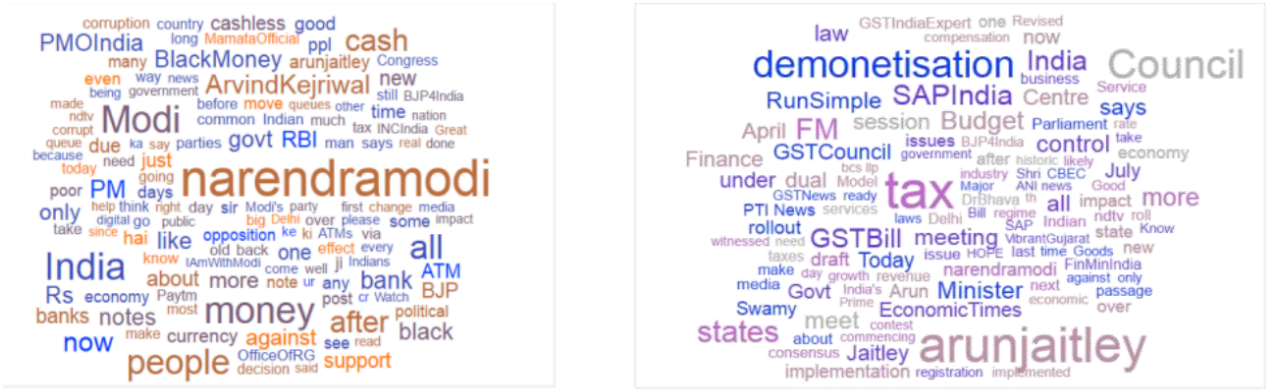


Figure 5: Word cloud for Demonetization and GST

Analyzing specific terms like "Banks" revealed public concerns such as cash shortages in ATMs, which can guide government actions to address such issues directly. Similarly, the mention of "hospitals" and issues with cash payments in medical contexts highlight areas where the government might need to ensure better compliance with new economic policies.

This method of analysis can be instrumental for the government not only to monitor public sentiment but also to swiftly respond to emerging problems, thereby enhancing service delivery and addressing public needs effectively. This proactive approach in analyzing social media data can also aid in crisis management and in preventing misinformation.

2.4.7(a) Inference for policy makers

The word cloud analysis offers policymakers valuable insights into public sentiment and the primary concerns surrounding policies like Demonetization and GST. For example, the frequent mention of "Paytm" indicates the growing relevance of digital payment systems, highlighting areas for potential service enhancement. Additionally, terms like "Banks" and "hospitals" reflect specific public issues, such as ATM cash shortages and payment difficulties in healthcare, which require government attention. By leveraging these insights, policymakers can develop targeted interventions to address public concerns, enhance communication strategies, and improve policy implementation. This approach helps in managing crises effectively, preventing misinformation, and ensuring that governmental actions are responsive to the needs of the citizenry.

2.4.8 Example

@HospitalsApollo when whole India is moving towards cashless economy your doctors are asking for cash #Demonetization

comments highlighting issues such as requests for cash payments at Apollo hospitals, even as India shifts towards a cashless economy. This situation indicates challenges in accessing healthcare services due to payment method discrepancies. To address these concerns, it is imperative for the government to closely monitor, analyze, and swiftly respond to these complaints to alleviate public distress. By utilizing this feedback, the government can enhance emergency management and crisis response capabilities. Effective management includes actively listening to the public, identifying and correcting misinformation, and taking appropriate actions to calm potentially disruptive situations, including issuing clear communications to clarify misunderstandings and dispel rumors.

2.5 The influence of Twitter on news consumption

2.5.1 Introduction

“A 2020 survey revealed that over 80 percent of respondents in India aged 16 to 70 years old used social networks as their main news outlet, along with close to 60 percent of those in Argentina and Australia and over 71 percent of Brazilian news consumers.”[7]

2.5.2 Why did we choose Twitter

This study examines the impact of using Twitter and Facebook on how much people learn about current events, considering their level of political interest. Through a panel survey with repeated measurements, the research clarifies that regular Twitter use positively influences the learning of current affairs. In contrast, frequent use of Facebook is associated with a decrease in knowledge acquisition. This decline is especially notable among individuals with less political interest, widening the knowledge gap between those who are politically engaged and those who are not.

The study explores how platforms like Facebook[8] and Twitter influence the acquisition of knowledge about current affairs. For effective learning from media, two conditions are crucial: the presence of relevant information and the audience's attention to it. Twitter meets both conditions well due to its design that emphasizes information dissemination and its use by many for news consumption, making it an effective platform for learning about current events.

On the other hand, Facebook does not necessarily facilitate the learning of current events as its structure focuses more on bi-directional social interactions and often prioritises personal content over news. Furthermore, even when users encounter news on Facebook, it does not typically lead to deeper engagement or learning. Instead, it gives users a superficial feeling of being informed, which might deter them from seeking out more substantial news sources. Therefore, while Twitter may enhance users' knowledge of current affairs, Facebook usage might actually impair it by creating a false sense of information acquisition.

2.5.3 Knowledge Curve

This analysis [8] examines the relationship between political interest, social network site usage, and the acquisition of current affairs knowledge over time, using a multilevel growth curve model with interaction terms. The findings reveal that political interest significantly impacts the level of current affairs knowledge (b=0.10, p< .001). However, when considering the interaction effects involving social media usage, the results differ between Twitter and Facebook.

For Twitter, the interaction effect between time, usage, and political interest is found to be insignificant (p= .236), suggesting that Twitter equally benefits all users, regardless of their political interest, in terms of acquiring knowledge. This uniform positive impact leads to the rejection of Hypothesis 3a, which might have suggested that political interest moderates the effect of Twitter usage on knowledge acquisition.[8]

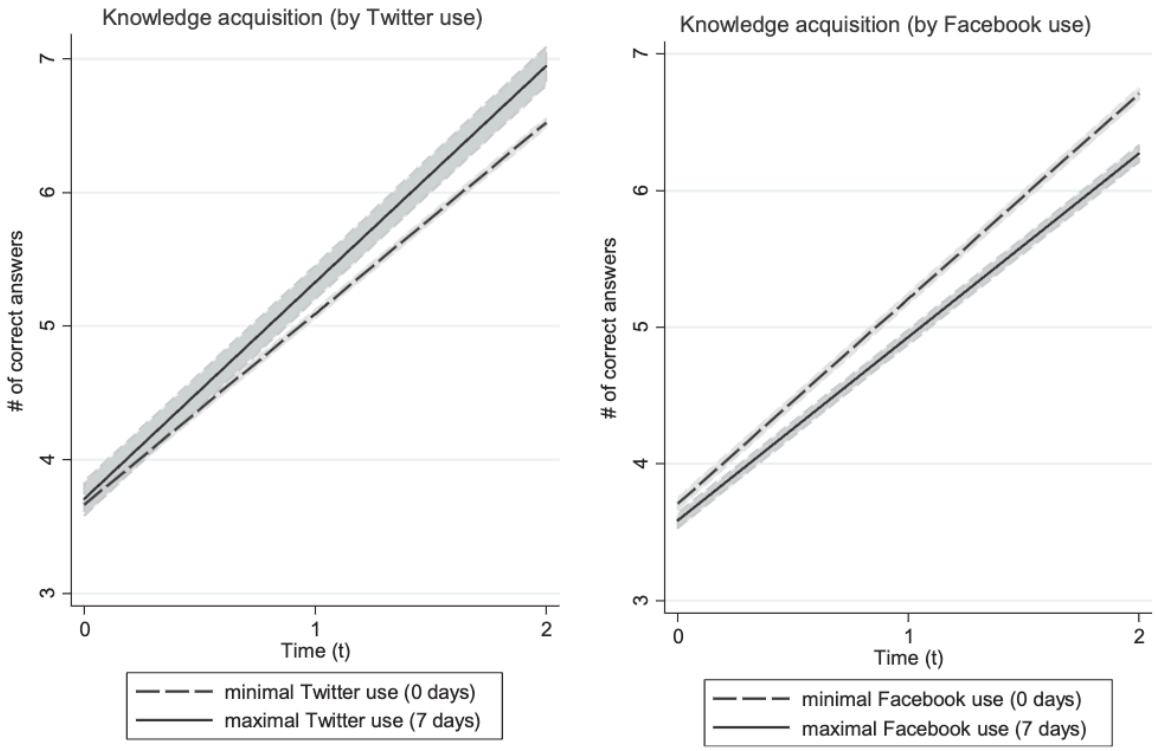


Figure 6: The average growth in current affairs knowledge over time for different levels of SNS use (on the left: Twitter; on the right:Facebook).The effect of time is shown together with its 95% confidence interval (at the mean values of other independent andcontrol variables).

In contrast, the interaction involving Facebook shows significant differences (b=0.01, p<.001). Analysis indicates that for individuals with high political interest, Facebook usage does not affect the amount of knowledge they gain about current affairs—they acquire similar levels of knowledge regardless of their usage frequency.

However, for those with less political interest, frequent Facebook usage negatively impacts their knowledge acquisition. Specifically, these users learn more when they use Facebook less frequently. This suggests that for the less politically interested, Facebook may act as a distraction from learning about current affairs rather than a facilitator.

Thus, the knowledge gap between citizens with high versus low political interest widens with increased Facebook usage, supporting Hypothesis 3b. This differential effect underscores how social network sites like Facebook can influence knowledge acquisition differently based on the user's level of political interest, potentially exacerbating existing disparities in knowledge among the population.

2.5.4 Growing user base in India

We chose Twitter for our sentiment analysis because of its extensive use as a primary source for news and current affairs. The platform's growing user base includes active participation from politicians and government agencies, making it a vital space for public discourse. Additionally, our research showed that tweets tagged with Goa government agencies garnered swift responses, highlighting Twitter's role as an effective platform for engaging with governmental stakeholders and obtaining real-time feedback. This combination of factors emphasizes Twitter's importance and growing influence in facilitating meaningful interactions and serving as a key communication channel in today's digital landscape.

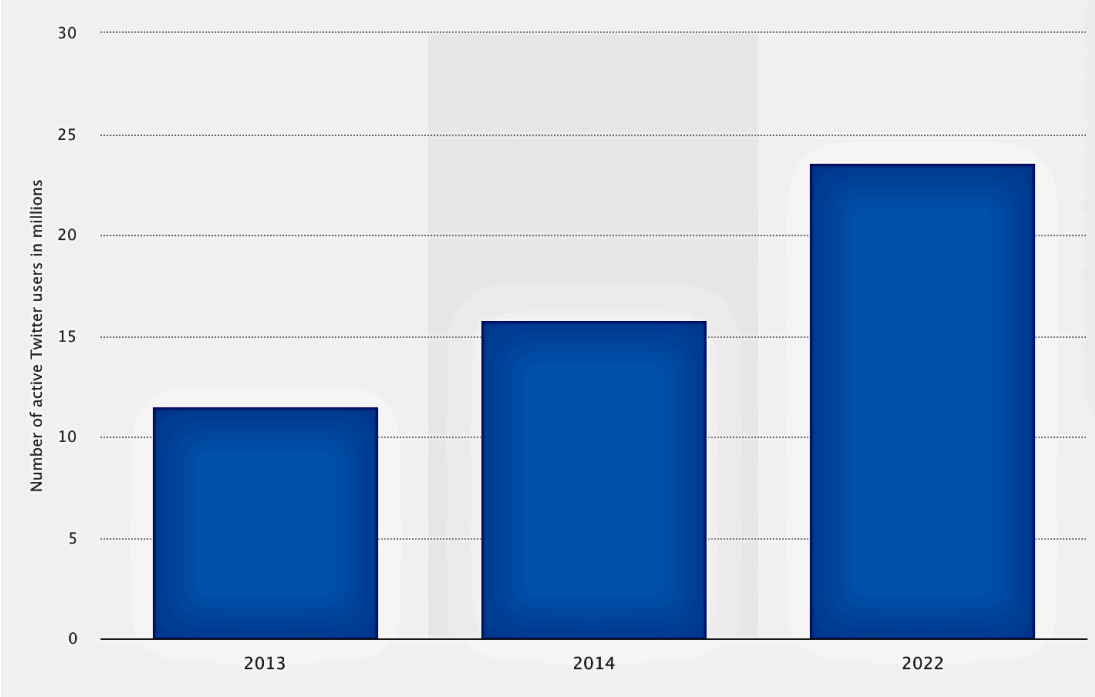


Figure 7 : Number of Twitter users in India from 2013 to 2022

People flock to Twitter for various reasons, with one of the primary motivations being to stay updated on current events. Twitter hosts vibrant discussions on a multitude of topics including politics, local and world news, entertainment, sports, technology, and health & wellness, making it a central hub for real-time information on global happenings.

2.5.5 Facts published by Twitter

People on Twitter are avid news consumers. Many of them are interested in politics and current events, and they regularly Tweet about it. [9]

- 94% of people on Twitter express interest in current events
- 85% of people on Twitter watch, read, or listen to the news at least once a day
- 83% of people on Twitter Tweet about news
- 3 in 4 people who come to Twitter for news do so at least once a day
- 55% of people on Twitter get their news from Twitter, more than other social media platforms
- 75% of people who come to Twitter for news follow news about politics and current events on Twitter
- In the first 6.5 months of 2022, there were 4.6B Tweets about news in the US (#1) and 10.4B Tweets about news globally (#2)

2.5.6 Goverment engagement

Another compelling reason for selecting Twitter was the observed effectiveness of the Goa government's engagement on the platform. Government officials were actively reading and responding to tweets from tourists, providing them with a sense of support and reassurance in emergencies. This proactive communication helped instill a feeling of safety and responsiveness, enhancing the overall experience for visitors.

2.6 Sentiment Analysis

2.6.1 Introduction

Sentiment analysis, often referred to as opinion mining, is a field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is a common practice in data science and artificial intelligence where it helps gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences.

In sentiment analysis, algorithms and models are typically applied to text data to determine whether the sentiment behind that text is positive, negative, or neutral. Advanced techniques can also detect more complex emotions such as anger, joy, or sadness. The process often involves extracting and classifying the sentiment of text data from various sources like social media posts, reviews, forums, news articles, and other internet content.

This technique is valuable for businesses and organizations as it provides insights into the consumer mindset, allowing them to tailor their strategies, products, or services to better meet their customers' needs. It also aids in identifying sentiment trends that can inform more effective marketing, customer service, and product development.

2.6.2 Word polarity

Negative polarity refers to tweets expressing emotions such as sadness, anger, fear, or disgust. Positive polarity includes tweets that convey feelings of joy and trust. Tweets that display emotions like surprise and anticipation are classified as neutral. These categories are labeled as P for Positive, N for Negative, and NTL for Neutral.

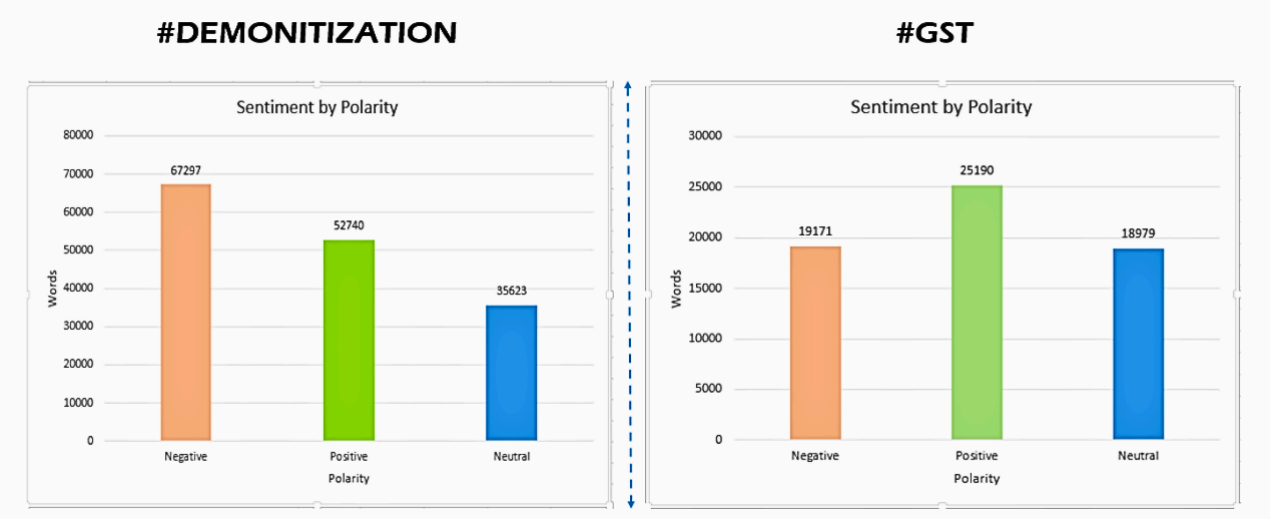


Figure 8 : Sentiment by words polarity

2.6.3 Tweet Polarity

Demonetization. • Negative tweets- 35.07% • Positive tweets- 36% • Neutral tweets- 28.9 Inference: Positive and negative tweet count by percentage is almost same suggesting that there is an ambiguity and people are not very much satisfied with the policy change.

GST. • Negative tweets- 26.98 % • Positive tweets- 40.33 % • Neutral tweets- 32.69 % Inference: Positive tweet percentage is 40.33% whereas negative tweet count by percentage is 26.98% suggesting that there is overall positive emotions among the people about the policy change/implementation. Sentiment by emotion for both the cases are depicted using Figure 9

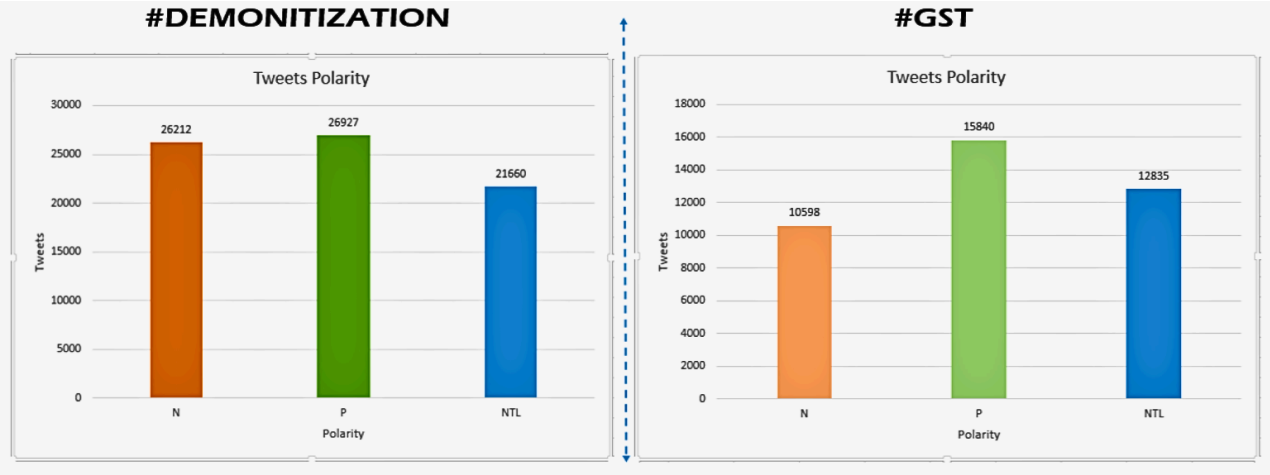


Figure 9: Sentiment by tweet polarity

2.6.4 Sentiment analysis by emotion

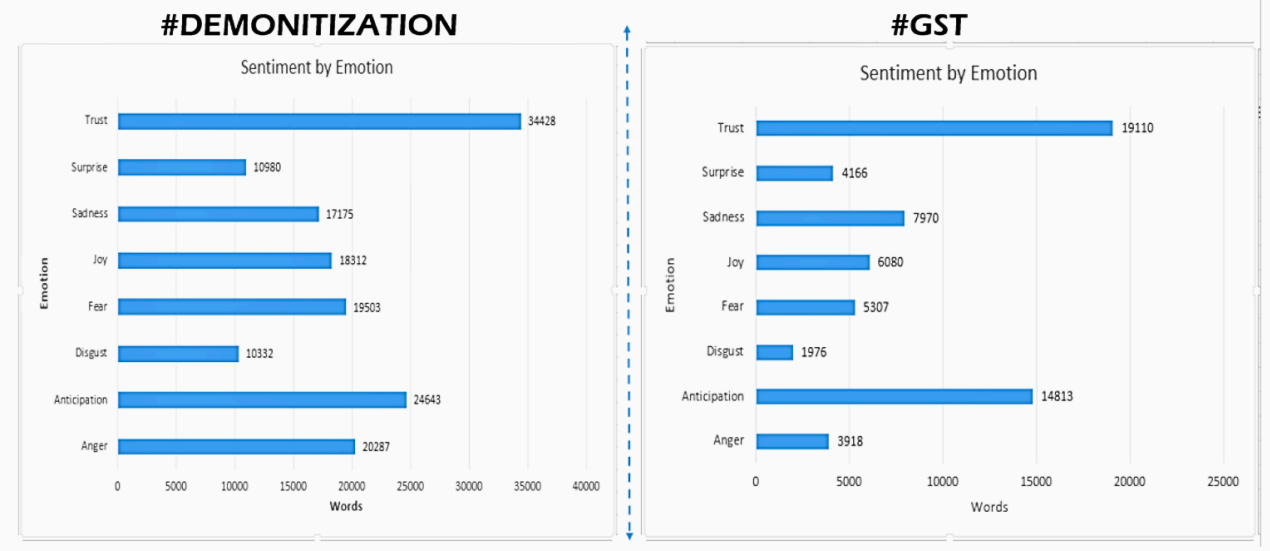


Figure 10: Sentiment Analysis by emotion

The analysis of both graphs clearly shows that the words associated with trust and anticipation are more frequent compared to other emotions. The prevalent use of trust-related words indicates public support for the policy. However, this isn't a definitive measure of sentiment as the number of negative tweets still predominates.

For example:

- Demonetization: Trust + Anticipation words total 34,428 + 24,643
- GST: Trust + Anticipation words total 19,110 + 14,813

For policymakers, sentiment analysis is a crucial tool to understand public reactions to government policies. By analyzing the emotions expressed in tweets or other content, policymakers can categorize public sentiment into groups and tailor their responses to these diverse reactions effectively. Figure 11 illustrates how sentiment analysis can be utilized in this context. [6]

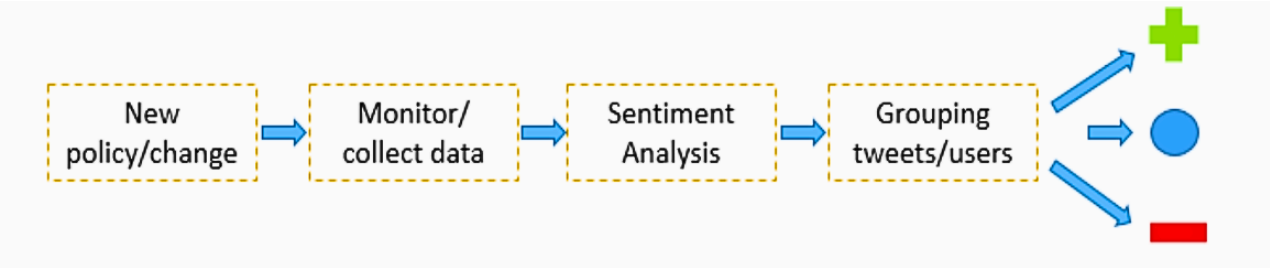


Figure 11: Sentiment analysis for post policy analysis

Some of the questions that can be very effectively answered through such analysis are:

- What do citizens feel about the new policies and initiatives?
- What are the most talked about the new policies/amendments?
- What are the most positively talked about attributes in the new policies?
- Who are advocates and skeptics of the new policies?
- Where the government should be actively listening?

2.6.5 Pre Policy Analytics

By conducting sentiment analysis on tweets related to tourism in Goa, the government can gauge overall sentiment (positive, negative, neutral) regarding various aspects of the tourist experience. For instance, if there is a prevalent negative sentiment about the cleanliness of beaches, this data can prompt targeted cleanliness initiatives or stricter enforcement of environmental regulations.

Example: If tweets frequently express dissatisfaction with beach pollution, the government might implement stricter waste management policies or initiate regular beach clean-up drives.

In addition to sentiment analysis, we can also conduct the following analyses to assess public opinion based on tweets from Twitter

2.6.5 (a) Trend Analysis

Analysing the frequency and patterns of certain keywords or hashtags (e.g., #GoaBeaches, #GoaTourism) can help identify trending issues or popular attractions. This information can guide the government in prioritizing which areas require more infrastructure support or promotional efforts.

Example: If there's a spike in tweets about a specific cultural festival in Goa, the government could support these events more robustly to enhance cultural tourism.

2.6.5 (b) Engagement and Response Strategy

Monitoring tweets allows the government to engage directly with tourists by addressing their complaints, queries, or suggestions in real-time. This direct engagement not only improves the tourist experience but also enhances the destination's image as responsive and tourist-friendly.

Example: If tourists tweet about difficulties in obtaining information about public transport, the government could respond directly with helpful information or consider improving signage and online resources.

2.6.5 (c) Policy Feedback Loop

This technique or tool can be particularly used in both the cases post policy and pre policy to gather insights**.** Tweets can serve as immediate feedback on newly implemented policies or changes. Analyzing public reaction on Twitter gives the government a rapid assessment tool to understand the efficacy and public reception of their decisions.

Example: Following a new visa-on-arrival policy, the increase in positive tweets can indicate successful reception, while criticisms can highlight areas for improvement.

2.6.5 (d) Predictive Analysis

Advanced analytics and machine learning models can be applied to predict future trends in tourism demand based on tweet patterns. This can aid in strategic planning for peak tourist seasons, helping to allocate resources more efficiently.

Example: Predictive analytics might forecast a high demand for certain types of tourism (e.g., eco-tourism) allowing the government to focus on enhancing the related infrastructure.

2.6.6 Trend analysis using hashtag

In our research, we were unable to perform trend analysis using hashtags for specific tourist destinations although we have generated an overall graph for all the tweets. For effective hashtag analysis, a substantial volume of tweets discussing a specific topic is necessary, which we did not have due to the narrower scope of our study.

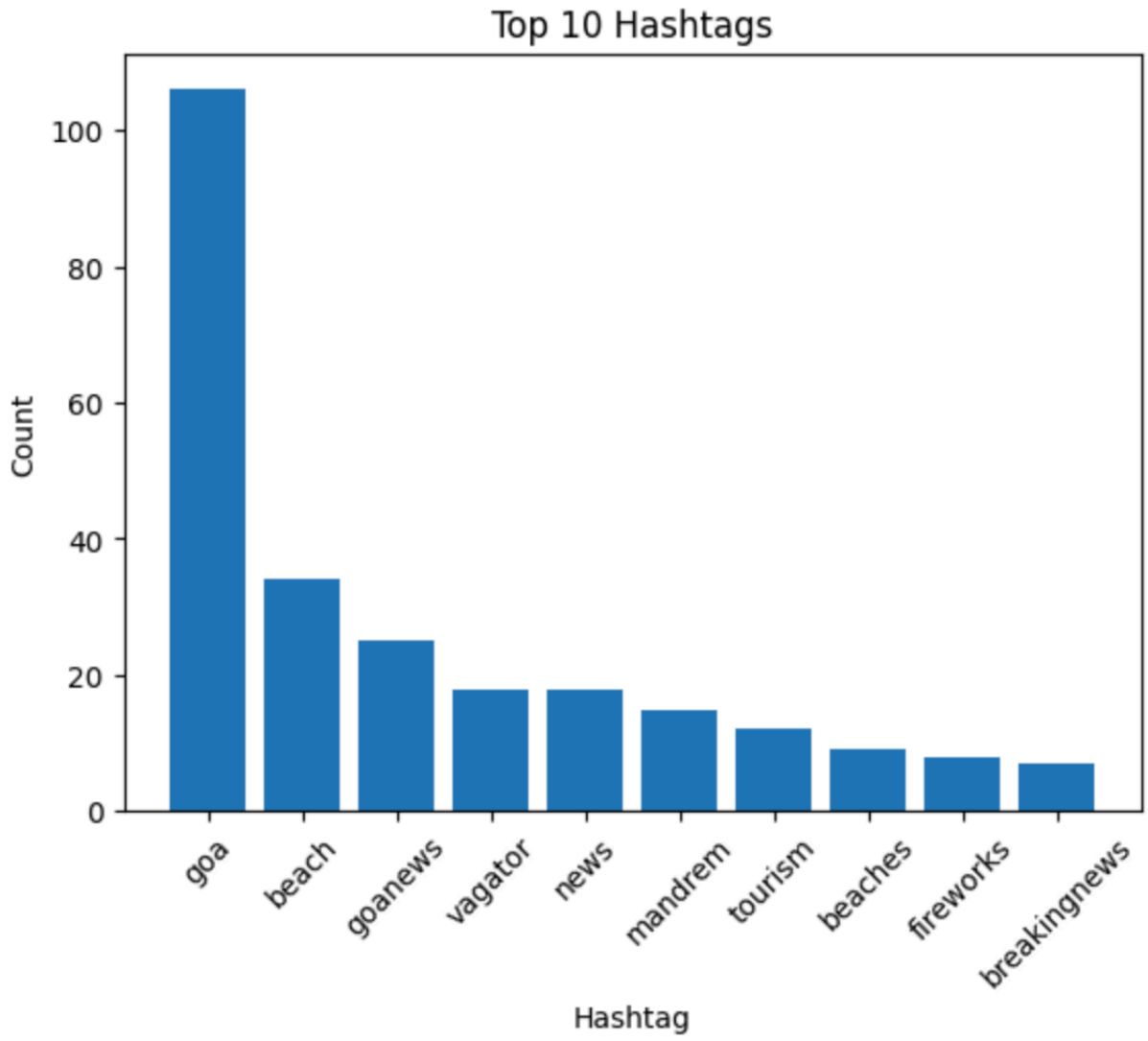


Figure 12 : Graph plotted for top 10 hashtags for the entire dataset of tweets

2.6.7 Word Cloud

A word cloud analysis can be beneficial as it allows us to determine the frequency of specific words in tweets, particularly when examining posts about a specific beach or a certain tourist destination.

Figure 13 : An overall word cloud for all the tweets

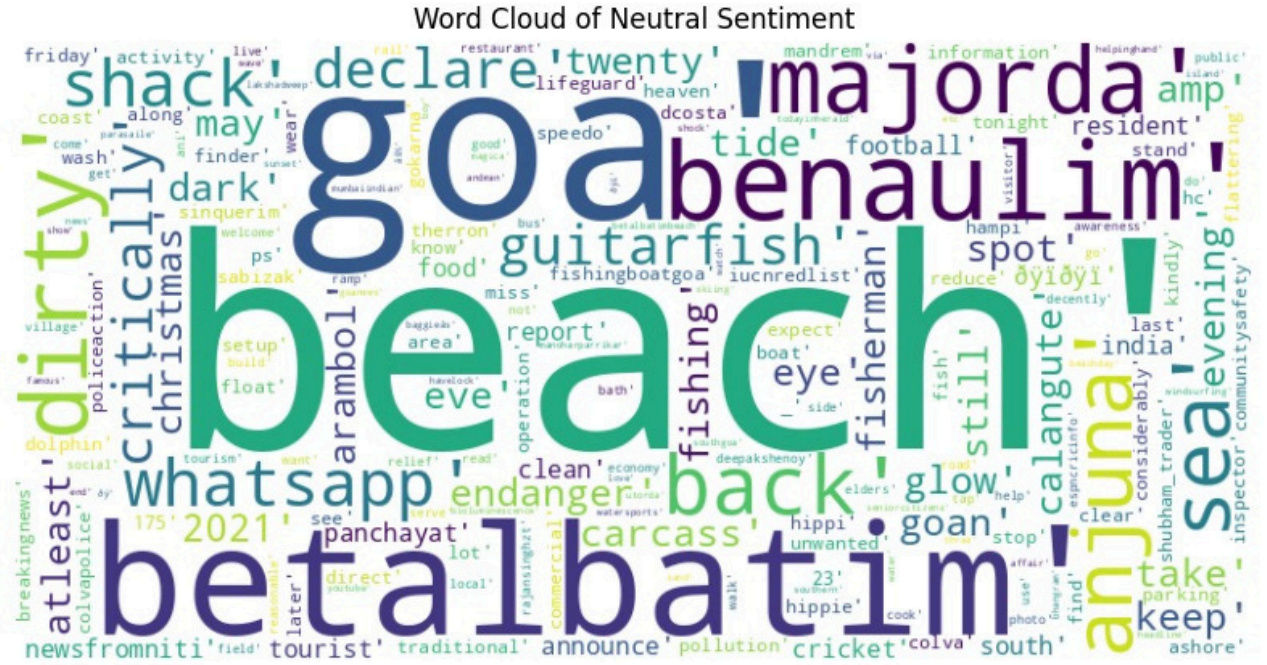


Figure 14 : An overall word cloud for all the neutral tweets

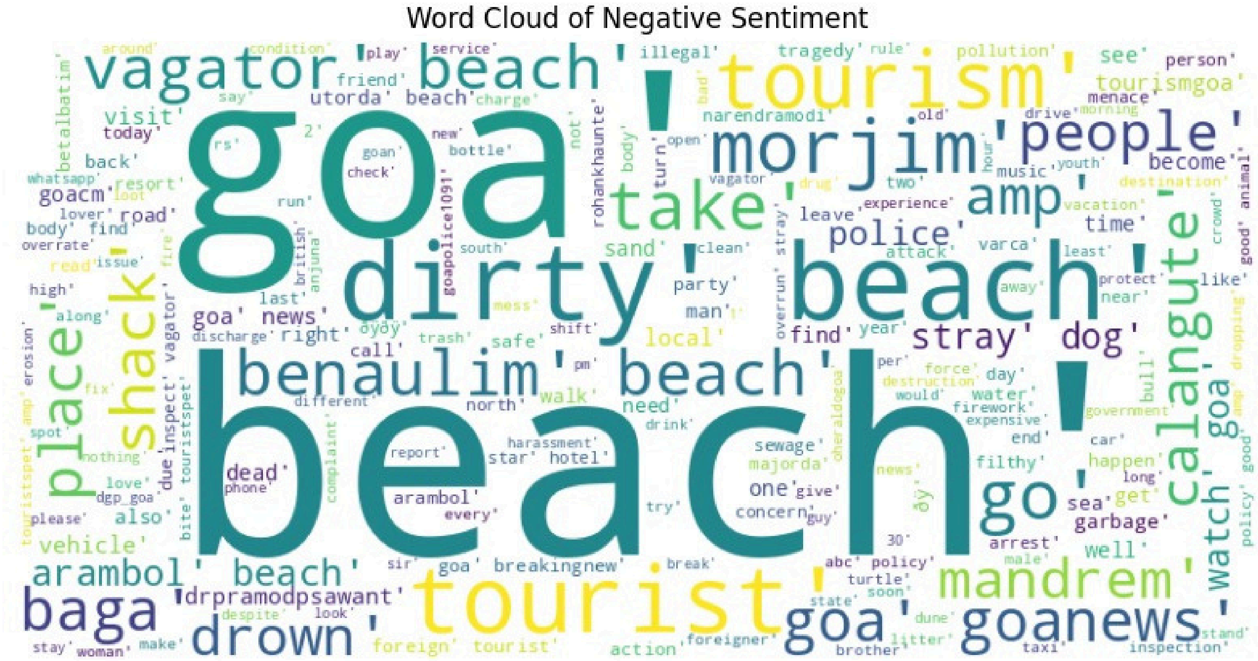


Figure 15 : An overall word cloud of all Negetive sentiments

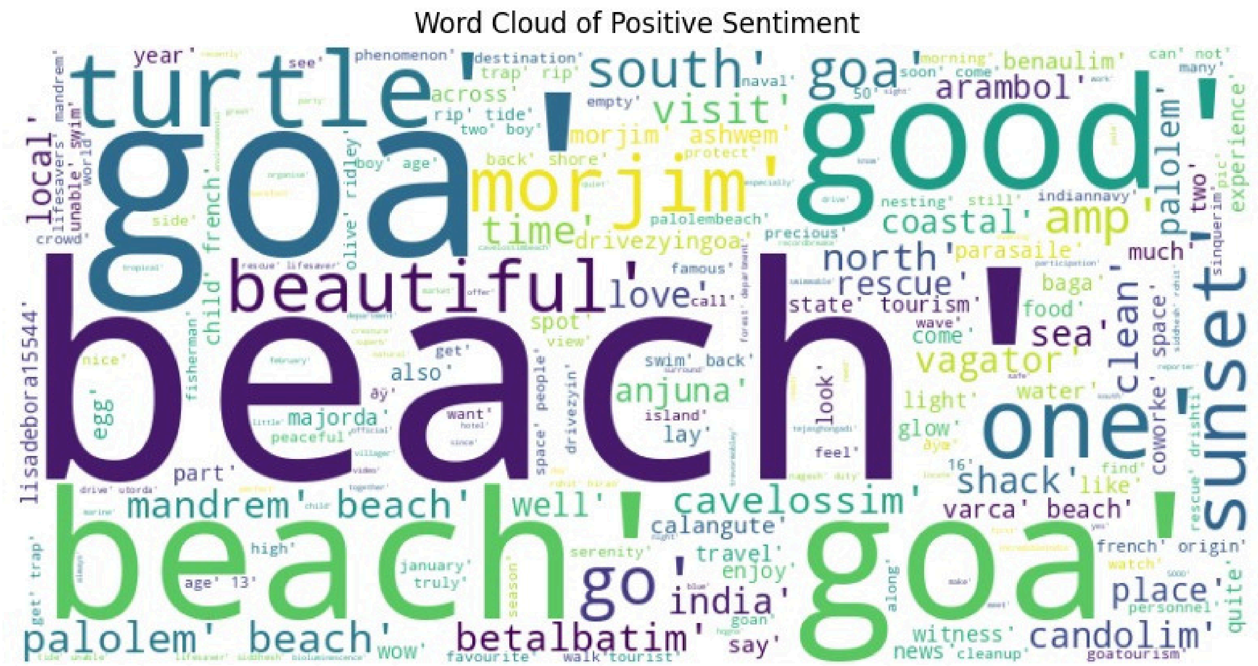


Figure 16 : An overall word cloud for all the Positive sentiments

3. METHODOLOGY

3.1 Approach

3.1.1 Survey

We conducted a survey that garnered over 500 responses, aimed at understanding public perceptions of Twitter as a source for news and current affairs. This survey confirmed that residents of Goa actively use Twitter for these purposes and generally view the platform as a reliable and influential source for information.

Goa Tourism Twitter Insights Survey
QUESTIONS AND PURPOSE

Section 1: Demographic Information
Collects basic demographic data such as age, gender and duration of residency in Goa.

- What is your age?
- What is your gender?
- Are you a Goan/Non-Goan?
- How long have you been residing in Goa?
- Do you have an active twitter account?

Section 2: Twitter Usage for Tourism Updates
Explores respondents' frequency and preferences in using Twitter for staying informed about tourism-related news and events in Goa and Investigates if respondents engage in sharing or retweeting tourism-related content about Goa on Twitter.

- How frequently do you use Twitter to stay updated on tourism-related news and events in Goa?
- *Purpose: This question aims to gauge the frequency of Twitter usage specifically for tracking tourism updates in Goa.*

Goa Tourism Twitter Insights Survey
QUESTIONS AND PURPOSE

- What types of tourism-related information do you typically seek on Twitter? (e.g., events, promotions, safety advisories, local attractions)
- *Purpose: This question helps to understand the specific types of tourism-related content that respondents are interested in accessing via Twitter.*

- Have you ever shared or retweeted tourism-related content about Goa on Twitter? If yes, please specify the types of content you shared.
- *Purpose: This question explores whether respondents actively engage with and share tourism-related content on Twitter, indicating their level of involvement and interest in promoting Goa as a tourist destination.*

Section 3: Perceptions of Twitter Influence
Explores the extent to which respondents believe positive or negative tweets about Goa impact tourists' decisions to visit the region.

- How influential do you perceive Twitter discussions and opinions to be in shaping the perception of Goa as a tourist destination?
- *Purpose: This question aims to assess the perceived impact of Twitter discussions and opinions on the reputation and appeal of Goa among both locals and tourists.*

- To what extent do you believe that positive or negative tweets about Goa influence tourists' decisions to visit the region?
- *Purpose: This question aims to assess the perceived impact of Twitter discussions and opinions on tourists' decisions to visit Goa, providing insights into the potential influence of positive and negative sentiment expressed on Twitter on tourist behaviour.*

- Do you think addressing concerns raised on Twitter could improve the image of Goa as a tourist destination? Why or why not?
- *Purpose: This question investigates whether respondents believe addressing Twitter-raised concerns could enhance Goa's tourism image, aiming to identify strategies for improving its perception as a tourist destination.*

Goa Tourism Twitter Insights Survey
QUESTIONS AND PURPOSE

Section 4: Barriers to Twitter Usage / Perception in terms of reliable information

- Assesses perceptions of the availability of trustworthy and reliable tourism-related information about Goa on Twitter.

- Do you feel that there is enough trustworthy and reliable tourism-related information available on Twitter about Goa?
- *Purpose: This question evaluates the perceived adequacy of reliable tourism information about Goa on Twitter, informing strategies to enhance credibility and accessibility of such information on the platform.*

Section 5: General Feedback and Suggestions

- Do you feel that Goa's tourism is on a decline lately?
- *Purpose: This question seeks to gauge participants' perceptions regarding the current state of Goa's tourism industry. By assessing whether respondents believe that tourism in Goa is experiencing a decline.*

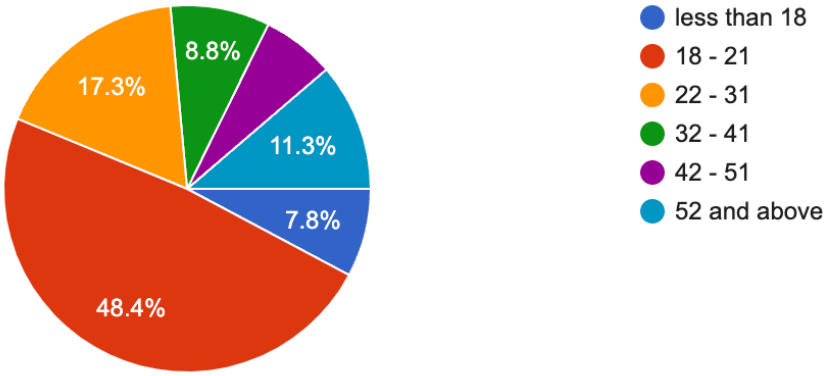
- Overall on a scale of 5, how satisfied are you with the quality and relevance of tourism-related information available on Twitter about Goa?
- (1 being Poor and 5 being Excellent)
- *Purpose:To assess respondents' satisfaction with the quality and relevance of tourism-related information about Goa on Twitter.*

3.1.2 Survey Results

(500+ Responses)

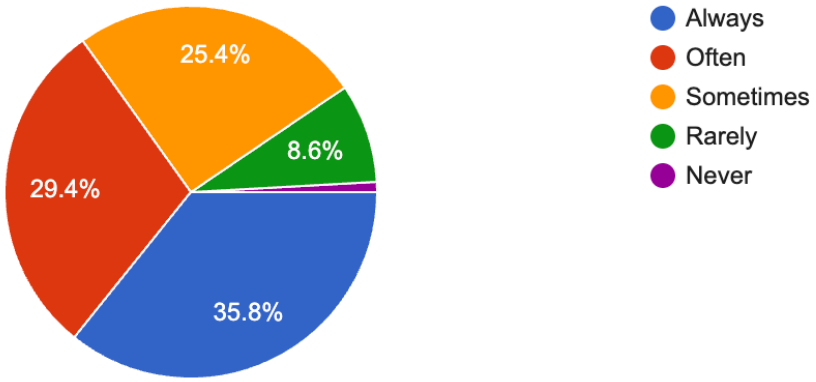
What is your age?

514 responses



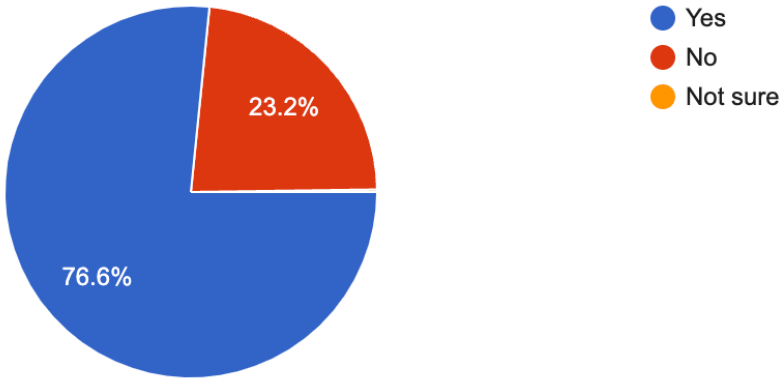
How frequently do you use Twitter to stay updated on tourism-related news and events in Goa?

453 responses



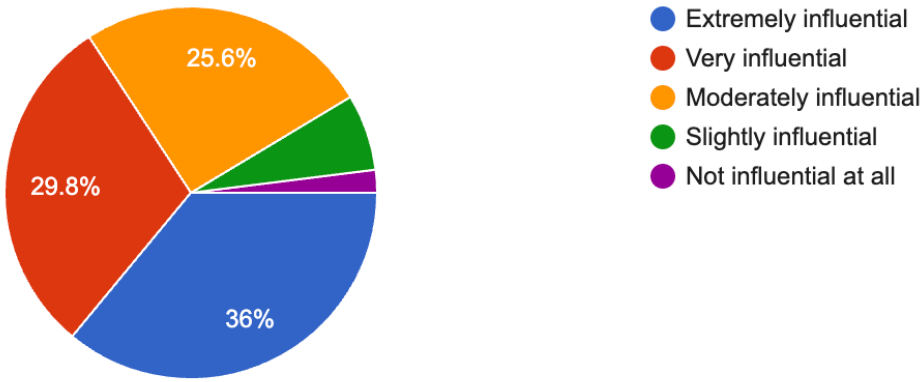
Have you ever shared or retweeted tourism-related content about Goa on Twitter?

453 responses



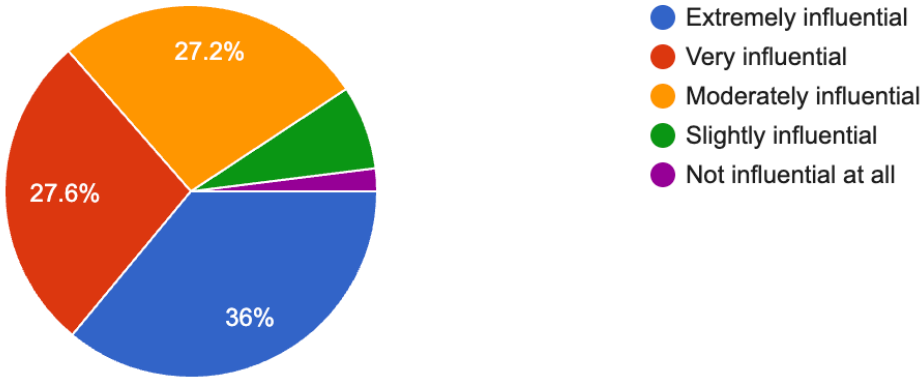
To what extent do you believe that positive or negative tweets about Goa influence tourists' decisions to visit the region?

453 responses



How influential do you perceive Twitter discussions and opinions to be, in shaping the perception of Goa as a tourist destination?

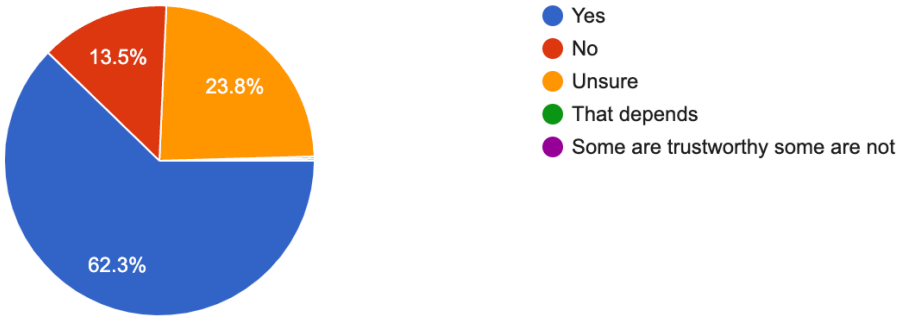
453 responses



Do you feel that there are enough trustworthy and reliable tourism-related information available on Twitter about Goa?

Copy

453 responses



3.2 Scraping/ Obtaining Data

3.2.1 APIFY

Apify [18] is a robust platform designed for web scraping, data extraction, and automation tasks. It offers a wide range of features tailored for developers and mid-sized organizations looking to streamline their data collection processes.

Some Key Features:

- Apify Store: Access a vast collection of over 1,600 pre-built scrapers covering sources like social media platforms, Google services, and more.
- Open-source tools: Utilize Crawlee and Apify Python SDK for flexible scraper development, supporting both Python and JavaScript languages.
- Actor deployment: Easily create, deploy, share, and integrate serverless micro apps (Actors) for various scraping needs.
- Proxy management: Benefit from a variety of proxies, including data center and residential proxies, with intelligent IP rotation to improve user experience and mitigate blocking risks.
- Data management: Seamlessly store and share crawling results in formats like Excel, CSV, JSON, etc., ensuring efficient data handling.
- Performance monitoring: Keep track of Actor performance, inspect runs, logs, and runtime costs, with automated alerts for efficient monitoring.
- Workflow integration: Integrate Apify with different workflows through ready-made integrations and custom workflows using webhooks and APIs.

In our project we used Apify's Tweet Scraper V2 (Pay Per Result) - X / Twitter Scraper Actor [19]. This actor is developed by API Dojo. This actor allows you to scrape data by using twitter handles, keywords, IDs, URLs, etc., with options to filter based on timeline, geolocation, and other criteria. You can also specify the minimum and maximum number of tweets to scrape, tailoring the process to your specific needs.

Overall, Apify provides a comprehensive solution for web scraping and data extraction tasks, catering to both experienced developers and organizations looking to streamline their data collection efforts. While it may require some learning for beginners and could pose pricing challenges for smaller users, its powerful features make it a valuable tool for various scraping projects.

3.2.2 Twitter API

The Twitter API allows users to programmatically interact with Twitter's platform and data. It provides endpoints and methods for reading tweets, writing tweets, accessing user information, sending direct messages, retrieving trends and location data, uploading media, and receiving real-time tweet streams.

One needs to register for a Twitter Developer Account and create an application to obtain API credentials. The API supports various response formats like JSON and has rate limits and usage policies.

The API enables developers to build applications that integrate with Twitter, such as social media monitoring tools, data analysis apps, bots, and automated tweet publishing systems.

Abbreviations used

A1 - Initial dataset (155 keywords)
A2 - Beach related dataset (36 keywords)
A3 - Beach related but more focused dataset (only certain beaches) (17 keywords)
A4 - Final dataset (Sample taken of A3)

V1 - First cleaned variation of A1
V2 - Second cleaned variation of A1
V3 - Third cleaned variation of A1
V4 - Fourth cleaned variation of A1
V5 - Fifth cleaned variation of A1

D1 - First cleaned variation of A4
D2 - Second cleaned variation of A4
D3 - Third cleaned variation of A4
D4 - Fourth cleaned variation of A4
D5 - Fifth cleaned variation of A4

3.3 Data Collection

3.3.1 Initial challenges

At first, when we started looking for labeled data, we couldn't find any datasets because our topic was very specific - we wanted tweets related to tourism in Goa.

Then, we searched through different sources to find data on this topic.

When we tried web scraping with ntscrapper, we hit a rate limit and couldn't get any tweets. Next, we attempted snscrape, but our connection requests didn't work and we couldn't retrieve data due to errors. Then, we experimented with tweepy, but we lacked the necessary credentials.

Finally, we turned to APIFY and managed to successfully gather the data we needed.

We also tried using Twitter's API and purchased their basic plan for a month. But, with this plan we could only retrieve 10K tweets in a months period. Another limitation of this is that the data retrieved from the API was only of a week back from when it was pulled.

So it was then decided to stick to APIFY itself to gather data.

3.3.2 Keyword based approach

Following engagements with stakeholders and a comprehensive study of the tourism sector in Goa, a corpus of 150+ keywords was meticulously curated (mentioned in table 10). These keywords were strategically employed to conduct targeted scraping of tweets from Twitter.

Goa Tourism	Goa Leisure	Goa Coastal Tourism	Goa Local Experiences
Goa Tourist Attractions	Goa Experience	Goa Island Hopping	Goa Culinary Tourism
Goa Sight Seeing	Goa Retreat	Goa Nature Reserve	Goa Food Tour
Goa Adventure	Goa Journey	Goa Wildlife Sanctuary	Goa Wine Tasting
Goa Travel	Goa Itinerary	Goa Bird Watching	Goa Handicraft Shopping
Goa Destination	Goa Cultural Heritage	Goa Scuba Diving	Goa Cultural Performances
Goa Vacation	Goa Nature Tourism	Goa Snorkeling	Goa Music Festivals
Goa Exploration	Goa Eco Tourism	Goa Water Sports	Goa Dance Festivals
Goa Wanderlust	Goa Sustainable Tourism	Goa Surfing	Goa Beach
Goa Excursion	Goa Travelogue	Goa Parasailing	Goa Sunset
Goa Holiday	Goa Trip Planning	Goa Adventure Activities	Goa Mountain
Goa Getaway	Goa Beach Resort	Goa Heritage Sites	Goa Church
Goa Discovery	Goa Beach Vacation	Goa Cultural Immersion	Goa Cultural Festival

Table 10: Keywords for initial iteration (for A1)

Goa Fort	Goa Virtual Meeting	Goa Optimism	Goa Flight Delay
Goa Waterfalls	Goa Hot Cocoa	Goa Gratitude	Goa Inner Peace
Goa Temples	Goa Book	Goa Friendship Reunion	Goa Pet Love
Goa Dance Workshop	Goa Rainy Day	Goa Work From Home	Goa Power Outage
Goa Caves	Goa Traffic Jam	Goa Cozy Nights	Goa Bookworm Joy
Goa Music Concert	Goa Cooking	Goa Personal Development	Goa Umbrella
Goa DIY Project	Goa Winter Evenings	Goa Achievement	Goa Community Love
Goa Road Trip Output	Goa Movie Night	Goa Digital Detox	Goa Busy Life
Goa Childhood	Goa Fitness Journey	Goa Boredom	Goa Friday Evening
Goa Park	Goa Weekend Getaway	Goa Skill Development	Goa Culture Lost
Goa Nature Walk	Goa Nostalgia	Goa Customer Experience	Goa Nature Relief
Goa Workout	Goa Tech Issues	Goa Surprise Gift	Responsible Goa
Goa Restaurant	Goa Presentation	Goa Sick Day	Goa Support Local

Table 10: Keywords for initial iteration (for A1)

Goa Shacks	Goa Agonda Beach	Goa Ashwem Beach	Goa Rent a Car
Clean Goa	Goa Baga Beach	Goa Majorda Beach	Goa Rent a Cab
Goa Tourist Problems	Goa Cavelossim Beach	Goa Sinquerim Beach	Panaji Smart City
Goa Food	Goa Varca Beach	Goa Anjuna Beach	Goa Pollution
Goa Culinary	Goa Benaulim Beach	Goa Betalbatim Beach	Goa Expensive
Goa Alcohol	Goa Candolim Beach	Goa Dona Paula Beach	Panjim Pay Parking
Goa Taxis	Goa Morjim Beach	Goa Dona Paula Jetty	Goa Roads
Goa Rentals	Goa Calangute Beach	Goa Miramar Beach	Goa Irregular Water Supply
Goa Ghats	Goa Colva Beach	Goa Butterfly Beach	Goa Electricity Cut
Goa Accidents	Goa Mandrem Beach	Goa Bambolim Beach	Goa Power Cut
Goa Murders	Goa Vagator Beach	Goa Casinos	Goa Irresponsible
Goa Shigmo	Goa Cola Beach	Goa Mandovi River	Goa Unsafe
Goa Palolem	Goa Utorda Beach	Goa Mhadei River	Goa Arambol Beach

Table 10: Keywords for initial iteration (for A1)

3.1.3 Data Collected with APIFY

With APIFY's Tweet Scraper V2 (Pay Per Result) - X / Twitter Scraper Actor, tweets were collected. The scraping was done using specific keywords (mentioned in table 10), a timeline from 2014 to 2024, and targeting tweets in English language.

On average, there were 100 tweets per keyword. The scrapped data for each keyword was then downloaded as JSON format files. All these JSON format files were then converted to CSV format files in batches and then appended together to form the master dataset, that is dataset A1. A1 contained 18,943 tweets. While converting the obtained files from JSON format to CSV format no changes were made in the dataset. The attributes that were present in the dataset are mentioned in table 11.

Attribute Name	Description
type	Indicates that its a tweet
id	The unique identifier for the tweet
url	This URL is a shortened link. It redirects to the original tweet on Twitter.
twitter url	This is the direct link to the tweet on the Twitter platform itself.
text	The actual text content of the tweet.
retweetCount	The number of times the tweet has been retweeted.
replyCount	The number of replies the tweet has received.
likeCount	The number of likes the tweet has received.
quoteCount	The number of times the tweet has been quoted.

Table 11: Attributes in all the Datasets

Attribute Name	Description
createdAt	The date and time when the tweet was created.
lang	The language of the tweet.
bookmarkCount	The number of times the tweet has been bookmarked by users.
isReply	Indicates whether the tweet is a reply to another tweet.
author	The user who authored the tweet.
extendedEntities	Additional entities such as images or videos associated with the tweet.
entities	Entities like hashtags, user mentions, URLs, etc. included in the tweet.
isRetweet	Indicates whether the tweet is a retweet of another tweet.
isQuote	Indicates whether the tweet is a quote tweet of another tweet.
media	Media content (images, videos, etc.) included in the tweet.
isConversationControlled	Indicates whether the conversation around the tweet is controlled (e.g., limited replies).
viewCount	The number of views the tweet has received.
inReplyToId	The ID of the tweet to which this tweet is replying (if applicable).

Table 11: Attributes in all the Datasets

Attribute Name	Description
quoteId	The ID of the tweet that this tweet is quoting (if applicable).
quote	The text content of the tweet being quoted (if applicable).

Table 11: Attributes in all the Datasets

Data Collected with the Twitter API was found to be a subset of the data we collected with the scrapper i.e., the Actor on APIFY. So, only data collected with APIFY was taken.

We initially focused on the entire tourism industry in Goa but later narrowed our scope to a specific area: beaches. We believed that if we could succeed in one segment, we could replicate that success in others.

To gather data, we selected 36 keywords related to beaches. Individual files per keyword were already present (done before). These individual CSV format files were appended and a new master dataset was created, dataset A2. The keywords are mentioned in table 12**.**

Goa Beach Resort	Goa Water Sports	Goa Sunset	Goa Baga Beach
Goa Beach Vacation	Goa Surfing	Goa Shacks	Goa Cavelossim Beach
Goa Scuba Diving	Goa Parasailing	Goa Palolem	Goa Varca Beach
Goa Snorkeling	Goa Beach	Goa Agonda Beach	Goa Benaulim Beach
Goa Arambol Beach	Goa Colva Beach	Goa Utorda Beach	Goa Anjuna Beach

Table 12: Keywords for beach related data (for A2)

Goa Candolim Beach	Goa Mandrem Beach	Goa Ashwem Beach	Goa Betalbatim Beach
Goa Morjim Beach	Goa Vagator Beach	Goa Majorda Beach	Goa Dona Paula Beach
Goa Calangute Beach	Goa Cola Beach	Goa Sinquerim Beach	Goa Miramar Beach
Goa Butterfly Beach	Goa Bambolim Beach	Clean Beaches Goa	Dirty Beaches Goa

Table 12: Keywords for beach related data (for A2)

Again it was observed that data collected with the Twitter API was found to be a subset of the data we collected with the scrapper i.e., the Actor on APIFY. So, only data collected with APIFY was taken.

The attributes that were present in the dataset were the same as before (mentioned in table 11).

Dataset A2 had 3782 tweets.

From A2, advertisements were filtered out and only the keywords mentioned in table 13 (17 keywords) were retained. This was done to narrow it down and obtain specific analysis. Dataset A3 was created in this manner.

In dataset A3, 600 tweets were present.

A random sample of A3 was taken and 50% tweets were randomly selected using sample from the random library in python. This resulted in the generation of Dataset A4.

So our final dataset, dataset A4 consisted of 300 tweets.

Dirty Beaches Goa	Goa Arambol Beach	Goa Palolem	Goa Majorda Beach
Goa Morjim Beach	Goa Betalbatim Beach	Goa Utorda Beach	Goa Varca Beach
Goa Benaulim Beach	Goa Mandrem Beach	Goa Baga Beach	Goa Cavelossim Beach
Goa Vagator Beach	Goa Calangute Beach	Goa Anjuna Beach	Goa Candolim Beach
Goa Sinquerim Beach			

Table 13: Keywords for final iteration (for A3 and A4)

3.4 Data Preparation

3.4.1 Dataset A1

The attributes of A1 are already mentioned in table 10. The attributes with their respective datatypes are mentioned in table 14.

Attribute Name	Data Type	Attribute Name	Data Type
type	object	bookmarkCount	int64
id	int64	isReply	bool
url	object	author	object
twitter url	object	extendedEntities	object
text	object	entities	object
retweetCount	int64	isRetweet	bool
replyCount	int64	isQuote	bool
likeCount	int64	media	object
quoteCount	int64	isConversationControlled	bool
viewCount	float64	inReplyToId	float64
createdAt	object	quoteId	float64
lang	object	quote	object

Table 14: Attributes in all the Datasets (A1, A2, A3, A4) with their datatypes

Table 15 contains the preparation done on dataset A1.

5 cleaned variations were made.

Libraries such as TextBlock, Spello, Autocorrect and Spellchecker were used to correct the spellings of the tweets in the column - text, but we weren't successful with the same as it was taking too long to traverse through the entire dataset.

All the other preparation procedures followed have been mentioned in table 15

Sr. no.	**Cleaned Dataset Variations of A1**	**Preparation done - cleaning and transformation**
1	V1	• Columns - type, url, isReply, author, extendedEntities, entities, isRetweet, isQuote, media, isConversationControlled, inReplyTold, quoteId and quote are dropped using drop from pandas. • Retained columns - id, twitterUrl, text, retweetCount, replyCount, likeCount, quoteCount, viewCount, createdAt, lang and bookmarkCount. • Missing values checked for, using missingno from matplotlib. • Columns containing the same include - viewCount, inReplyTold, quoteId and quote. • viewCount dropped. • Duplicates removed using drop_duplicates from pandas. • URLs removed using re library. • All the tweets are in english, crosschecked using detect from the langdetect library.

Table 15: Data Preparation done on Dataset A1 to obtain different datasets with variations.

Sr. no.	Cleaned Dataset Variations of A1	Preparation done - cleaning and transformation
2	V2	• Columns - type, url, isReply, author, extendedEntities, entities, isRetweet, isQuote, media, isConversationControlled, inReplyTold, quoteId and quote are dropped using drop from pandas. • Retained columns - id, twitterUrl, text, retweetCount, replyCount, likeCount, quoteCount, viewCount, createdAt, lang and bookmarkCount. • Missing values checked for, using missingno from matplotlib. Columns containing the same include - viewCount, inReplyTold, quoteId and quote. • viewCount dropped. • Duplicates removed using drop_duplicates from pandas. • URLs removed using re library. • All the tweets are in english, crosschecked using detect from the langdetect library. • Punctuations removed using the re library. • Text converted to lowercase using str.lower.

Table 15: Data Preparation done on Dataset A1 to obtain different datasets with variations.

Sr. no.	Cleaned Dataset Variations of A1	Preparation done - cleaning and transformation
3	V3	• Columns - type, url, isReply, author, extendedEntities, entities, isRetweet, isQuote, media, isConversationControlled, inReplyTold, quoteId and quote are dropped using drop from pandas. • Retained columns - id, twitterUrl, text, retweetCount, replyCount, likeCount, quoteCount, viewCount, createdAt, lang and bookmarkCount. • Missing values checked for, using missingno from matplotlib. Columns containing the same include - viewCount, inReplyTold, quoteId and quote. • viewCount dropped. • Duplicates removed using drop_duplicates from pandas. • URLs removed using re library. • All the tweets are in english, crosschecked using detect from the langdetect library. • Punctuations removed using the re library. • Text converted to lowercase using str.lower. • Emojis converted to their text equivalent using the demoji library. • Unicode normalization performed using the unicodedata library.

Table 15: Data Preparation done on Dataset A1 to obtain different datasets with variations.

Sr. no.	Cleaned Dataset Variations of A1	Preparation done - cleaning and transformation
4	V4	• Columns - type, url, isReply, author, extendedEntities, entities, isRetweet, isQuote, media, isConversationControlled, inReplyTold, quoteId and quote are dropped using drop from pandas. • Retained columns - id, twitterUrl, text, retweetCount, replyCount, likeCount, quoteCount, viewCount, createdAt, lang and bookmarkCount. • Missing values checked for, using missingno from matplotlib. Columns containing the same include - viewCount, inReplyTold, quoteId and quote. • viewCount dropped. • Duplicates removed using drop_duplicates from pandas. • URLs removed using re library. • All the tweets are in english, crosschecked using detect from the langdetect library. • Punctuations removed using the re library. • Text converted to lowercase using str.lower. • Emojis converted to their text equivalent using the demoji library. • Unicode normalization performed using the unicodedata library. • Words tokenized using split from pandas.

Table 15: Data Preparation done on Dataset A1 to obtain different datasets with variations.

Sr. no.	Cleaned Dataset Variations of A1	Preparation done - cleaning and transformation
5	V5	• Columns - type, url, isReply, author, extendedEntities, entities, isRetweet, isQuote, media, isConversationControlled, inReplyTold, quoteId and quote are dropped using drop from pandas. • Retained columns - id, twitterUrl, text, retweetCount, replyCount, likeCount, quoteCount, viewCount, createdAt, lang and bookmarkCount. • Missing values checked for, using missingno from matplotlib. Columns containing the same include - viewCount, inReplyTold, quoteId and quote. viewCount dropped. • Duplicates removed using drop_duplicates from pandas. • URLs removed using re library. • All the tweets are in english, crosschecked using detect from the langdetect library. • Punctuations removed using the re library. • Text converted to lowercase using str.lower. • Emojis converted to their text equivalent using the demoji library. • Unicode normalization performed using the unicodedata library. • Words tokenized using split from pandas. • Stopwords removed using stopwords from the nltk library. • Lemmatization performed using the spacy library.

Table 15: Data Preparation done on Dataset A1 to obtain different datasets with variations.

3.4.2 Dataset A2 and A3

The attributes of A2 along with its datatypes are mentioned in table 14.

A2 consisted of 3782 tweets. Duplicates were removed from dataset A2.

From A2, A3 was created. Advertisements were manually removed and tweets related to 17 selected keywords were retained in A3. Only certain columns were retained. The columns kept in A3 are provided in table 16

A3 consisted of 600 tweets.

In A3, a new label was added - sentiment. Each tweet in this dataset was manually labelled positive, negative or neutral depending on the sentiment it carried.

The new label - sentiment was of datatype object in A3.

A random sample was taken from dataset A3 and that led to the creation of dataset A4.

Attribute Name	Data Type	Attribute Name	Data Type
id	int64	quoteCount	int64
text	object	viewCount	float64
retweetCount	int64	createdAt	object
replyCount	int64	lang	object
likeCount	int64	sentiment	object

Table 16: Attributes present in dataset A3

3.4.3 Dataset A4

Dataset A4 contained the same attributes as A3 (mentioned in table 16), along with the inclusion of another attribute.

An additional label - class, was added. Each tweet was classified into one of the 17 classes (obtained with the help of the keywords). Classes are the same as the keywords mentioned in table 13.

Table 17 contains the attributes present in dataset A4

Table 18 contains the preparation done on dataset A4.

5 cleaned variations were made.

Libraries such as TextBlock, Spello, Autocorrect and Spellchecker were used to correct the spellings of the tweets in the column - text, but it was observed that names of places were losing their meaning. For example vagator was being converted to aviator. So it was decided not to apply the spell correction methods.

All the other preparation procedures followed have been mentioned in table 16.

Attribute Name	Data Type	Attribute Name	Data Type
id	int64	quoteCount	int64
text	object	viewCount	float64
retweetCount	int64	createdAt	object
replyCount	int64	lang	object
likeCount	int64	sentiment	object
class	object		

Table 17: Attributes present in dataset A4

Sr. no.	Cleaned Dataset Variations of A4	Preparation done - cleaning and transformation
1	D1	• Columns - twitterUrl, retweetCount, replyCount, likeCount, quoteCount, viewCount, createdAt, lang and bookmarkCount are dropped using drop from pandas. • Retained columns - id, text, class and sentiment • Missing values checked for, using missingno from matplotlib. None found. • URLs removed using re library. • All the tweets are in english, crosschecked using detect from the langdetect library.
2	D2	• Columns - twitterUrl, retweetCount, replyCount, likeCount, quoteCount, viewCount, createdAt, lang and bookmarkCount are dropped using drop from pandas. • Retained columns - id, text, class and sentiment • Missing values checked for, using missingno from matplotlib. None found. • URLs removed using re library. • All the tweets are in english, crosschecked using detect from the langdetect library. • User handles removed. Tweets containing @(mentions) were removed from tweets using the re library.

Table 18: Data Preparation done on Dataset A4 to obtain different datasets with variations.

Sr. no.	Cleaned Dataset Variations of A4	Preparation done - cleaning and transformation
3	D3	• Columns - twitterUrl, retweetCount, replyCount, likeCount, quoteCount, viewCount, createdAt, lang and bookmarkCount are dropped using drop from pandas. • Retained columns - id, text, class and sentiment • Missing values checked for, using missingno from matplotlib. None found. • URLs removed using re library. • All the tweets are in english, crosschecked using detect from the langdetect library. • Punctuations removed using the re library.
4	D4	• Columns - twitterUrl, retweetCount, replyCount, likeCount, quoteCount, viewCount, createdAt, lang and bookmarkCount are dropped using drop from pandas. • Retained columns - id, text, class and sentiment • Missing values checked for, using missingno from matplotlib. None found. • URLs removed using re library. • All the tweets are in english, crosschecked using detect from the langdetect library. • Punctuations removed using the re library. • Text converted to lowercase using str.lower. • Words tokenized using split from pandas.

Table 18: Data Preparation done on Dataset A4 to obtain different datasets with variations.

Sr. no.	Cleaned Dataset Variations of A4	Preparation done - cleaning and transformation
5	D5	• Columns - twitterUrl, retweetCount, replyCount, likeCount, quoteCount, viewCount, createdAt, lang and bookmarkCount are dropped using drop from pandas. • Retained columns - id, text, class and sentiment • Missing values checked for, using missingno from matplotlib. None found. • URLs removed using re library. • All the tweets are in english, crosschecked using detect from the langdetect library. • Punctuations removed using the re library. • Text converted to lowercase using str.lower. • Words tokenized using split from pandas. • Stopwords removed using stopwords from the nltk library. • Lemmatization performed using the spacy library.

Table 18: Data Preparation done on Dataset A4 to obtain different datasets with variations.

3.5 Exploratory Data Analysis

3.5.1 EDA on A1 and its variations

Although all that was needed from dataset A1 was the column - text. Basic EDA was done to gather more insights into the data.

Figure 21: Word Cloud on the column - text on dataset A1

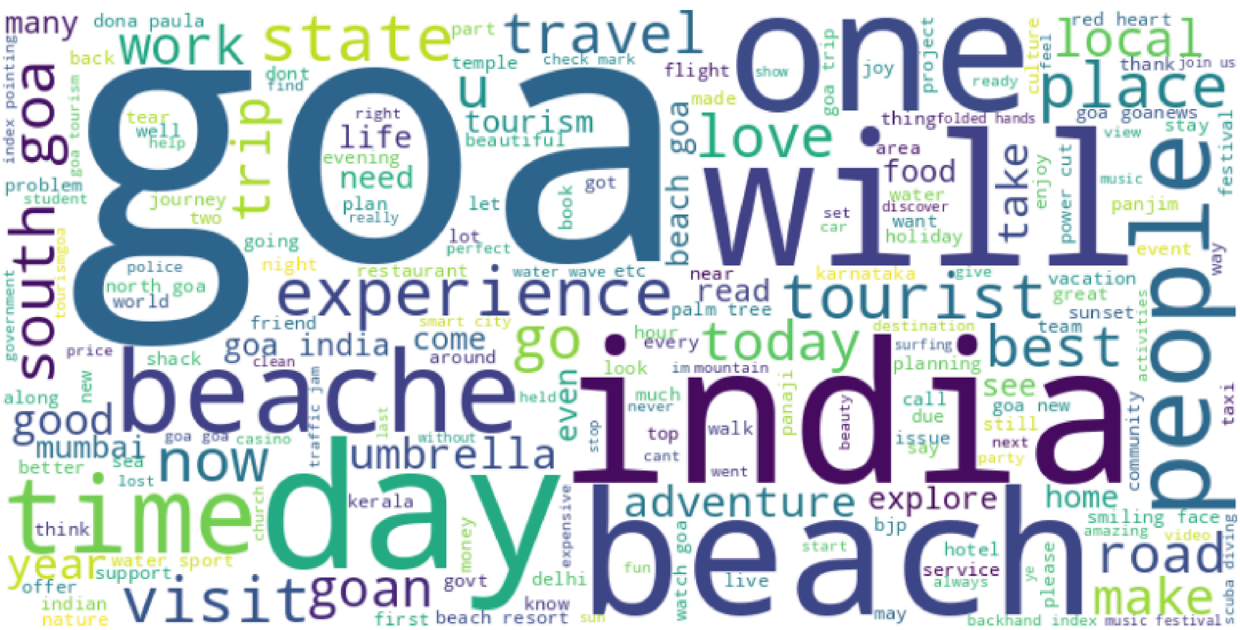


Figure 22: Word Cloud on the column - text on V3

```
Word count: 441524
Character count: 2787209
Average word length: 5.261507868201955
```

Figure 23: Basic general stats on the column - text on V3

```
Summary Statistics:
                 id  retweetCount    replyCount     likeCount    quoteCount
count  1.390300e+04  13903.000000  13903.000000  13903.000000  13903.000000
mean   1.642774e+18      3.684169      1.049845     14.778105      0.168381
std    2.149253e+17     49.409421      8.388240    181.684458      2.189005
min    4.526943e+17      0.000000      0.000000      0.000000      0.000000
25%    1.619304e+18      0.000000      0.000000      0.000000      0.000000
50%    1.740358e+18      0.000000      0.000000      1.000000      0.000000
75%    1.767229e+18      0.000000      1.000000      3.000000      0.000000
max    1.776396e+18   4728.000000    400.000000  13683.000000    118.000000

       bookmarkCount     viewCount
count   13903.000000  1.058500e+04
mean        0.676760  1.134619e+03
std        20.179598  1.682894e+04
min         0.000000  1.000000e+00
25%         0.000000  2.600000e+01
50%         0.000000  7.000000e+01
75%         0.000000  2.660000e+02
max      1962.000000  1.305446e+06
```

Figure 24: Summary Statistics on V3

```
Missing Values:
id                  0
twitterUrl          0
text                0
retweetCount        0
replyCount          0
likeCount           0
quoteCount          0
createdAt           0
lang              105
bookmarkCount       0
viewCount        3318
dtype: int64
```

Figure 25: Missing values Statistics on V3

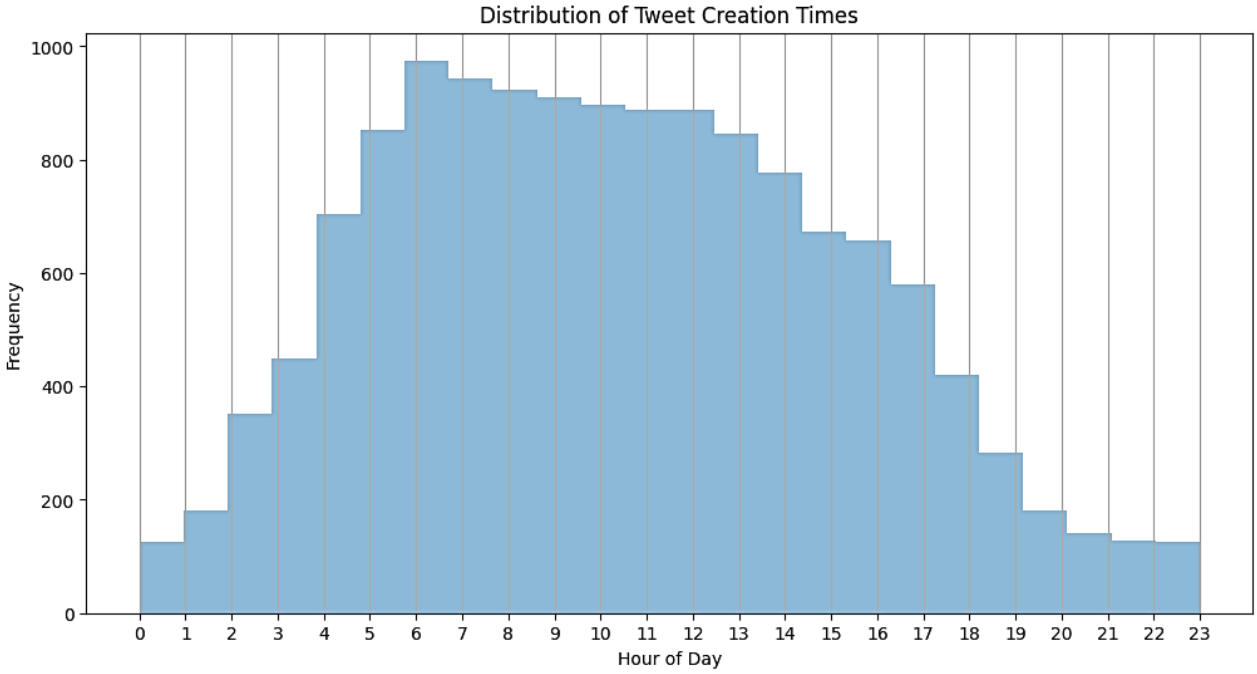


Figure 26: Distribution of tweet creation times on V3

```
Topics generated by LDA:
Topic 1:
['goa', 'beach', 'day', 'vacation', 'rainy', 'spending', 'agonda', 'leisure', 'dance', 'india']

Topic 2:
['goa', 'butterfly', '2024', 'shigmotsav', 'beach', 'festival', 'carrental', 'exploregoa', 'car', 'tourism']

Topic 3:
['goa', 'face', 'beach', 'smiling', 'joy', 'tears', 'sunset', 'dance', 'day', 'people']

Topic 4:
['flight', 'goa', 'delay', 'indigo', 'delhi', 'beach', 'mumbai', 'passengers', 'wine', 'passenger']

Topic 5:
['goa', 'minister', 'shri', 'mhadei', 'city', 'govt', 'smart', 'bjp', 'beach', 'development']

Topic 6:
['goa', 'beach', 'tourism', 'power', 'electricity', 'floor', 'local', 'primegoa', 'rolling', 'tv_channel']

Topic 7:
['beach', 'goa', 'beaches', 'water', 'travel', 'adventure', 'goas', 'india', 'best', 'explore']

Topic 8:
['goa', 'paula', 'dona', 'beach', 'traffic', 'jetty', 'jam', 'accident', 'news', 'music']

Topic 9:
['goa', 'tourism', 'tour', 'india', 'live', 'hub', 'friends', 'cab', 'family', 'travel']

Topic 10:
['goa', 'beach', 'life', 'river', 'mandovi', 'boredom', 'fitness', 'india', 'utorda', 'murders']
```

Figure 27: LDA on the column - text on V3

```
Top words for each topic:
Topic 1: beach, resort, goa, candolim, anjuna, majorda, sunset, south, utorda, colva
Topic 2: goa, india, beaches, trip, day, tourist, like, just, planning, visit
Topic 3: tour, tourism, travel, cabbazar, trending, cab, photo, cabs, hub, coastal
Topic 4: face, smiling, tears, joy, eyes, hearteyes, crying, loudly, grinning, heart
Topic 5: water, sports, adventure, activities, scuba, diving, parasailing, ride, snorkeling, watersports
Topic 6: tourism, goas, development, modikiguarantee_goa, cultural, journey, heritage, sustainable, pm, future
Topic 7: paula, dona, jetty, minister, rohankhaunte, donapaula, miramar, dec, public, open
Topic 8: vacation, spending, agonda, leisure, beach, holiday, travel, hotel, hospitality, summer
Topic 9: music, dance, festivals, festival, experience, imvoyager, food, goa, cultural, workshop
Topic 10: city, panjim, news, smart, panaji, traffic, parking, power, goanews, goa
```

Figure 28: NMF on the column - text on V3

3.5.2 EDA on A4

Figure 29: Checking for missing data with A4

	id	retweetCount	replyCount	likeCount	quoteCount
count	3.000000e+02	300.000000	300.000000	300.000000	300.000000
mean	1.636143e+18	5.333333	0.863333	22.983333	0.166667
std	2.509781e+17	43.308363	3.984668	237.456512	1.007774
min	5.160000e+17	0.000000	0.000000	0.000000	0.000000
25%	1.660000e+18	0.000000	0.000000	0.000000	0.000000
50%	1.740000e+18	0.000000	0.000000	0.000000	0.000000
75%	1.760000e+18	0.000000	1.000000	3.000000	0.000000
max	1.780000e+18	635.000000	57.000000	4046.000000	12.000000

	viewCount	bookmarkCount
count	245.000000	300.000000
mean	1262.326531	1.016667
std	10893.128914	16.399580
min	6.000000	0.000000
25%	40.000000	0.000000
50%	105.000000	0.000000
75%	408.000000	0.000000
max	168465.000000	284.000000

Figure 30: Summary Statistics of A4

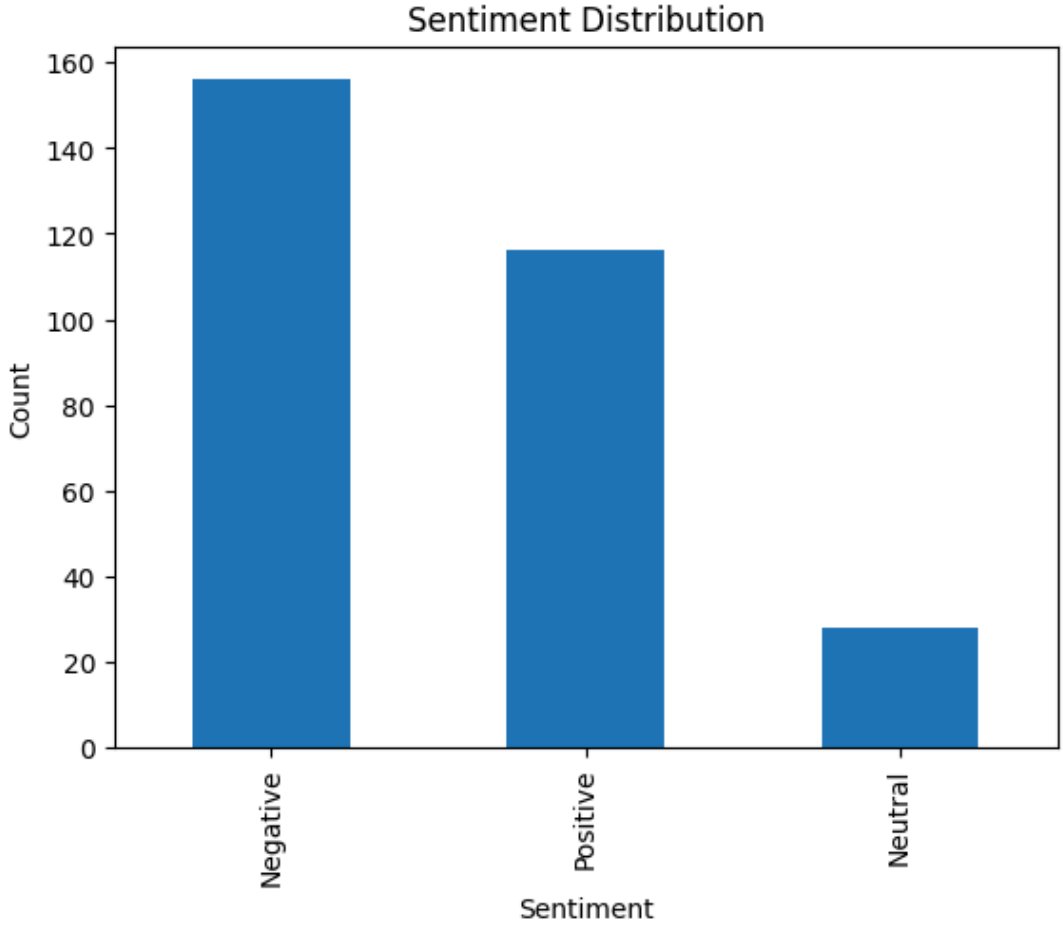


Figure 31: Sentiment Distribution of A4

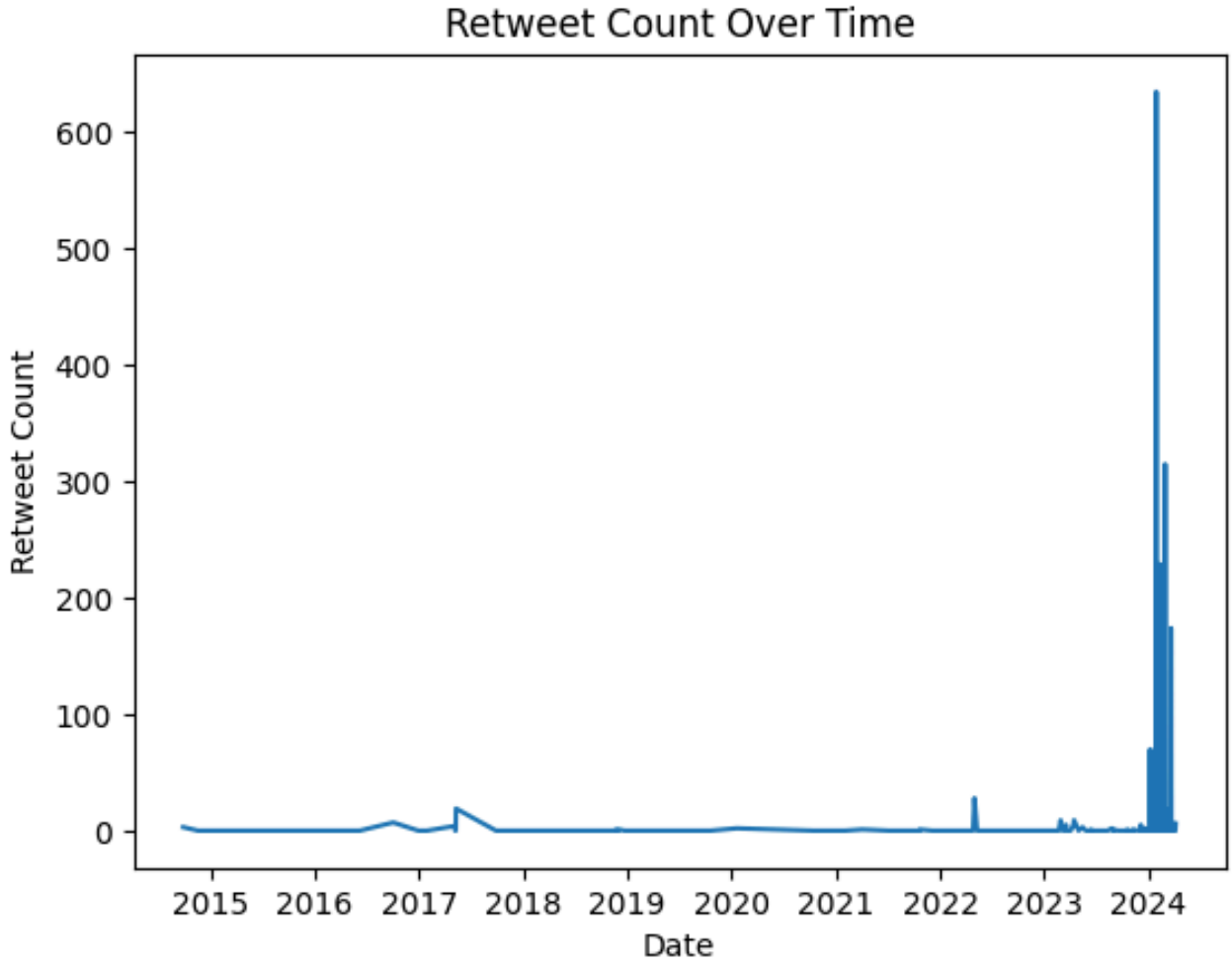


Figure 32: Retweet Count Over Time of A4

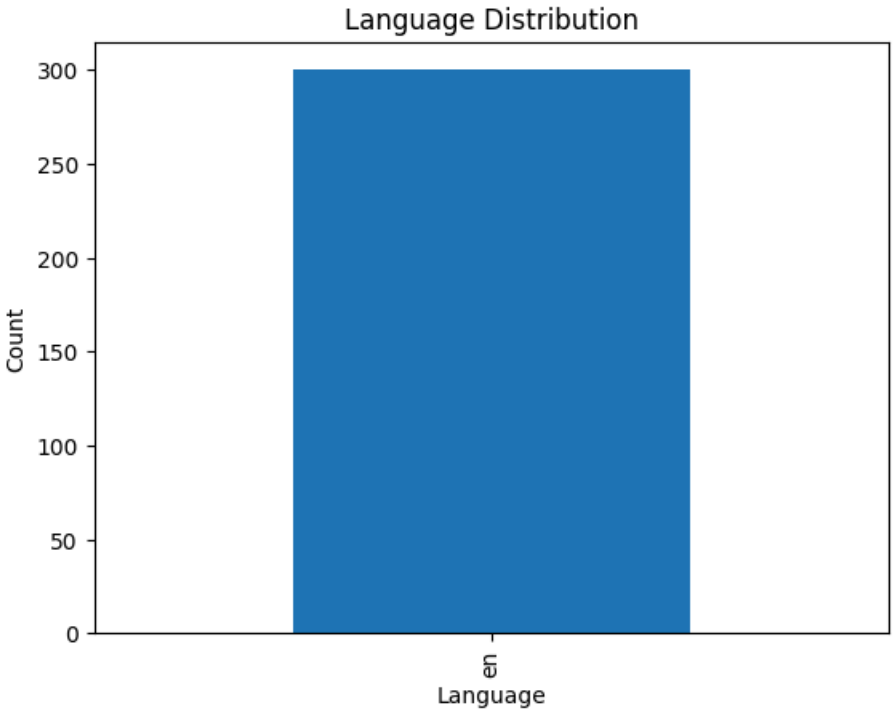


Figure 33: Language Distribution of A4

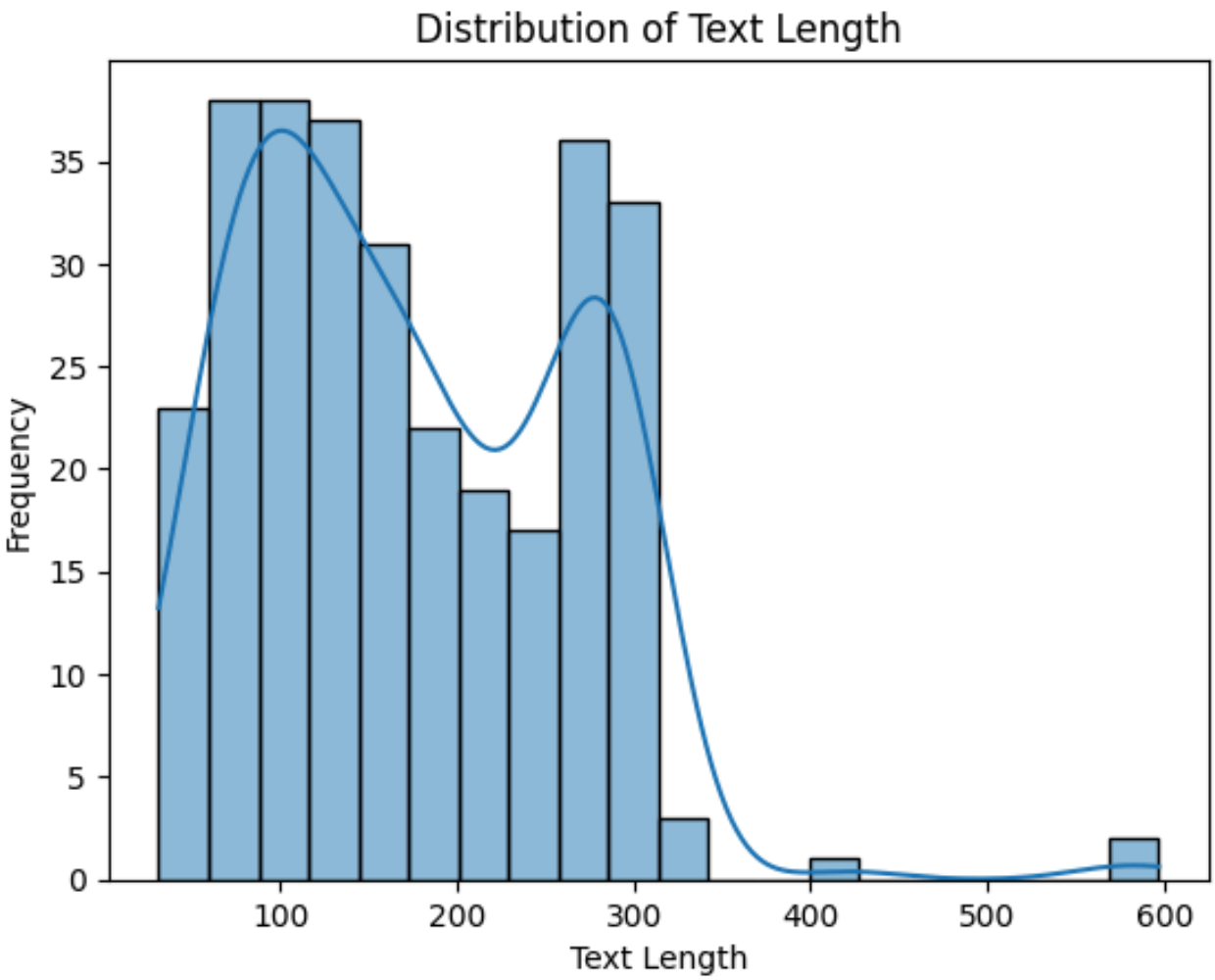


Figure 34: Distribution of Text LengthA4

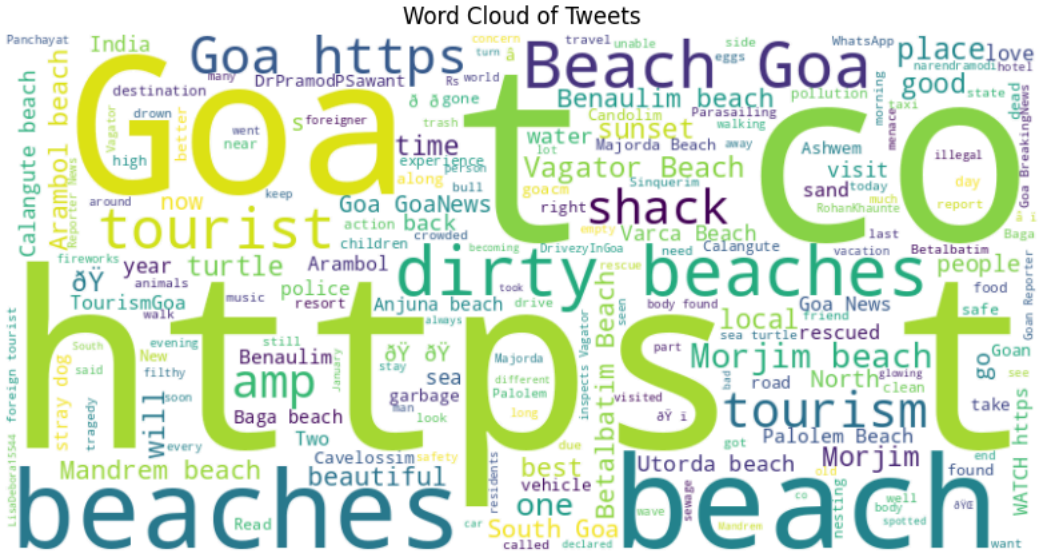


Figure 35: Word Cloud of column - text of A4

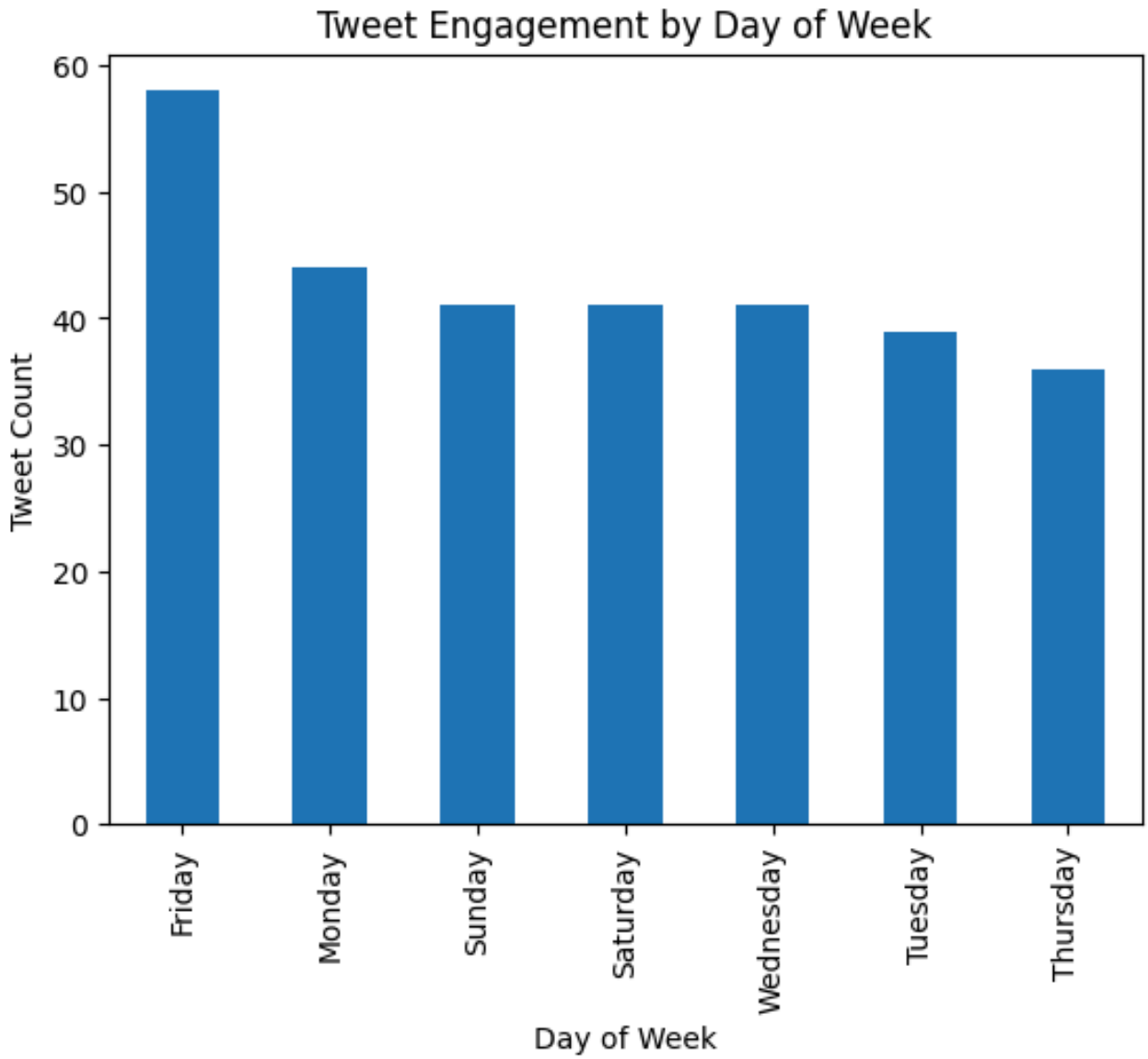


Figure 36: Tweet Engagement by Day of Week of A4

	text	engagement
0	#Arambol or #Harmal Beach is full of Russians....	4750
1	Recently held a meeting with local shack owner...	737
2	23. Betalbatim Beach, Goa https://t.co/EP4WxSpbgs	613
3	Record-breaking news from Vagator Beach, Goa! ...	370
4	Major milestone in Goa's sea turtle conservati...	350
5	Benaulim MLA Capt Venzy Viegas expressed conce...	231
6	I was in Goa for five days, & was appalled...	217
7	Tourist drives his vehicles on protected turtl...	124
8	Sunset at Betalbatim Beach. #Goa https://t.co/...	120
9	#NewsfromNITI: Goa to announce a WhatsApp no. ...	101

Figure 37: Popular Tweet Analysis of A4 (top 10 most engaged tweets)

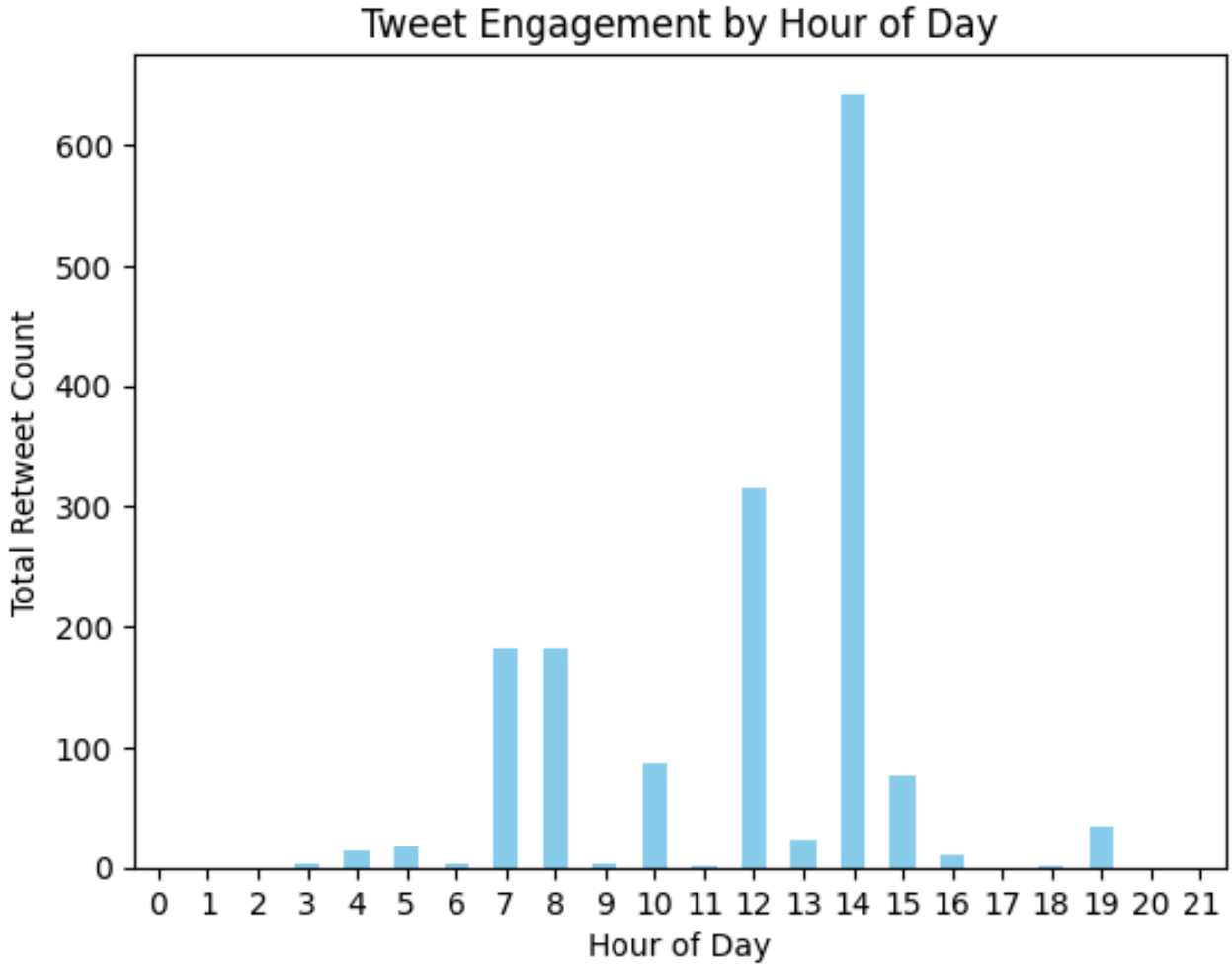


Figure 38: Tweet Engagement by Hour of Day of A4

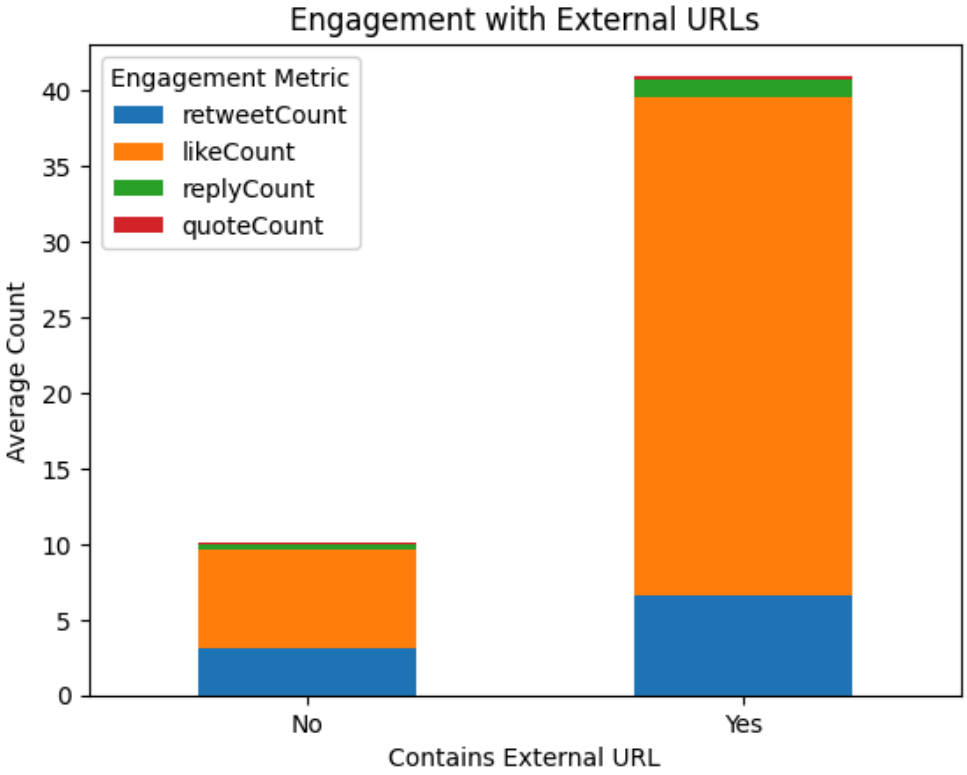


Figure 39: Engagement with External URLs

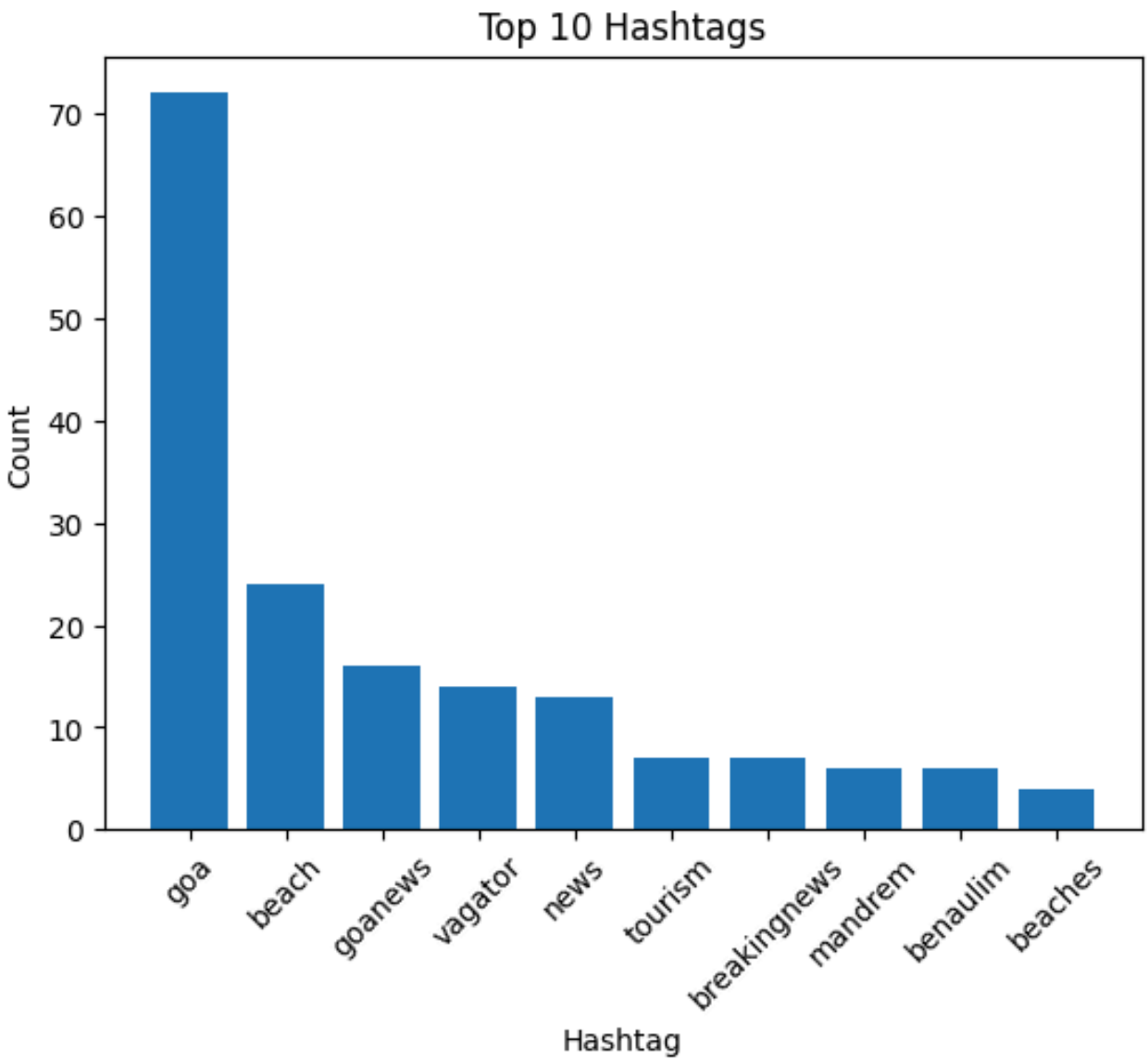


Figure 40: Top 10 Hashtags of A4

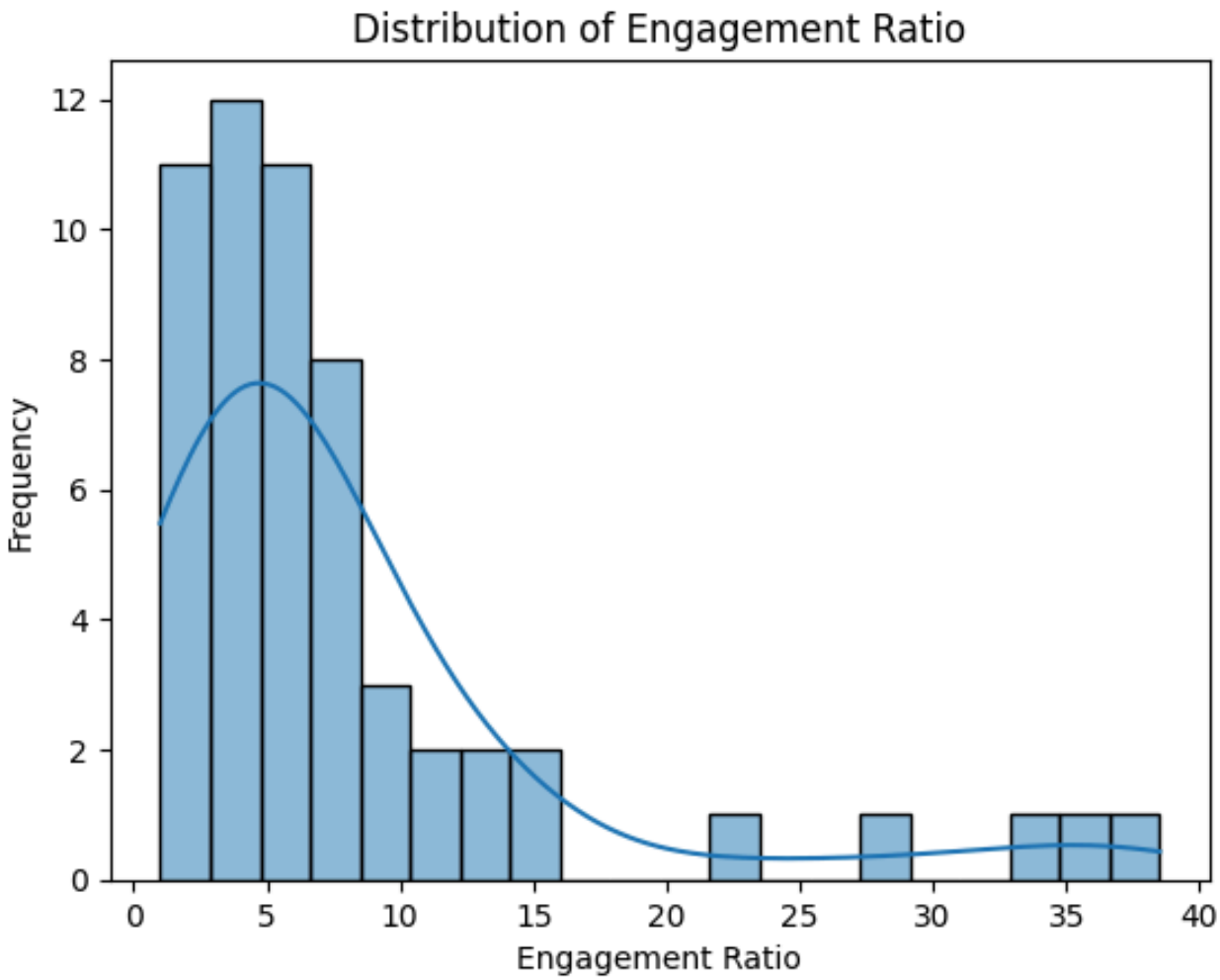


Figure 41: Distribution of Engagement Ratio of A4

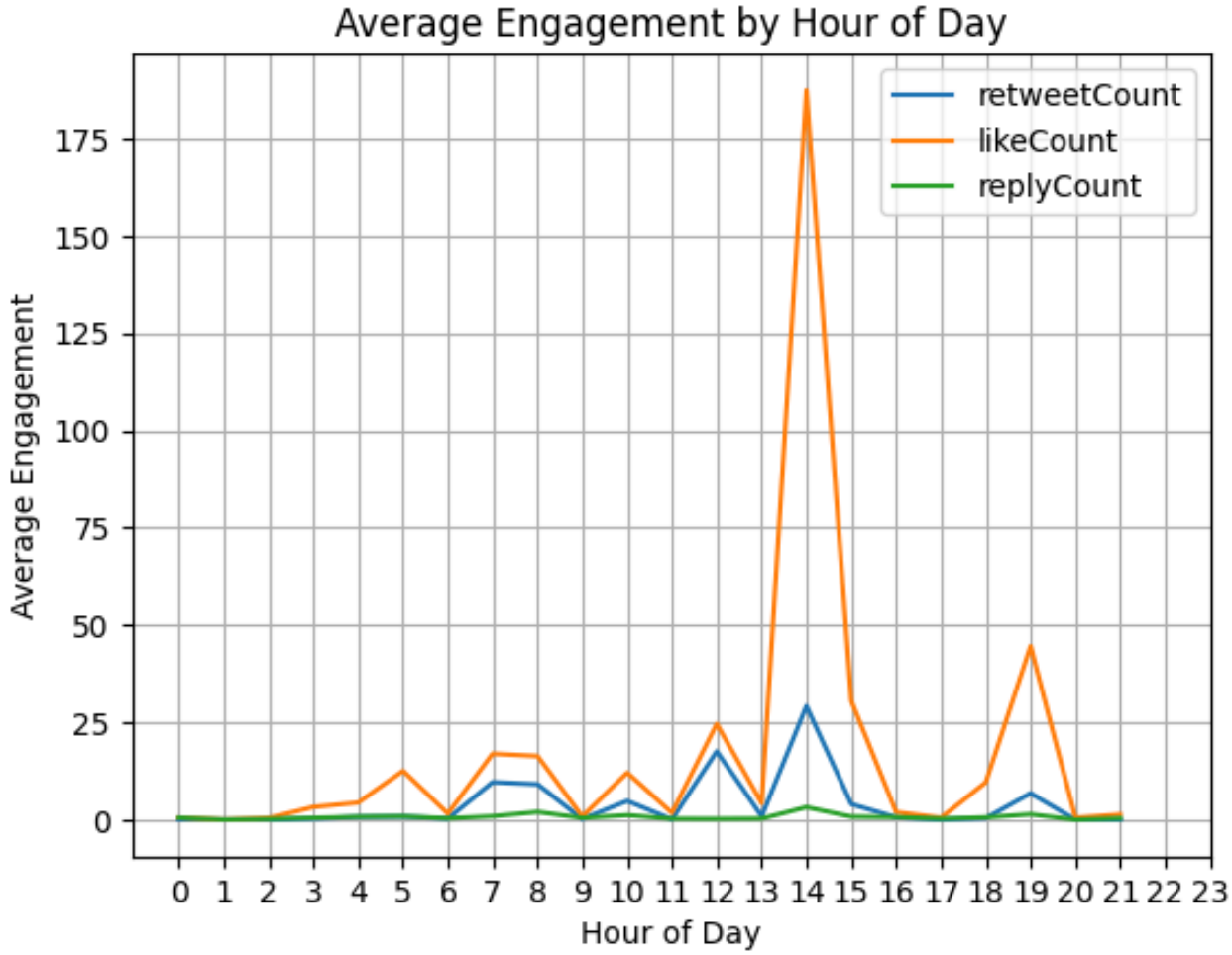


Figure 42: Average Engagement By Hour of Day of A4

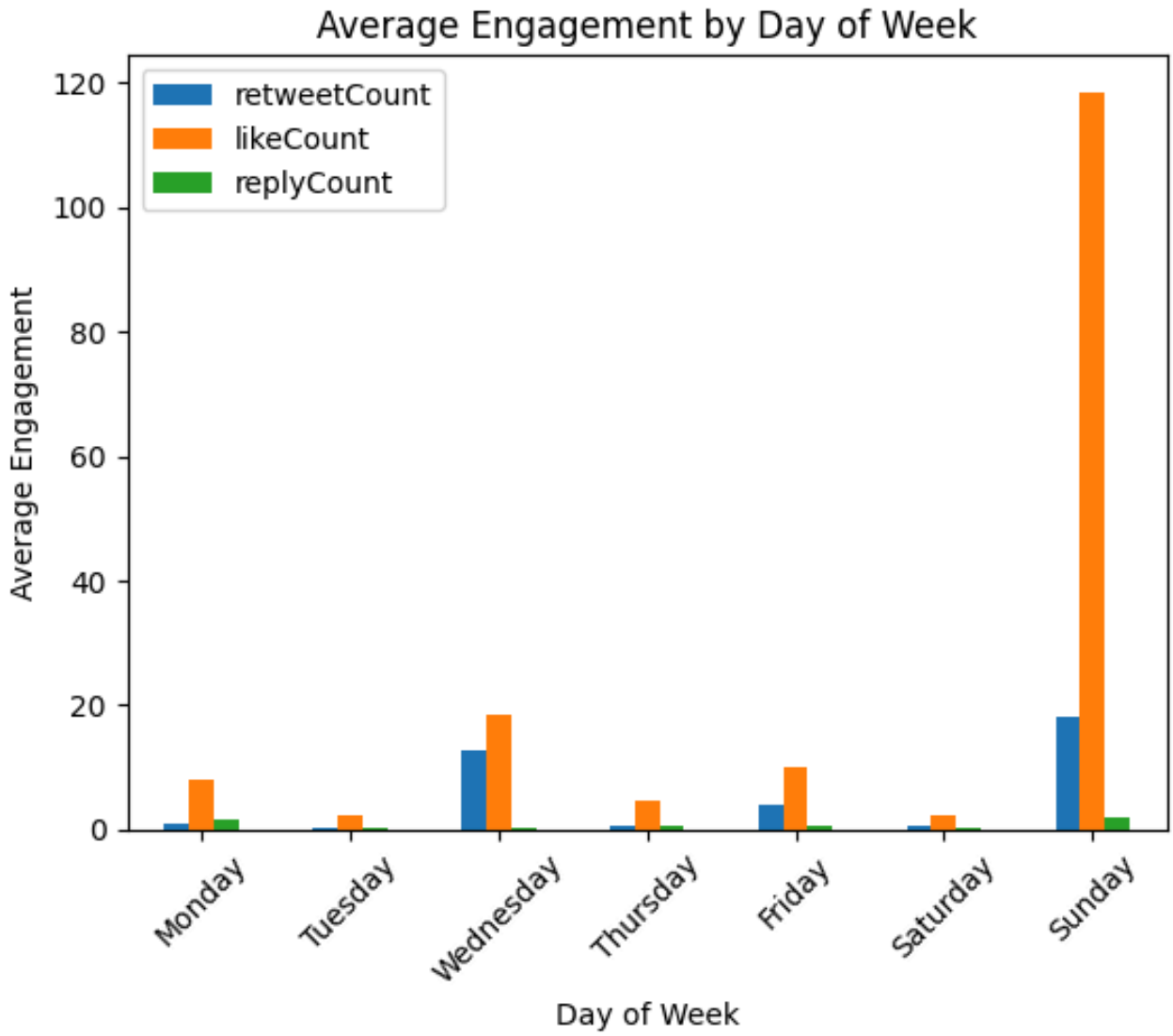


Figure 43: Average Engagement By Day of Week of A4

```
Topics generated by LDA:
Topic 1:
['beach', 'dogs', 'goa', 'gone', 'expensive', 'environment', 'india', 'favourite', 'https', 'natural']

Topic 2:
['https', 'read', 'beach', 'goa', 'news', 'varca', 'betalbatim', 'goanews', 'turtle', 'headlines']

Topic 3:
['arambol', 'drown', 'beach', 'https', 'thebarmyarmy', 'weather', 'body', 'brothers', 'goa', 'mobile']

Topic 4:
['ðÿ', 'beach', 'sunset', 'https', 'goa', 'mandrem', 'cavelossim', 'betalbatim', 'water', 'india']

Topic 5:
['sea', 'beaches', 'goa', 'vagator', 'https', 'dirty', 'turtle', 'beach', 'department', 'health']

Topic 6:
['goa', 'beaches', 'https', 'beach', 'dirty', 'majorda', 'palolem', 'tourism', 'sunset', 'beautiful']

Topic 7:
['beach', 'goa', 'shacks', 'baga', 'candolim', 'https', 'drivezyingoa', 'rescued', 'palolem', 'lisadebora15544']

Topic 8:
['goa', 'ðÿ', 'https', 'beach', 'morjim', 'beaches', 'tourists', 'goan', 'stray', 'coastal']

Topic 9:
['beach', 'sinquerim', 'goa', 'https', 'betalbatim', 'tourist', 'best', 'beaches', 'ðÿ', 'goan']

Topic 10:
['beach', 'known', 'arambol', 'goa', 'dirty', 'beaches', 'https', 'help', 'amp', 'place']
```

Figure 44: LDA on the column - text of A4

```
Top words for each topic:
Topic 1: turtle, benaulim, https, morjim, news, beach, varca, sea, fishermen, read
Topic 2: beaches, dirty, goa, tourism, whatsapp, https, place, tourist, tourists, garbage
Topic 3: ðÿ, o8pl94pv4r, anjuna, sinquerim, beach, goa, œðÿ, ðÿœ, crazy, evening
Topic 4: sunset, palolem, goa, https, beach, anjuna, beautiful, b67qmub8f7, south, cavelossim
Topic 5: shacks, baga, drivezyingoa, beach, drivezyin, beautiful, crowded, music, time, weather
Topic 6: majorda, beach, betalbatim, https, goa, cricket, places, erosion, candolim, needed
Topic 7: mandrem, rescued, children, lifesavers, https, france, minor, beach, goan, origin
Topic 8: dogs, stray, morjim, tourists, policy, feces, overrun, pets, biting, dropping
Topic 9: arambol, drown, beach, best, utorda, https, amp, north, body, goa
Topic 10: vagator, inspects, health, sewage, https, breakingnews, goa, beach, inspection, amid
```

Figure 45: NMF on the column - text of A4

3.5.3 EDA on D5

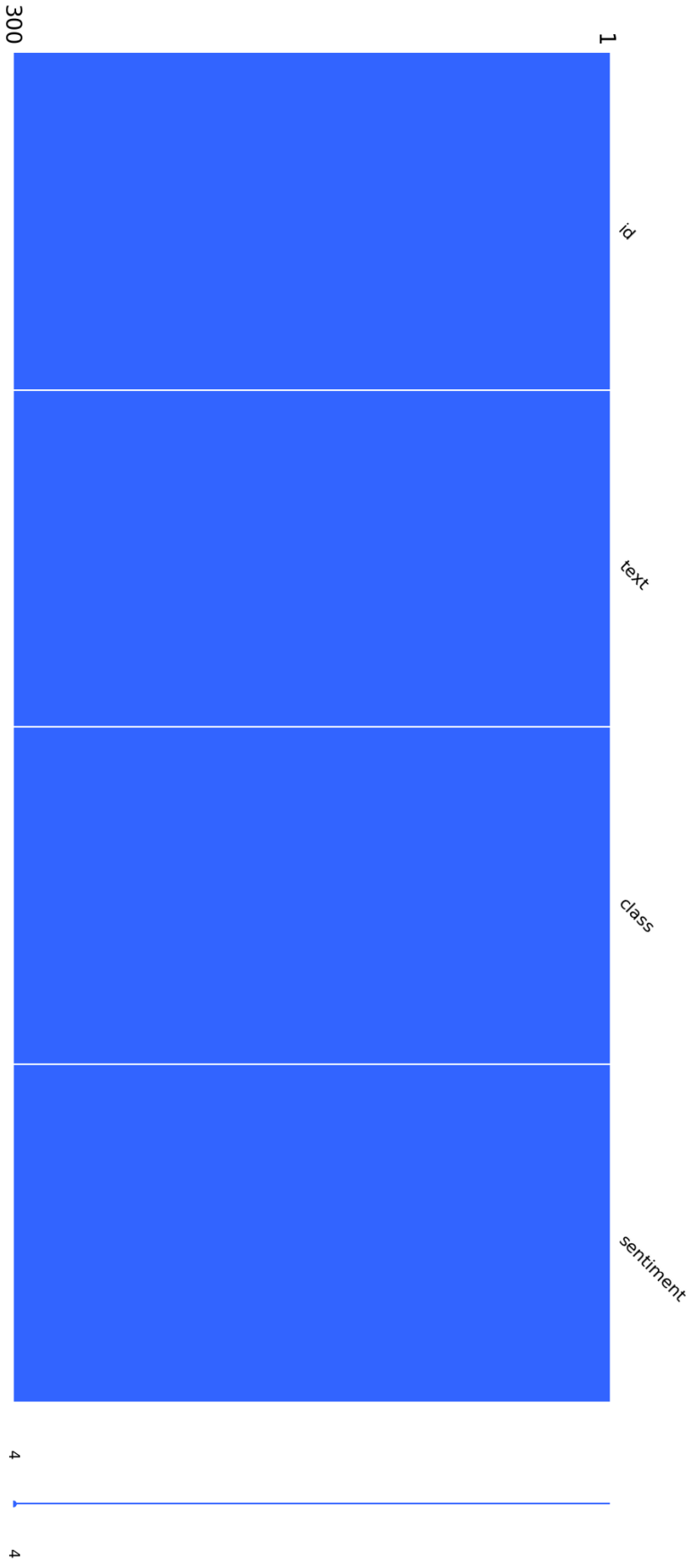


Figure 46: Checking for missing data with D5

	id
count	3.000000e+02
mean	1.636143e+18
std	2.509781e+17
min	5.160000e+17
25%	1.660000e+18
50%	1.740000e+18
75%	1.760000e+18
max	1.780000e+18

Figure 47: Summary Statistics of D5

```
Unique classes: ['Dirty beach' 'Vagator' 'Palolem' 'Cavelossim' 'Baga' 'Betalbatim'
 'Calangute' 'Morjim' 'Mandrem' 'Utorda' 'Benaulim' 'Sinquerim' 'Anjuna'
 'Arambol' 'Candolim' 'Varca' 'Majorda']

Unique sentiments: ['Negative' 'Positive' 'Neutral']
```

Figure 48: Unique values of D5

```
Class Counts:
class
Dirty beach    41
Morjim         37
Benaulim       24
Vagator        22
Arambol        21
Betalbatim     20
Mandrem        18
Calangute      17
Palolem        16
Utorda         13
Baga           12
Anjuna         12
Majorda        12
Varca          11
Cavelossim      9
Candolim        9
Sinquerim       6
Name: count, dtype: int64
```

Figure 49: Count of values in column - Class of D5

```
Sentiment Counts:
sentiment
Negative    156
Positive    116
Neutral      28
Name: count, dtype: int64
```

Figure 50: Count of Each Sentiment of D5

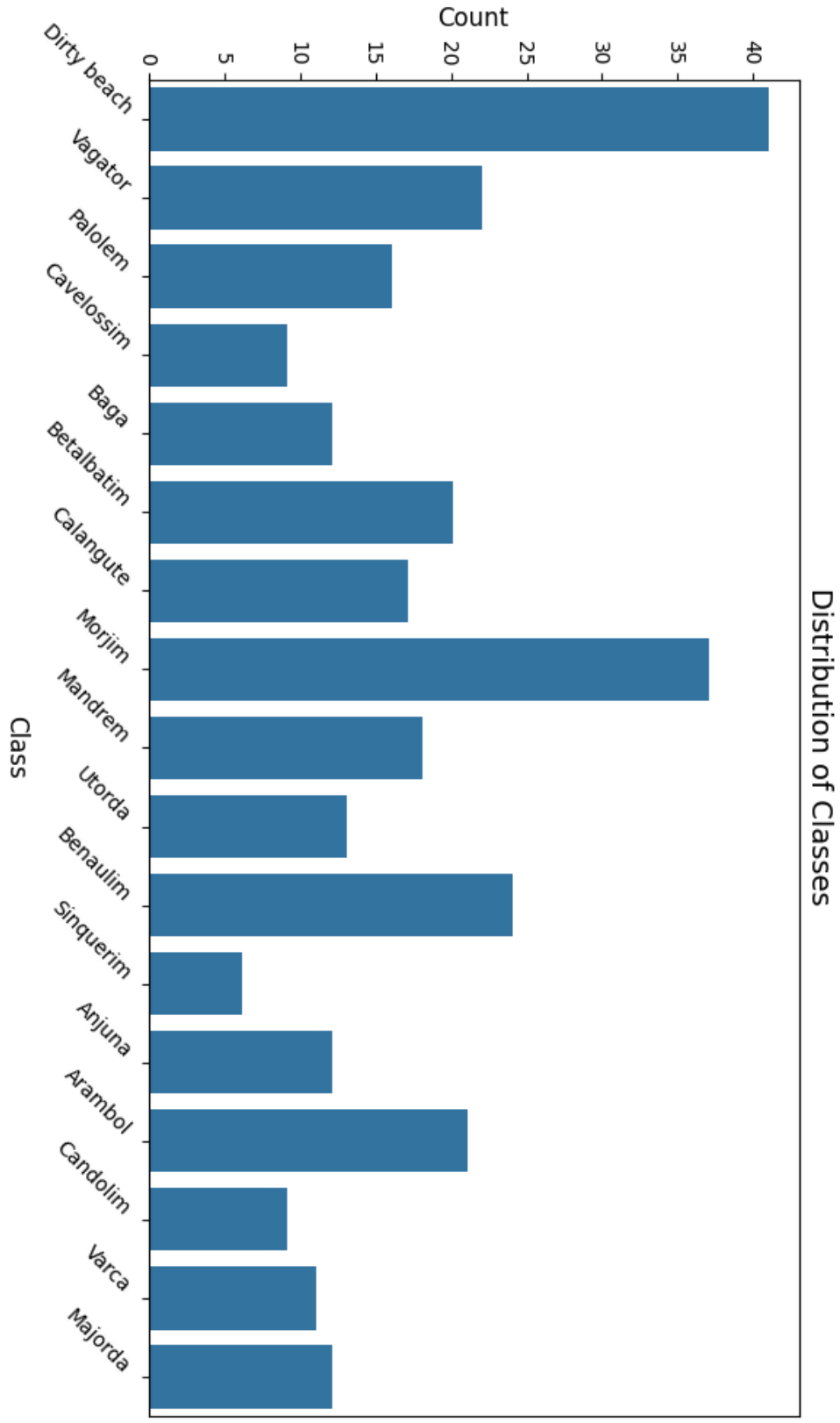


Figure 51: Distribution of Classes of D5

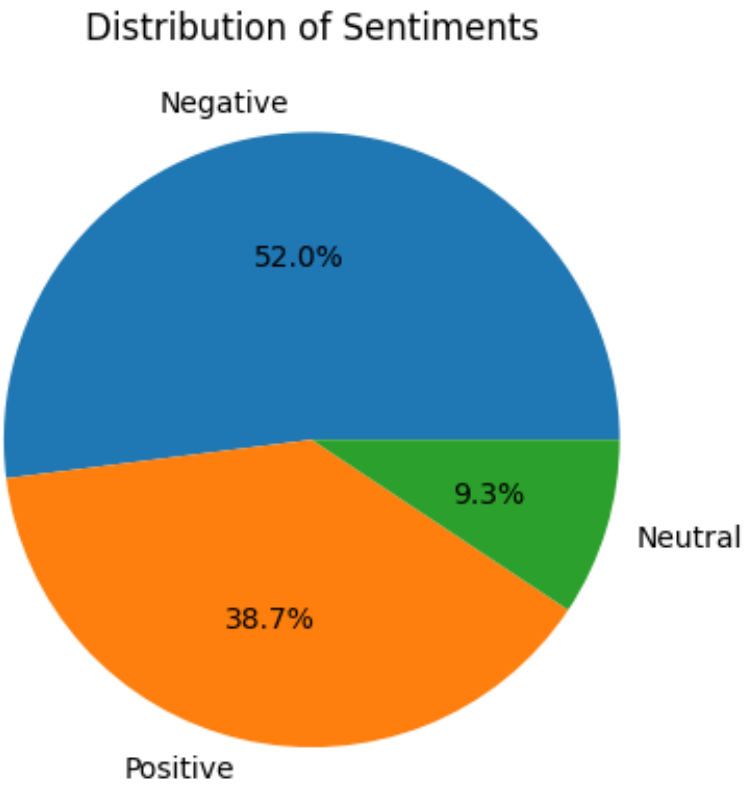


Figure 52: Distribution of Sentiments of D5

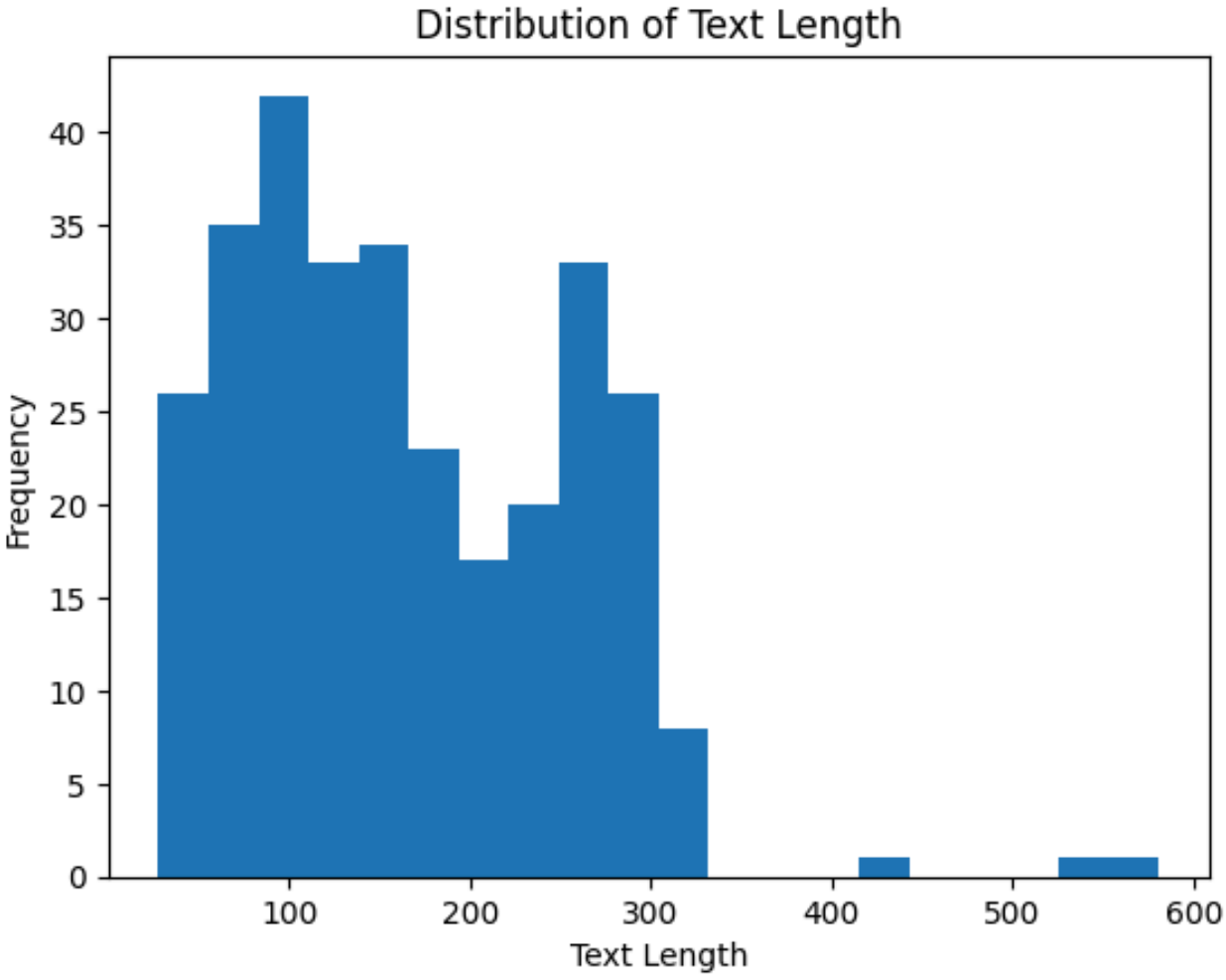


Figure 53: Distribution of Text Length of D5

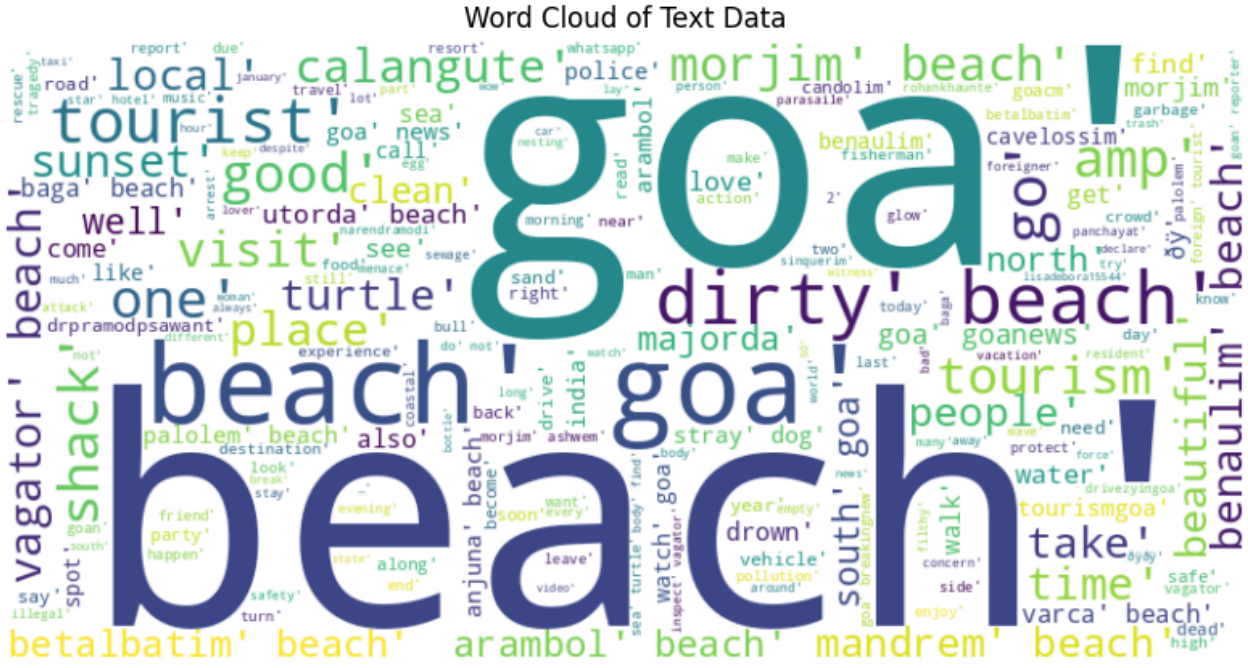


Figure 54: Word Cloud of column - Text of D5

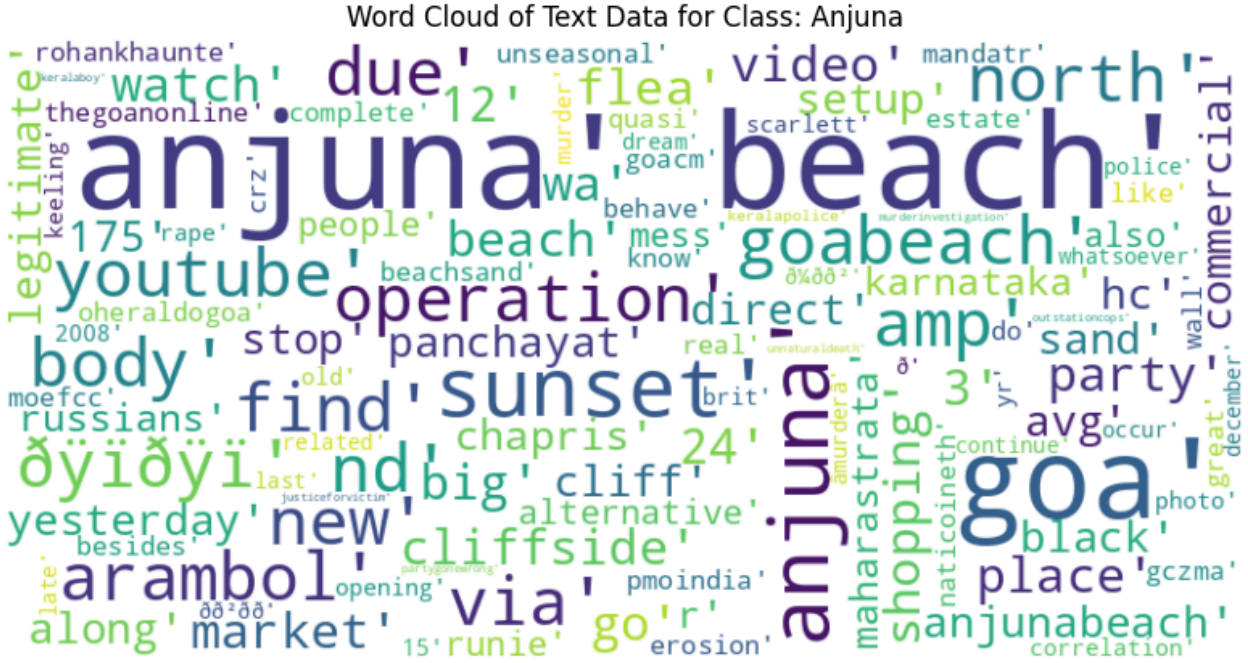


Figure 55: Word Cloud of class - Anjuna of D5

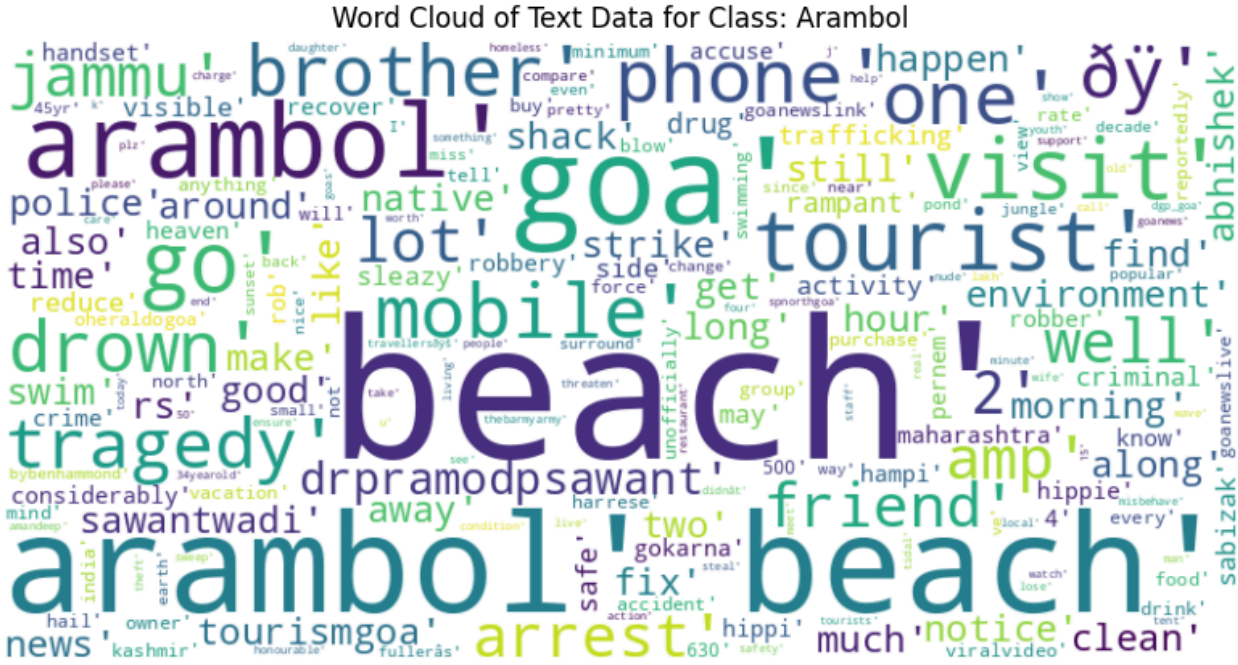


Figure 56: Word Cloud of class - Arambol of D5

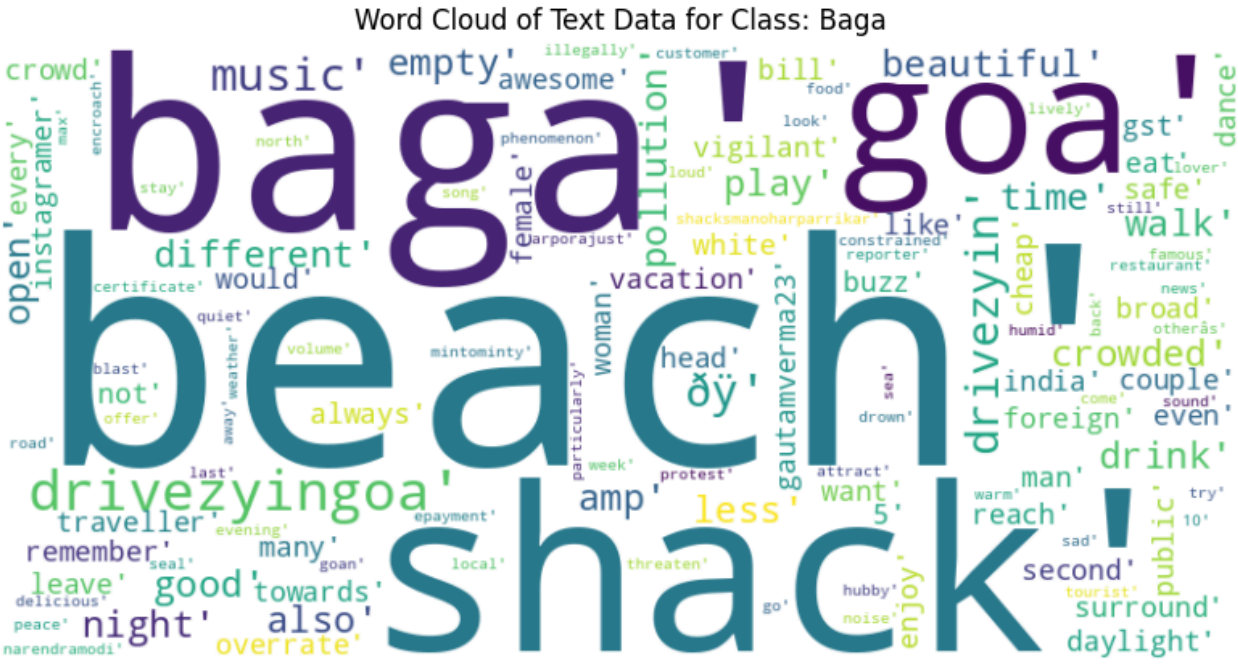


Figure 57: Word Cloud of class - Baga of D5

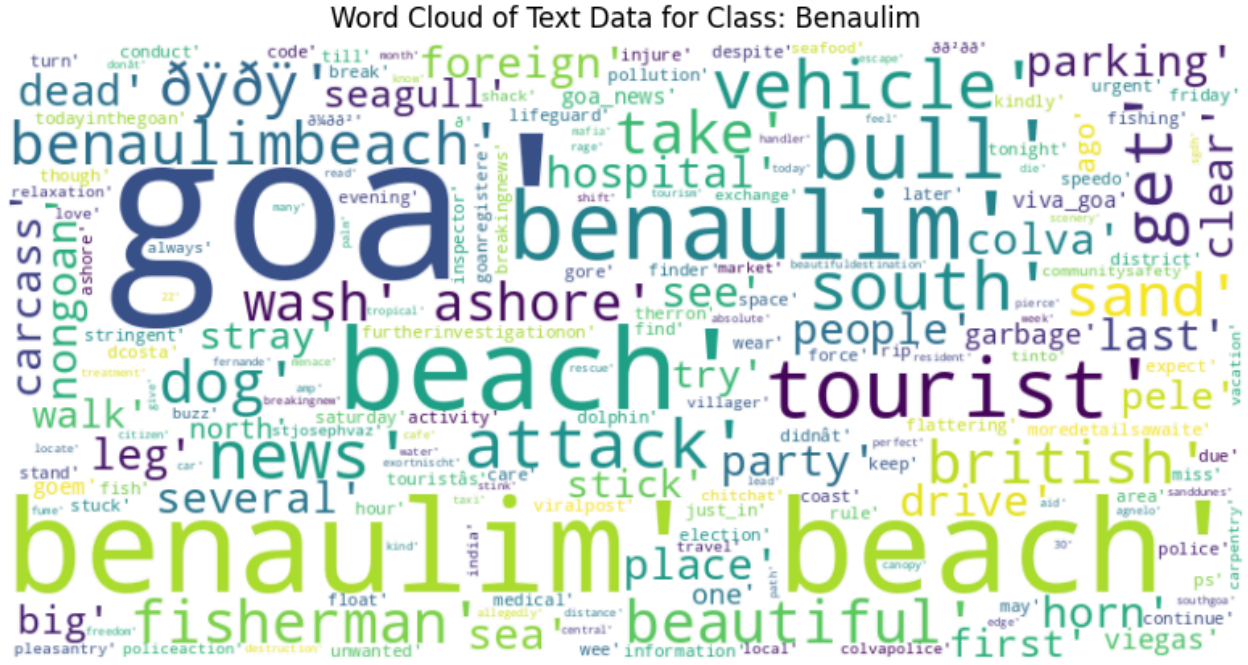


Figure 58: Word Cloud of class - Benaulim of D5

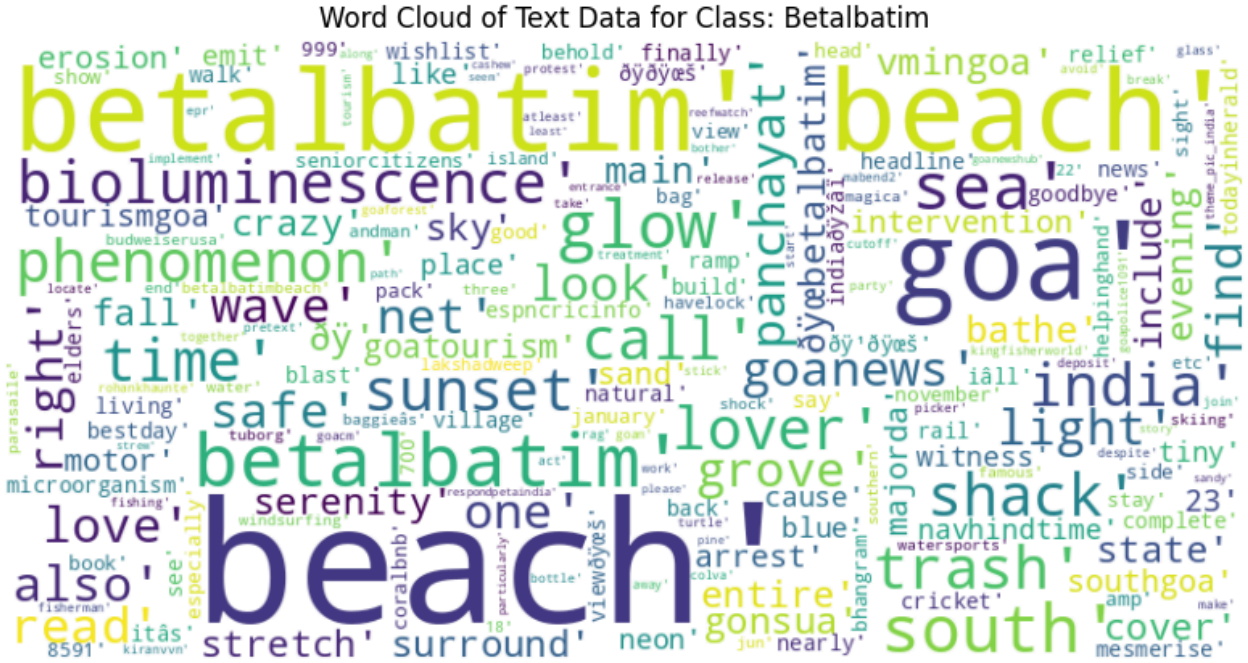


Figure 59: Word Cloud of class - Betalbatim of D5

Figure 60: Word Cloud of class - Calangute of D5

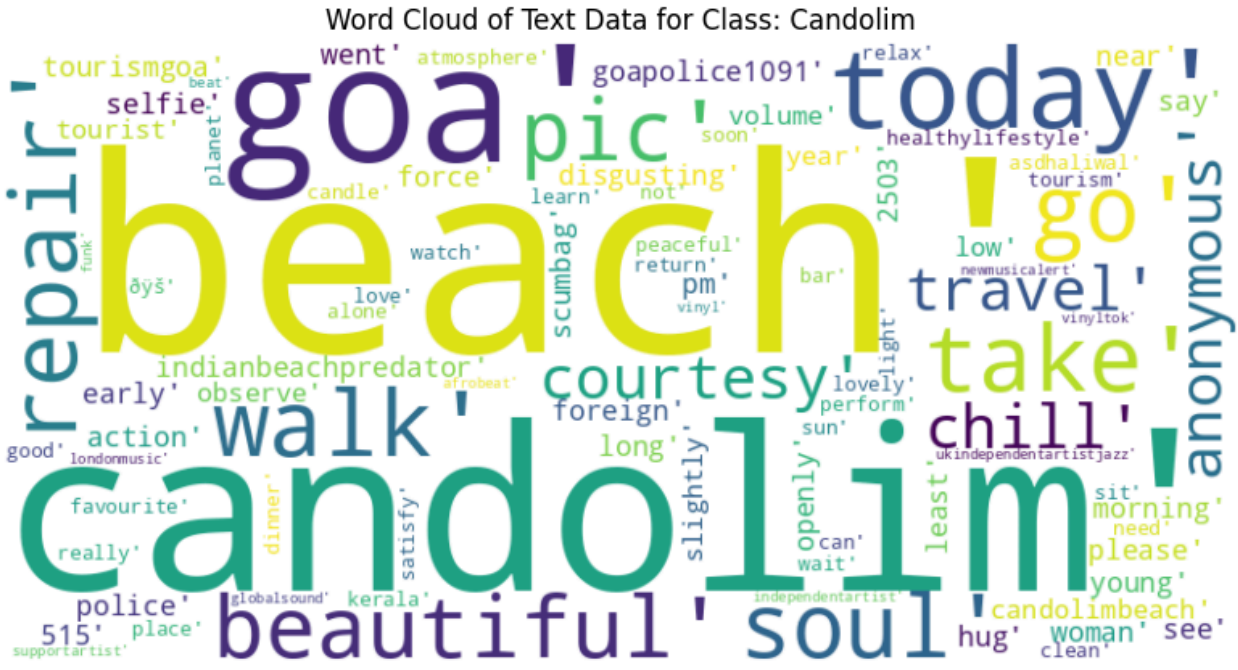


Figure 61: Word Cloud of class - Candolim of D5

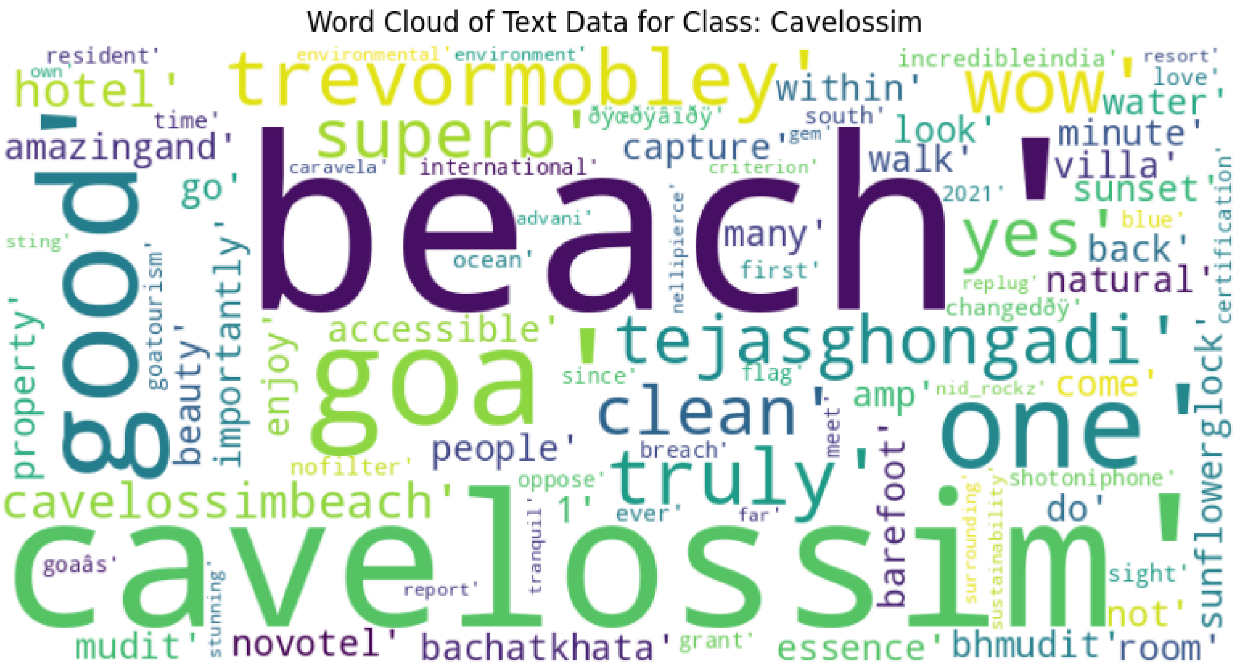


Figure 62: Word Cloud of class - Cavelossim of D5

Figure 63: Word Cloud of class - Dirty beach of D5

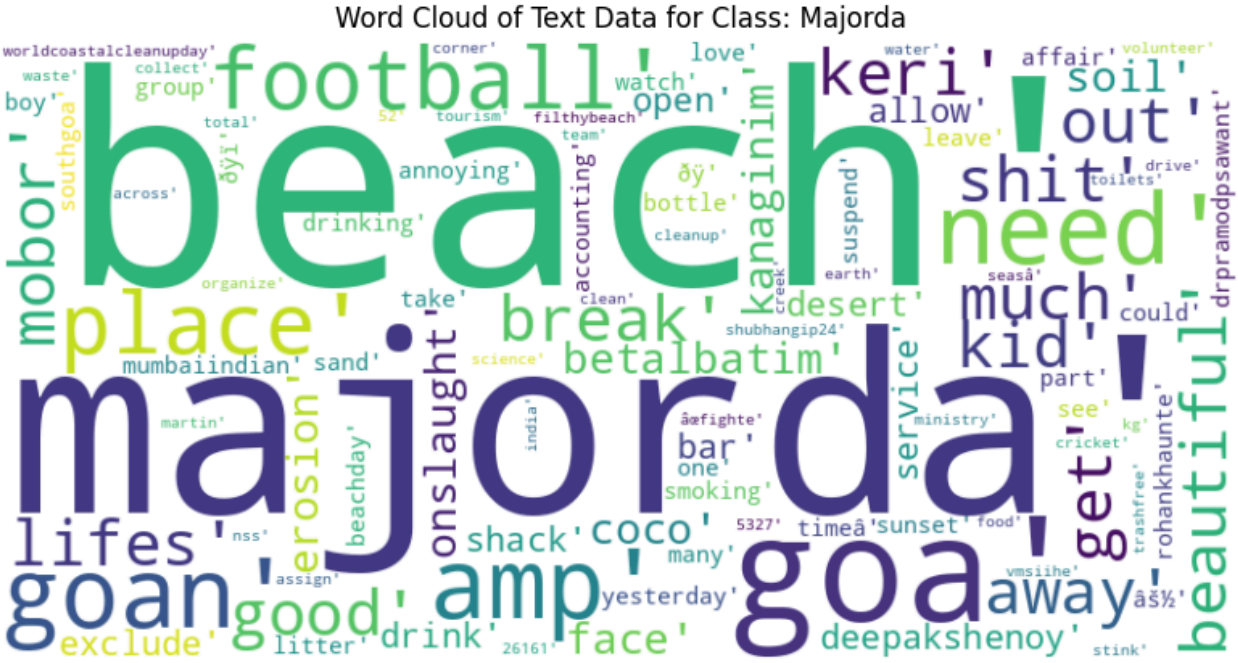


Figure 64: Word Cloud of class - Majorda of D5

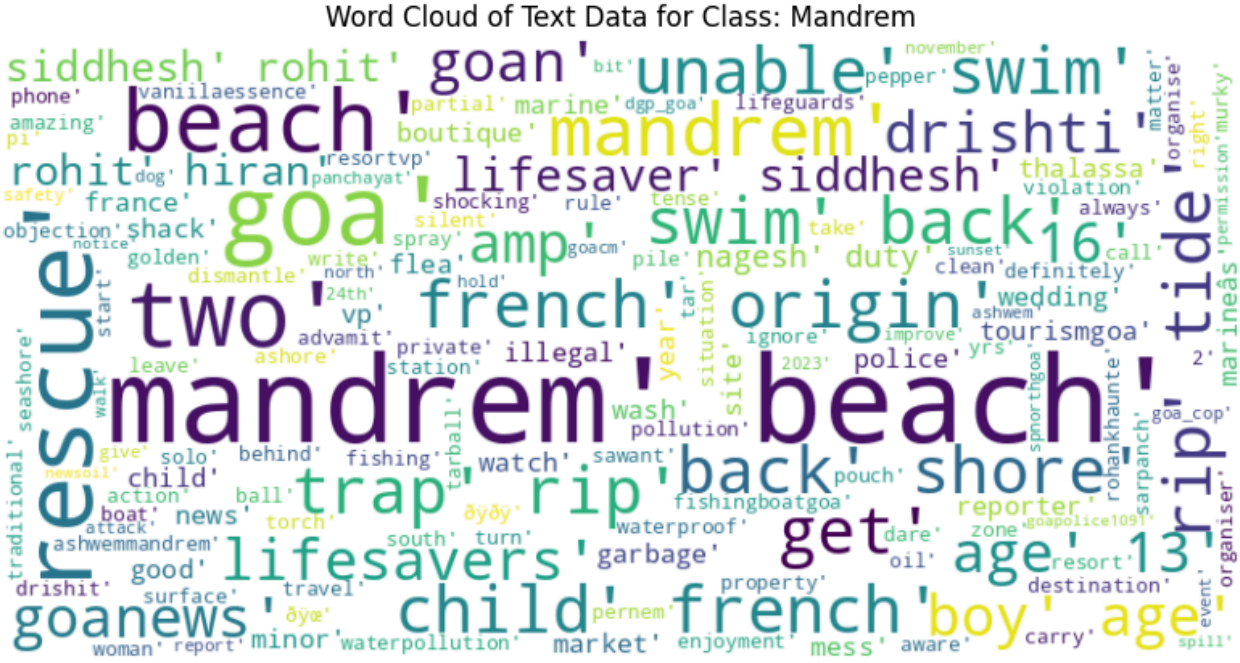


Figure 65: Word Cloud of class - Mandrem of D5

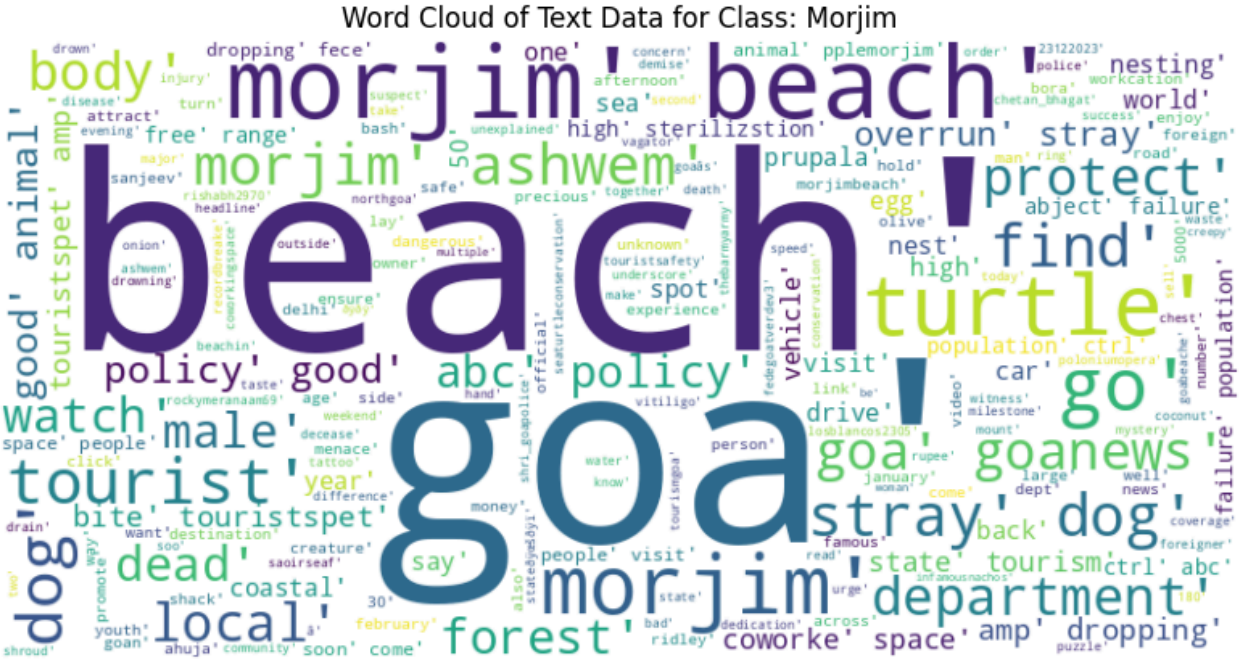


Figure 66: Word Cloud of class - Morjim of D5

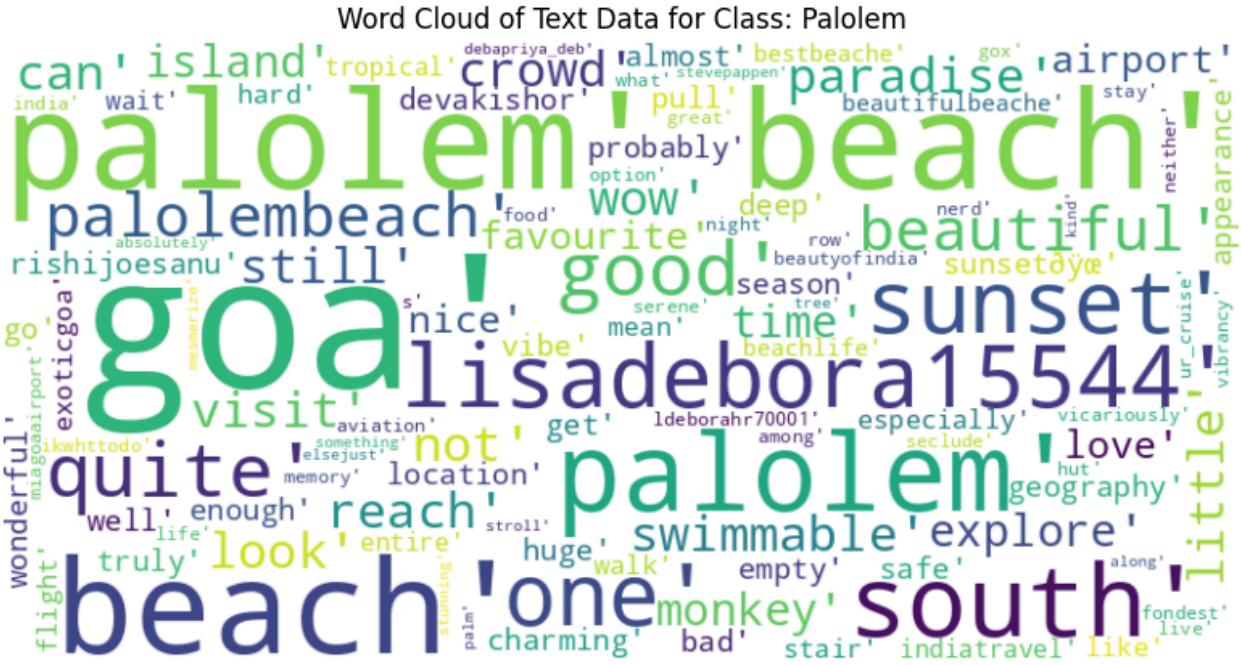


Figure 67: Word Cloud of class - Palolem of D5

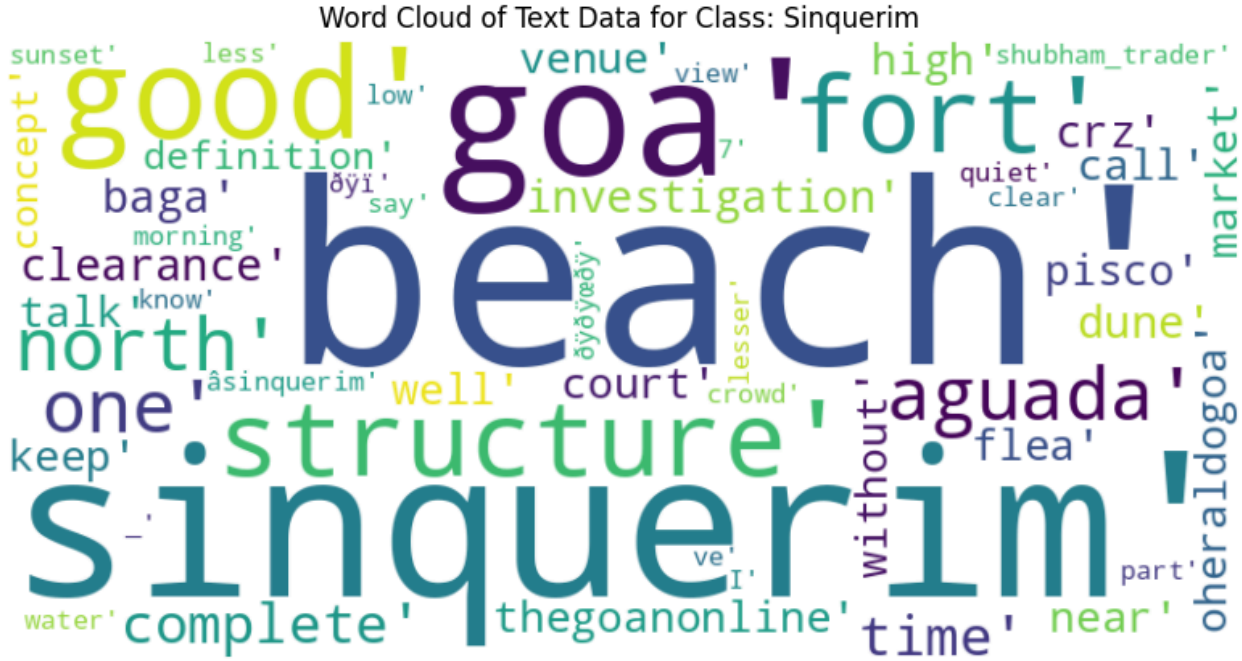


Figure 68: Word Cloud of class - Sinquerim of D5

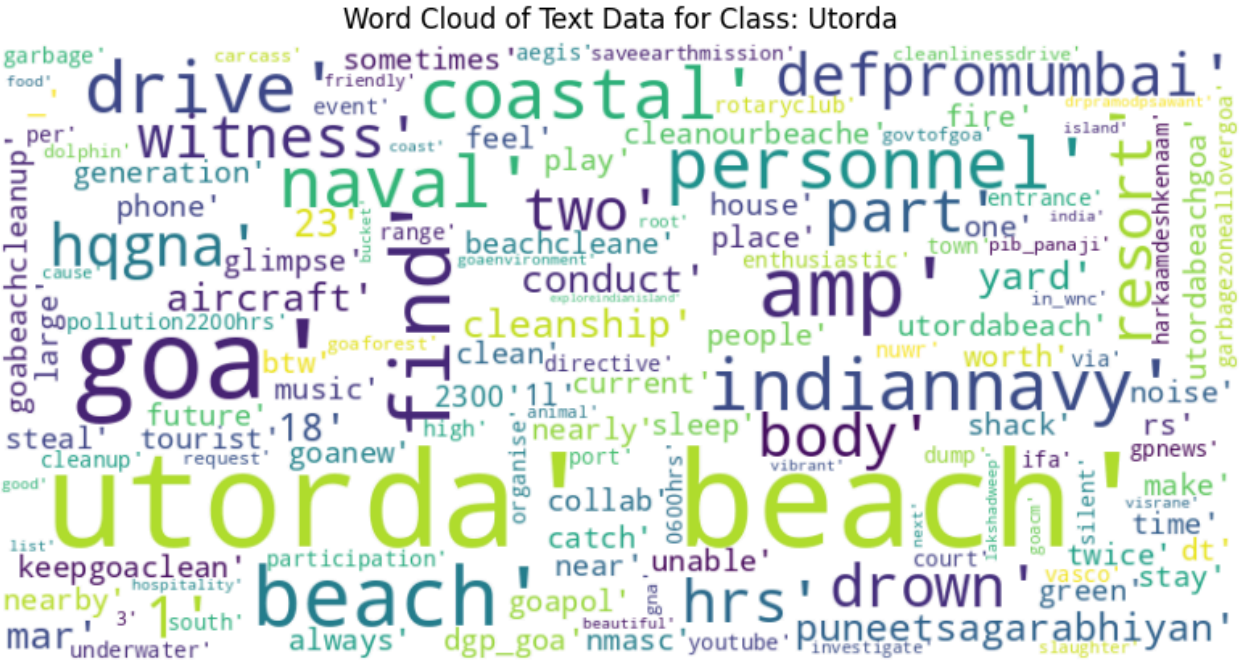


Figure 69: Word Cloud of class - Utorda of D5

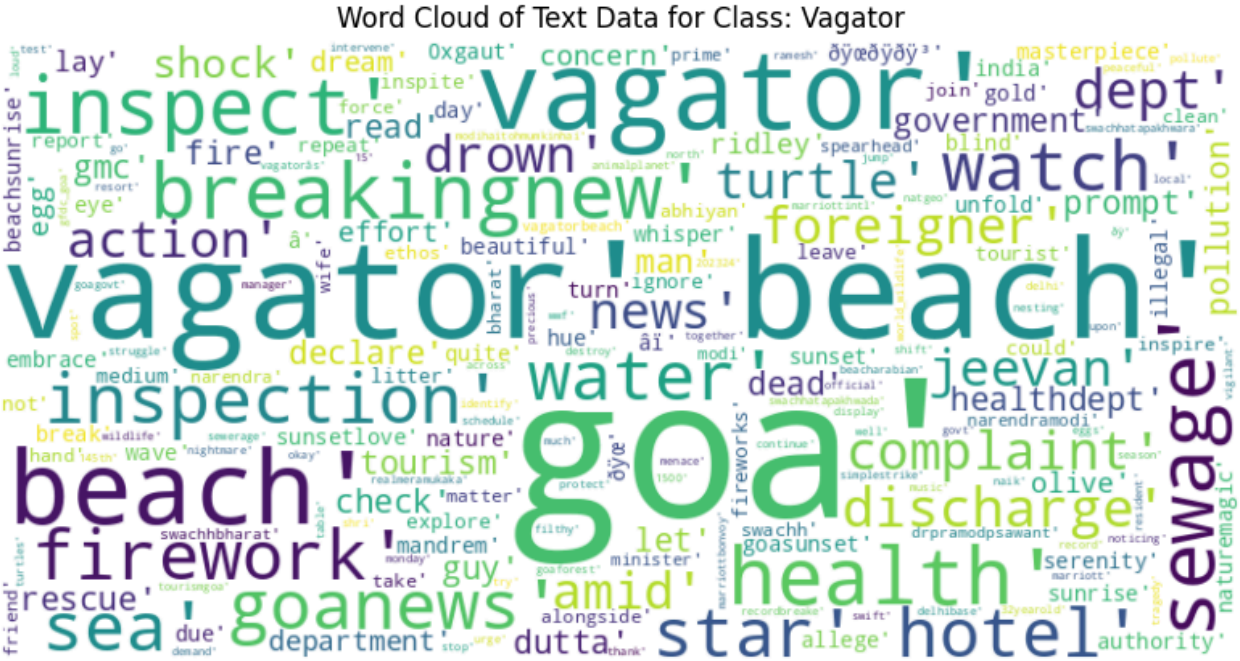


Figure 70: Word Cloud of class - Vagator of D5

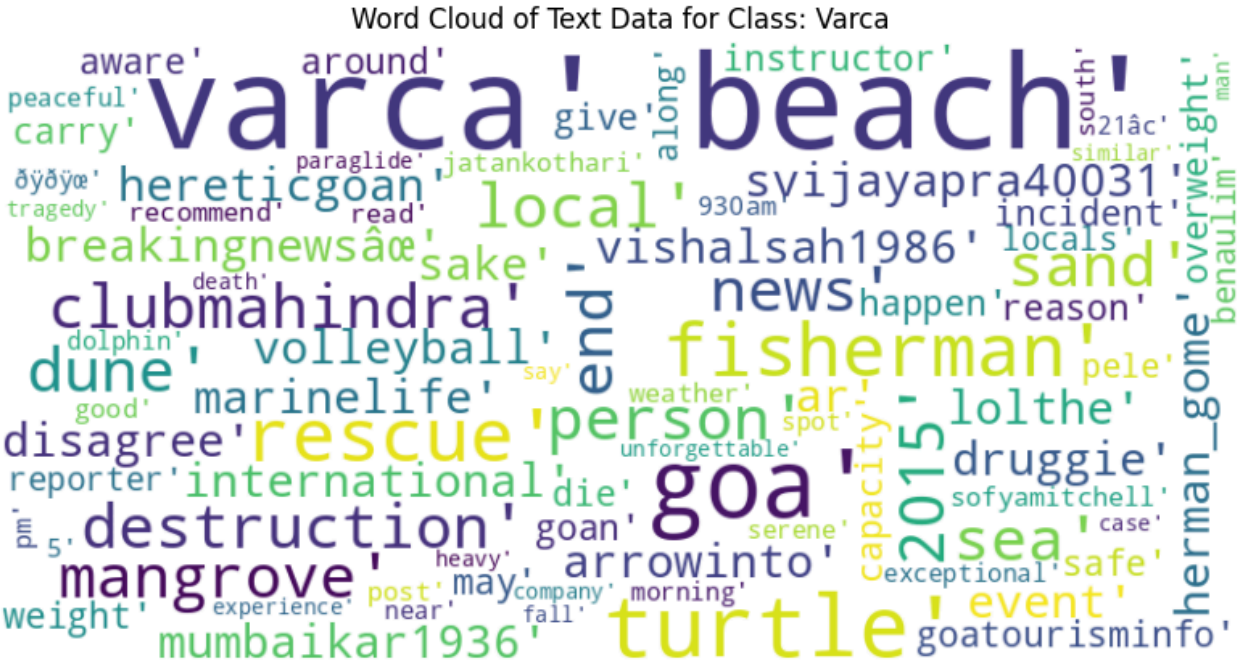


Figure 71: Word Cloud of class - Varca of D5

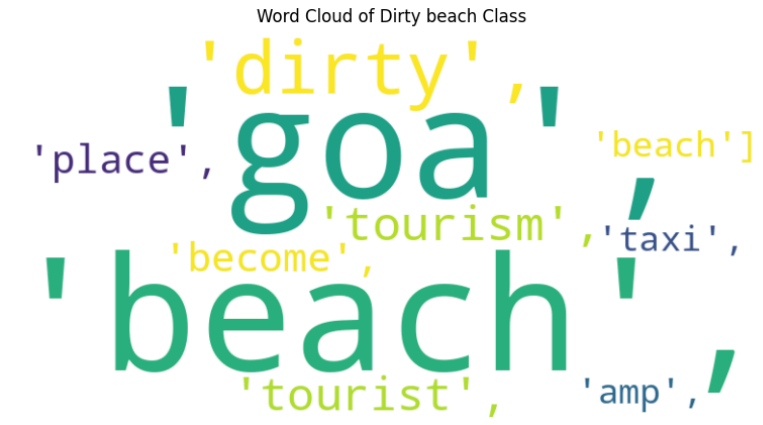


Figure 72: Word Cloud of Top 10 words of class - Dirty Beach of D5

Class: Dirty beach

	Word	Count
0	'beach',	43
1	'goa',	41
2	'dirty',	40
3	'tourism',	12
4	'tourist',	11
5	'become',	8
6	'place',	8
7	'beach']	6
8	'amp',	6
9	'taxi',	6

Figure 73: Count of Top 10 words of class - Dirty Beach of D5

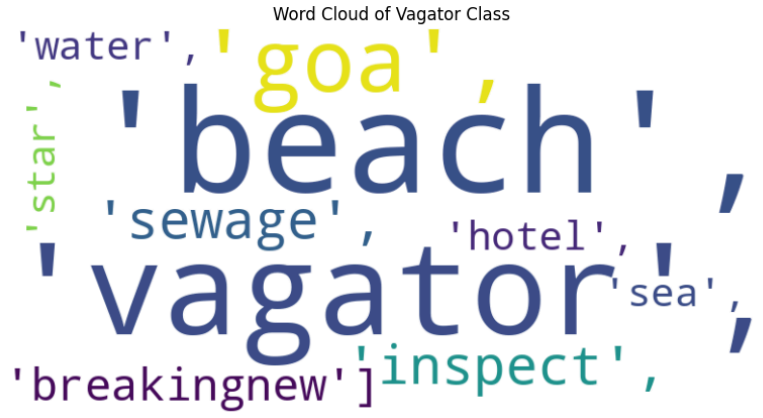


Figure 74: Word Cloud of Top 10 words of class - Vagator of D5

Class: Vagator

	Word	Count
0	'beach',	26
1	'vagator',	26
2	'goa',	23
3	'inspect',	5
4	'sewage',	5
5	'breakingnew']	5
6	'star',	4
7	'hotel',	4
8	'water',	4
9	'sea',	4

Figure 75: Count of Top 10 words of class - Vagator of D5

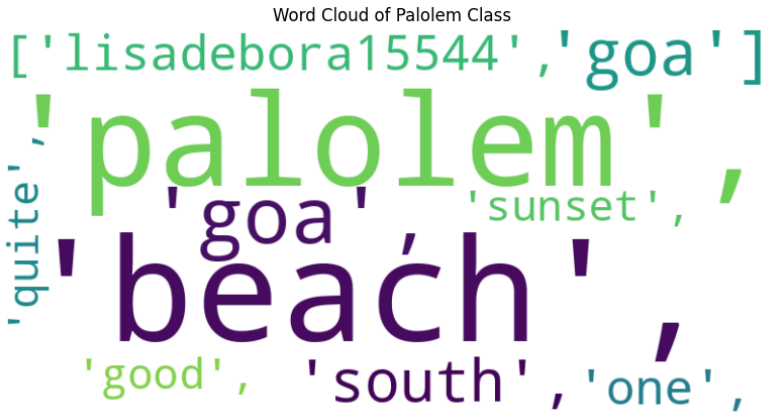


Figure 76: Word Cloud of Top 10 words of class - Palolem of D5

Class: Palolem

	Word	Count
0	'beach',	18
1	'palolem',	14
2	'goa',	10
3	'goa']	6
4	'south',	5
5	['lisadebora15544',	4
6	'one',	4
7	'good',	3
8	'quite',	3
9	'sunset',	3

Figure 77: Count of Top 10 words of class - Palolem of D5

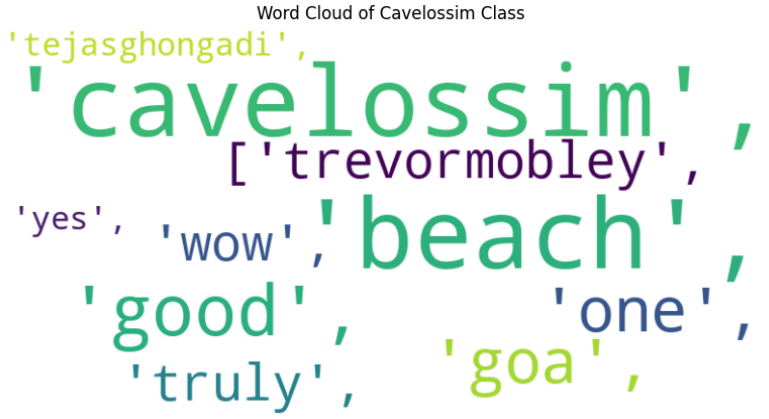


Figure 78: Word Cloud of Top 10 words of class - Cavelossim of D5

	Word	Count
0	'cavelossim',	9
1	'beach',	9
2	'good',	4
3	'one',	3
4	'goa',	3
5	['trevormobley',	2
6	'wow',	2
7	'truly',	2
8	'tejasghongadi',	2
9	'yes',	2

Class: Cavelossim

Figure 79: Count of Top 10 words of class - Cavelossim of D5

Figure 80: Word Cloud of Top 10 words of class - Baga of D5

	Word	Count
0	'beach',	17
1	'shack',	10
2	'baga',	7
3	'goa',	5
4	['baga',	4
5	'music',	3
6	'drivezyingoa',	3
7	'good',	2
8	'walk',	2
9	'also',	2

Class: Baga

Figure 81: Count of Top 10 words of class - Baga of D5

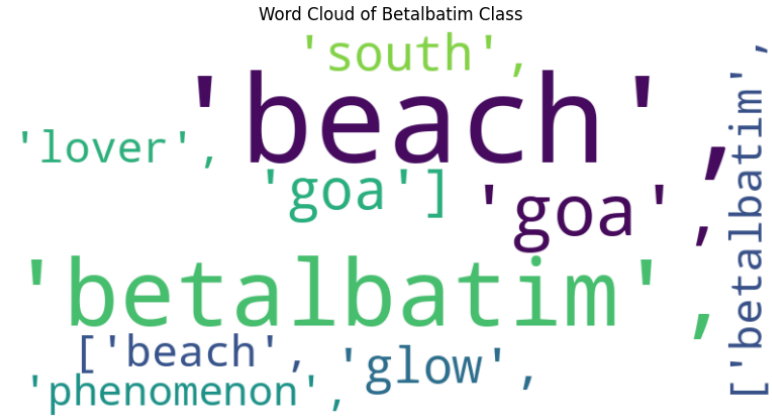


Figure 82: Word Cloud of Top 10 words of class - Betalbatim of D5

	Word	Count
0	'beach',	33
1	'betalbatim',	17
2	'goa',	9
3	'goa']	5
4	'glow',	4
5	'south',	4
6	'lover',	3
7	['beach',	3
8	['betalbatim',	3
9	'phenomenon',	3

Class: Betalbatim

Figure 83: Count of Top 10 words of class - Betalbatim of D5

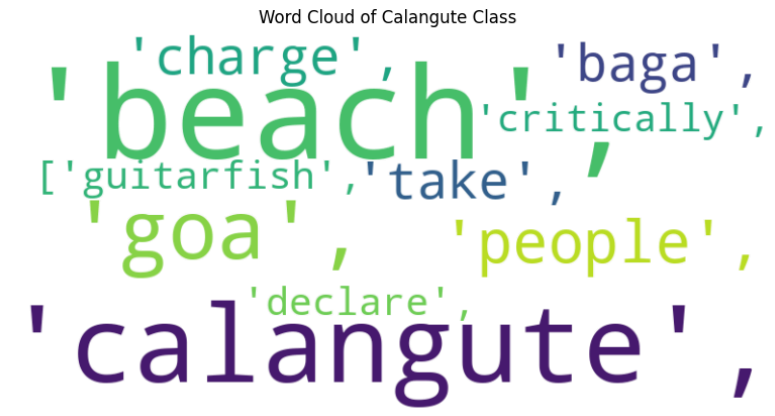


Figure 84: Word Cloud of Top 10 words of class - Calangute of D5

```
Class: Calangute
             Word  Count
0         'beach',     18
1     'calangute',     17
2           'goa',      9
3        'people',      4
4          'take',      3
5        'charge',      3
6          'baga',      3
7   ['guitarfish',      2
8       'declare',      2
9    'critically',      2
```

Figure 85: Count of Top 10 words of class - Calangute of D5

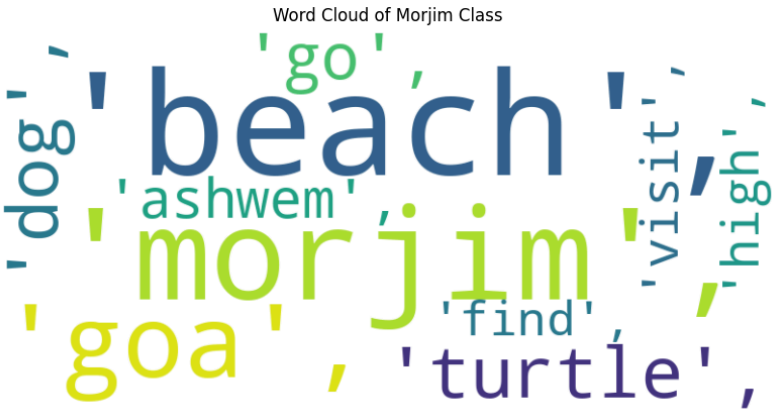


Figure 86: Word Cloud of Top 10 words of class - Morjim of D5

```
Class: Morjim
          Word  Count
0    'beach',     46
1   'morjim',     36
2      'goa',     29
3   'turtle',     12
4      'dog',     12
5       'go',     10
6   'ashwem',      7
7     'high',      6
8    'visit',      6
9     'find',      6
```

Figure 87: Count of Top 10 words of class - Morjim of D5

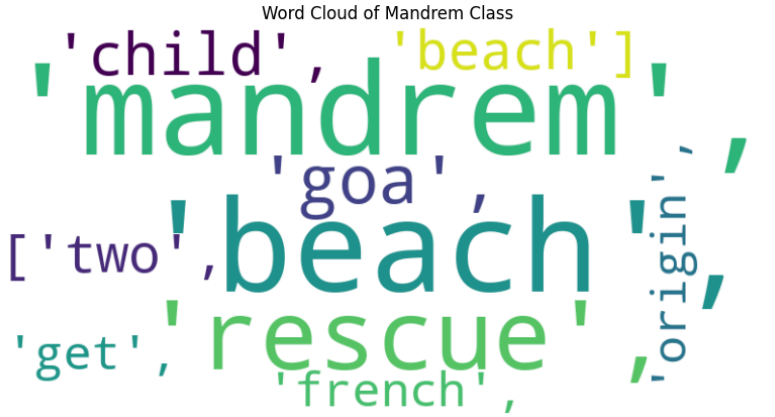


Figure 88: Word Cloud of Top 10 words of class - Mandrem of D5

```
Class: Mandrem
          Word  Count
0   'mandrem',     17
1     'beach',     17
2    'rescue',      9
3       'goa',      8
4     'child',      6
5     'beach']      5
6      ['two',      5
7    'french',      4
8    'origin',      4
9       'get',      4
```

Figure 89: Count of Top 10 words of class - Mandrem of D5

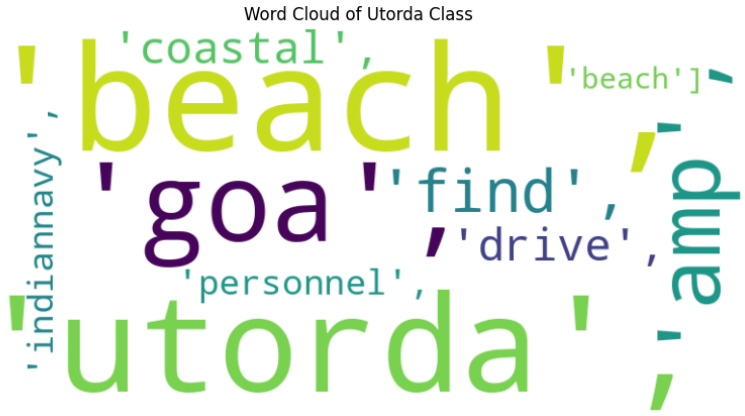


Figure 90: Word Cloud of Top 10 words of class - Utorda of D5

Class: Utorda

	Word	Count
0	'beach',	13
1	'utorda',	11
2	'goa',	7
3	'amp',	4
4	'find',	3
5	'coastal',	3
6	'drive',	3
7	'personnel',	3
8	'indiannavy',	3
9	'beach']	2

Figure 91: Count of Top 10 words of class - Utorda of D5

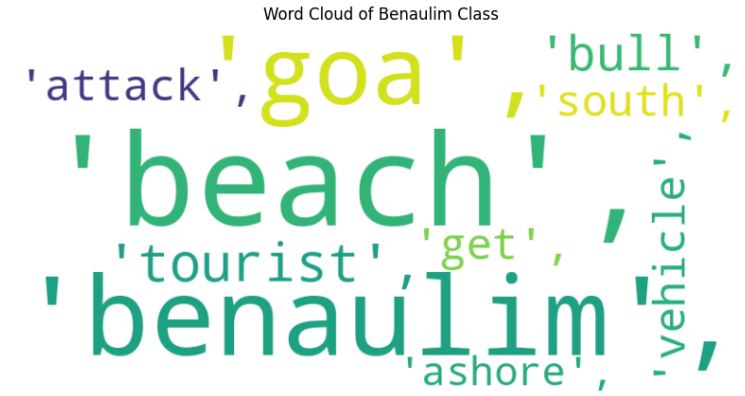


Figure 92: Word Cloud of Top 10 words of class - Benaulim of D5

Class: Benaulim

	Word	Count
0	'beach',	24
1	'benaulim',	20
2	'goa',	15
3	'tourist',	6
4	'bull',	5
5	'attack',	4
6	'south',	4
7	'vehicle',	4
8	'get',	4
9	'ashore',	4

Figure 93: Count of Top 10 words of class - Benaulim of D5

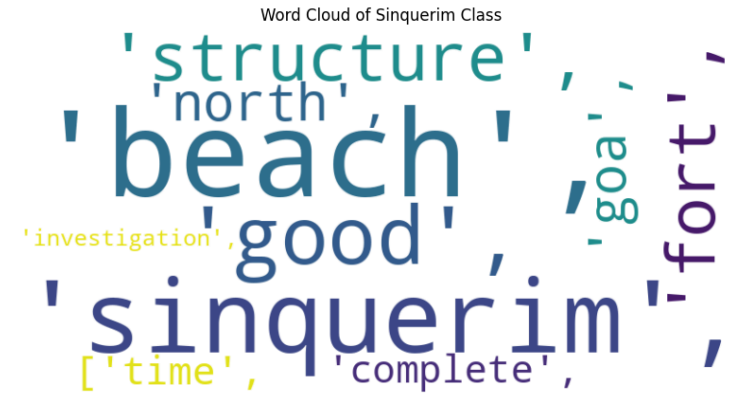


Figure 94: Word Cloud of Top 10 words of class - Sinquerem of D5

Class: Sinquerim

	Word	Count
0	'beach',	6
1	'sinquerim',	5
2	'good',	3
3	'structure',	2
4	'fort',	2
5	'north',	2
6	'goa',	2
7	['time',	1
8	'complete',	1
9	'investigation',	1

Figure 95: Count of Top 10 words of class - Sinquerim of D5

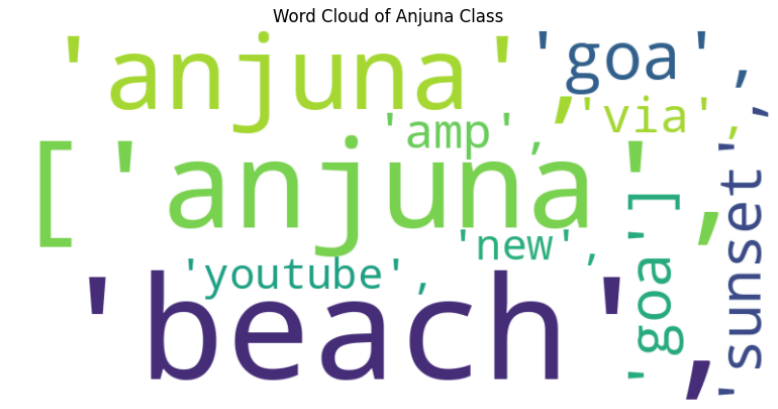


Figure 96: Word Cloud of Top 10 words of class - Anjuna of D5

	Word	Count
0	'beach',	10
1	['anjuna',	7
2	'anjuna',	6
3	'goa',	4
4	'sunset',	3
5	'goa']	3
6	'amp',	2
7	'via',	2
8	'youtube',	2
9	'new',	2

Class: Anjuna

Figure 97: Count of Top 10 words of class - Anjuna of D5

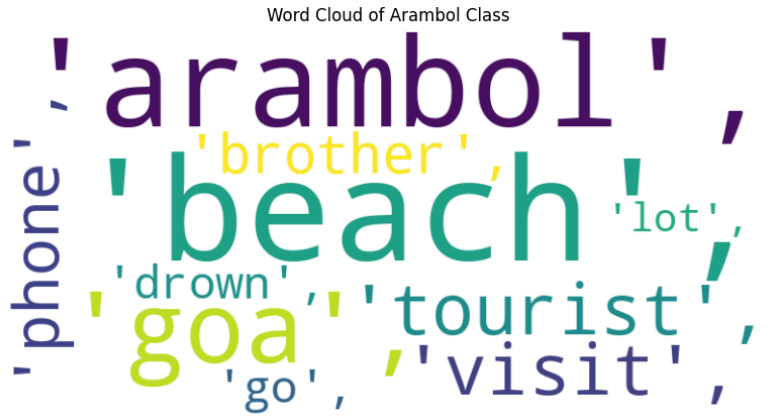


Figure 98: Word Cloud of Top 10 words of class - Arambol of D5

	Word	Count
0	'beach',	27
1	'arambol',	20
2	'goa',	12
3	'tourist',	5
4	'visit',	5
5	'phone',	4
6	'brother',	4
7	'drown',	4
8	'go',	4
9	'lot',	3

Class: Arambol

Figure 99: Count of Top 10 words of class - Arambol of D5

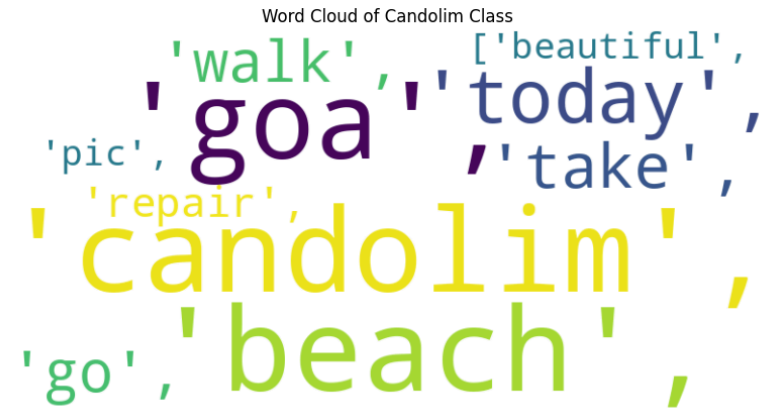


Figure 100: Word Cloud of Top 10 words of class - Candolim of D5

	Word	Count
0	'candolim',	7
1	'beach',	7
2	'goa',	6
3	'today',	3
4	'take',	2
5	'walk',	2
6	'go',	2
7	'repair',	2
8	['beautiful',	2
9	'pic',	2

Class: Candolim

Figure 101: Count of Top 10 words of class - Candolim of D5

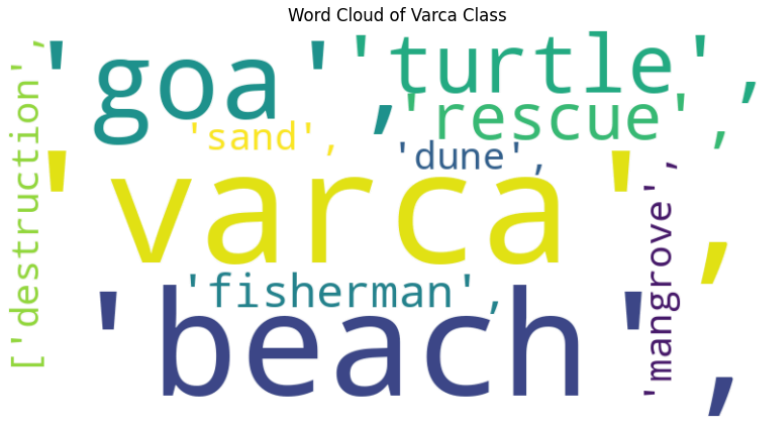


Figure 102: Word Cloud of Top 10 words of class - Varca of D5

Class: Varca

	Word	Count
0	'varca',	11
1	'beach',	10
2	'goa',	5
3	'turtle',	3
4	'rescue',	3
5	'fisherman',	3
6	['destruction',	2
7	'mangrove',	2
8	'sand',	2
9	'dune',	2

Figure 103: Count of Top 10 words of class - Varca of D5

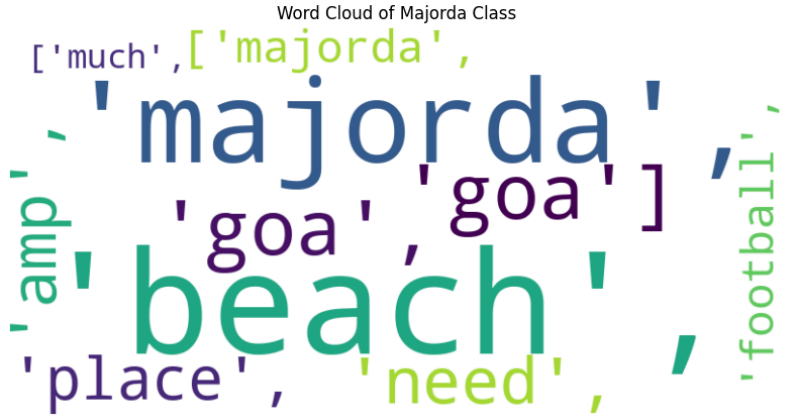


Figure 104: Word Cloud of Top 10 words of class - Majorda of D5

Class: Majorda

	Word	Count
0	'beach',	15
1	'majorda',	9
2	'goa']	3
3	'goa',	3
4	'need',	2
5	'amp',	2
6	'place',	2
7	['majorda',	2
8	'football',	2
9	['much',	1

Figure 105: Count of Top 10 words of class - Majorda of D5

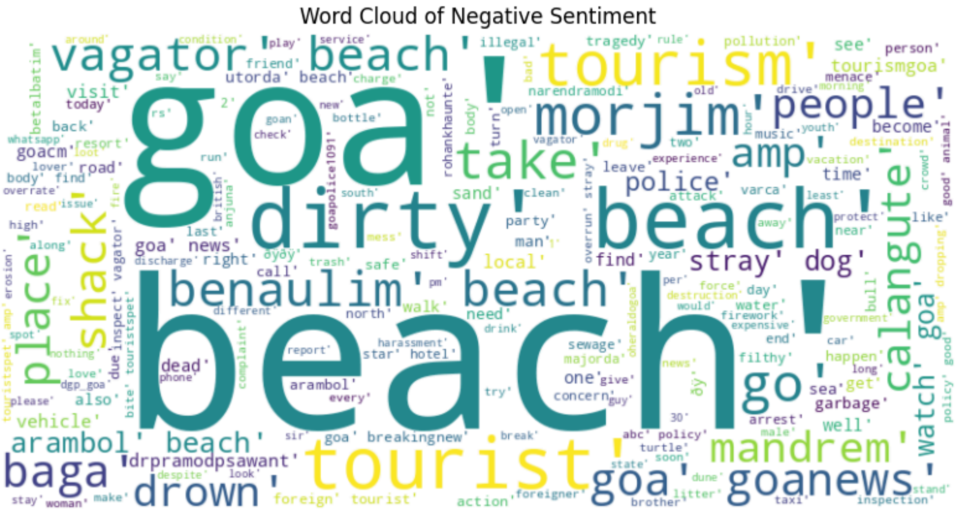


Figure 106: Word Cloud of Negative Sentiment of D5

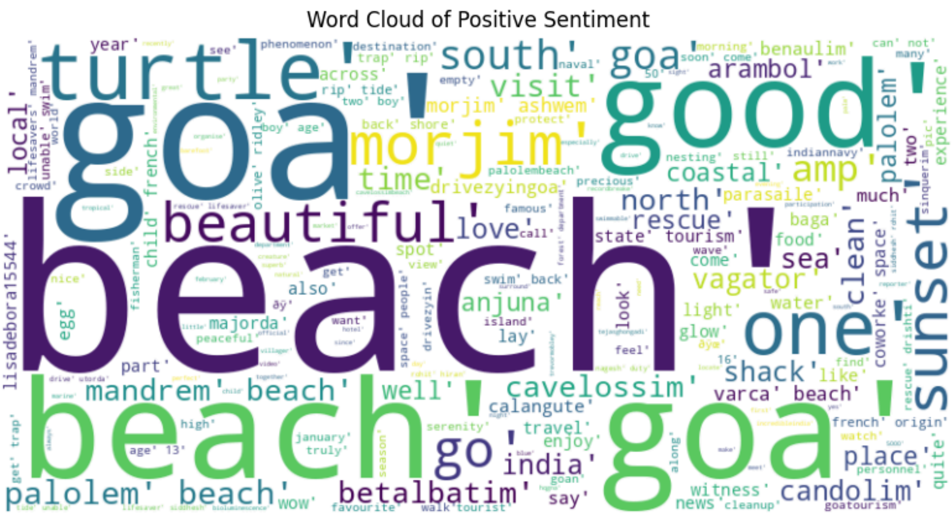


Figure 107: Word Cloud of Positive Sentiment of D5

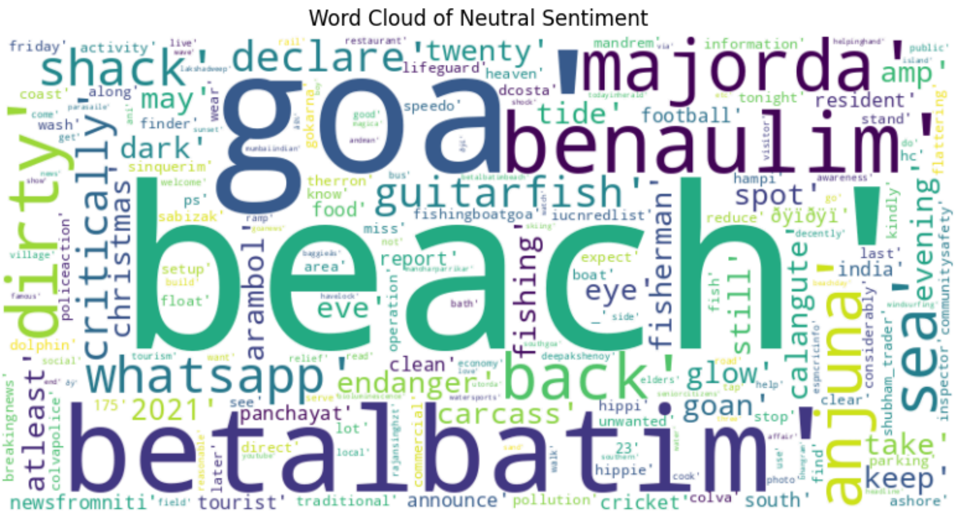


Figure 108: Word Cloud of Neutral Sentiment of D5

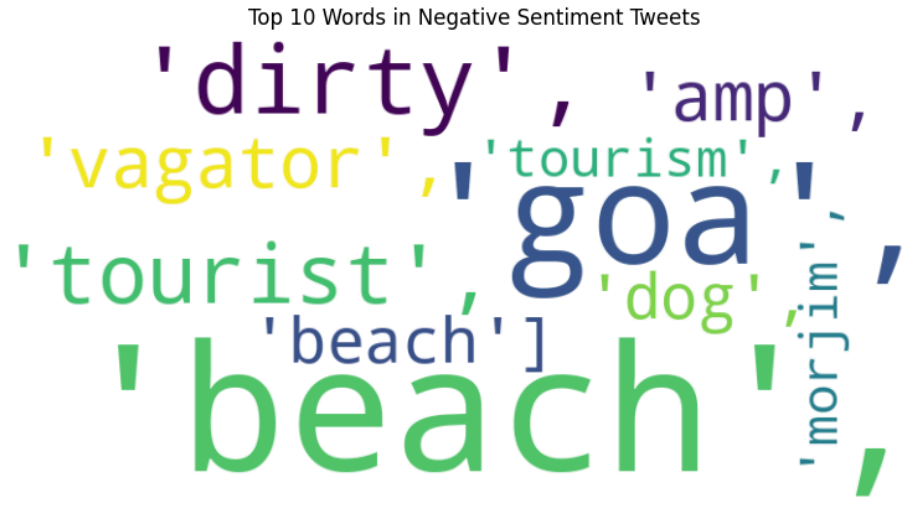


Figure 109: Word Cloud of Top 10 words of Negative Sentiment of D5

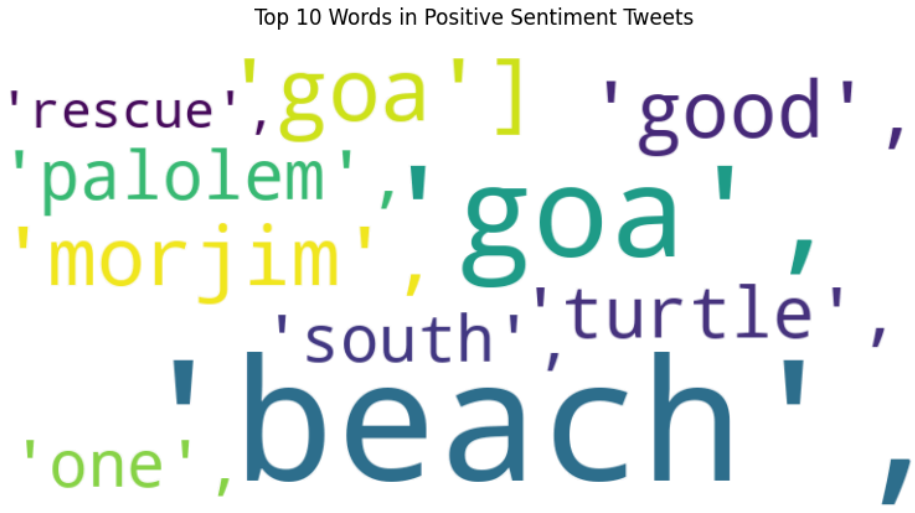


Figure 110: Word Cloud of Top 10 words of Positive Sentiment of D5

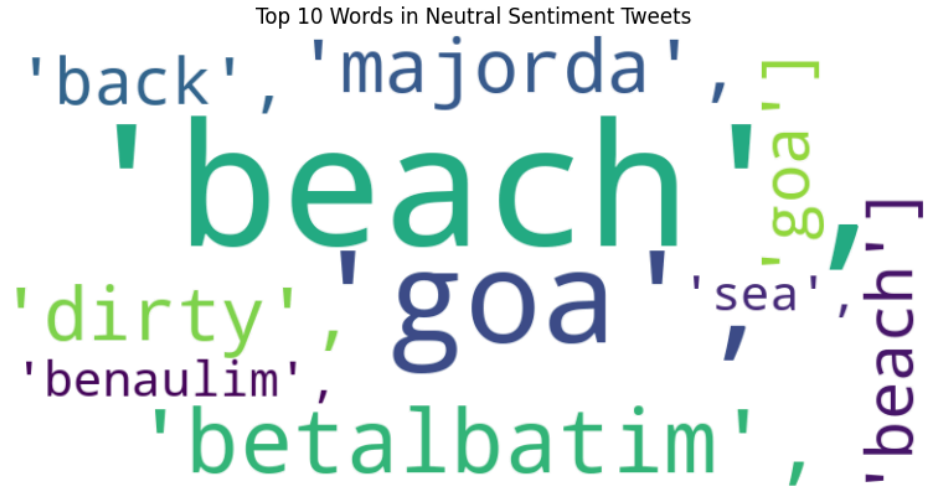


Figure 111: Word Cloud of Top 10 words of Neutral Sentiment of D5

Most common words for Negative sentiment:

	Word	Count
0	'beach',	183
1	'goa',	112
2	'dirty',	37
3	'tourist',	28
4	'vagator',	20
5	'amp',	19
6	'beach']	18
7	'dog',	17
8	'tourism',	17
9	'morjim',	16

Figure 112: Count of Top 10 words of Negative Sentiment of D5

Most common words for Positive sentiment:

	Word	Count
0	'beach',	125
1	'goa',	63
2	'goa']	23
3	'morjim',	20
4	'good',	19
5	'turtle',	15
6	'palolem',	14
7	'south',	13
8	'one',	13
9	'rescue',	12

Figure 113: Count of Top 10 words of Positive Sentiment of D5

Most common words for Neutral sentiment:

	Word	Count
0	'beach',	31
1	'goa',	16
2	'betalbatim',	6
3	'majorda',	4
4	'dirty',	4
5	'back',	3
6	'goa']	3
7	'beach']	3
8	'benaulim',	3
9	'sea',	3

Figure 114: Count of Top 10 words of Neutral Sentiment of D5

```
Topics generated by LDA:
Topic 1:
['goa', 'dog', 'shack', 'tourism', 'calangute', 'india', 'dirty', 'yesterday', 'amp', 'good']

Topic 2:
['betalbatim', 'morjim', 'goa', 'turtle', 'water', 'cricket', 'vehicle', 'benaulim', 'south', 'airport']

Topic 3:
['goa', 'palolem', 'good', 'clean', 'cavelossim', 'dirty', 'weather', 'report', 'tourism', 'newsfromniti']

Topic 4:
['goa', 'candolim', 'wash', 'mandrem', 'varca', 'news', 'ashore', 'beautiful', 'peaceful', 'south']

Topic 5:
['ðÿ', 'wow', 'news', 'goa', 'miss', 'whatsapp', 'rescue', 'lisadebora15544', 'reporter', 'cavelossim']

Topic 6:
['majorda', 'goa', 'utorda', 'youtube', 'spot', 'ðÿ', 'mandrem', 'child', 'good', 'visit']

Topic 7:
['drown', 'goa', 'vagator', 'body', 'sunset', 'utorda', 'inspection', 'watch', 'goanews', 'tourismgoa']

Topic 8:
['sunset', 'goa', 'vagator', 'mandrem', 'drivezyingoa', 'beautiful', 'benaulim', 'amid', 'shack', 'palolem']

Topic 9:
['goa', 'arambol', 'good', 'dirty', 'north', 'paradise', 'mandrem', 'expensive', 'morning', 'different']

Topic 10:
['goa', 'anjuna', 'dirty', 'tourist', 'sunset', 'ðÿïðÿï', 'visit', 'morjim', 'indians', 'north']
```

Figure 115: LDA on the column - text of D5

```
Top words for each topic:
Topic 1: dirty, goa, tourism, tourist, place, whatsapp, clean, garbage, soon, taxi
Topic 2: sunset, anjuna, goa, betalbatim, cavelossim, ðÿïðÿï, yesterday, sinquerim, know, âï
Topic 3: turtle, morjim, varca, sea, fisherman, local, benaulim, rescue, protect, news
Topic 4: palolem, goa, beautiful, good, south, lisadebora15544, time, palolembeach, paradise, serene
Topic 5: majorda, goa, place, betalbatim, cricket, erosion, need, corner, southgoa, break
Topic 6: vagator, inspect, health, sewage, breakingnew, inspection, goa, amid, discharge, complaint
Topic 7: arambol, drown, good, ðÿ, utorda, body, north, brother, tragedy, amp
Topic 8: dog, stray, good, goa, dropping, overrun, touristspet, bite, abc, policy
Topic 9: mandrem, rescue, child, goan, minor, france, reporter, french, origin, shore
Topic 10: shack, baga, drivezyingoa, drivezyin, drink, beautiful, time, betalbatim, music, crowded
```

Figure 116: NMF on the column - text of D5

3.6 BERT for Classification

3.6.1 Introduction

Natural Language Processing (NLP) has seen significant growth, with numerous research papers addressing various tasks like text classification, named entity recognition, and summarisation. There are generally two approaches to solving NLP problems:

1. Linguistic Approaches: These rely on specific features of the text considered relevant by domain experts, such as word combinations, n-grams, grammatical categories, and more. These features can either be manually crafted for a specific problem or derived from linguistic resources like ontologies. Linguistic methods are highly precise but tend to have lower recall because they work well only within limited contexts.

1. Machine Learning (ML) and Deep Learning Approaches: These methods automatically identify relevant features from annotated text corpora, typically using techniques like bag-of-words or n-grams. With the advancement of computing power, machine learning has become dominant in handling large volumes of text. These approaches typically require substantial data and have a statistical basis, with models like BERT and Transformers being prominent examples.

Traditional NLP faces challenges with multilingualism, as designing rules for one language doesn't necessarily transfer to another due to differences in sentence structure and alphabets. Approaches like the Universal Networking Language (UNL) try to address this, but they are complex to develop and require extensive expertise.

The paper focuses on comparing the BERT model, which pretrains deep bidirectional representations from unlabelled text before fine-tuning on labeled text for various NLP tasks, with traditional machine learning methods that utilize Term Frequency-Inverse Document Frequency (TF-IDF). This comparison aims to evaluate the trade-offs between precision and recall in these approaches.

We conducted four experiments on text classification using two classifiers: BERT and a traditional classifier based on TF-IDF, assessing their performance across different tasks. This work begins with a review of related studies, describes the models and experiments, presents the results, and concludes with insights and potential future research directions.

3.6.2 Experiments Overview

To evaluate the effectiveness of the BERT model compared to traditional machine learning approaches in NLP, we have set up four experiments.

Traditional NLP Approach: For the experiments performed in the study, they are utilizing the TfidfVectorizer from the sklearn Python 3 library to preprocess the text. In experiments three, they employ the Predictor from the auto ml module, and in the second experiment, they use H2OAutoML from the h2o module to identify the optimal model for the data. The first experiment aims to demonstrate the amount of effort required to achieve results comparable to those effortlessly obtained with the BERT model. This involves experimenting with various sklearn models and analyzing their outcomes in detail.

BERT Implementation: For the BERT model, they have used the pre-trained version available in the ktrain Python 3 module. This model requires a specific directory structure for operation: a main directory containing two subdirectories labeled 'train' and 'test'. Each of these should have subdirectories corresponding to different classes, named after the class they represent. Within each class directory, there should be '.txt' files containing the texts associated with that class, regardless of the filenames. [10]

3.6.3 Expiriment 1: Movie sentiment classification

In the first experiment, the IMDB dataset has been used, which is obtained from a specific website[10]. This dataset includes 50,000 movie reviews, split evenly with 25,000 reviews used for training the model and 25,000 for testing. The aim is to conduct sentiment analysis, a widely used supervised learning task for text classification.

Model	Accuracy
BERT	**0.9387**
Voting Classifier	0.9007
Logistic Regression	0.8949
Linear SVC	0.8989
Multinomial NB	0.8771
Ridge Classifier	0.8990
Passive Aggresive Classifier	0.8931

Table 6: Accuracy retrieved by the different methodologies in the IMDB experiment over the validation set.

3.6.4 Experiment 2: RealOrNot tweets experiment

In the second experiment, the RealOrNot dataset has been utilised consisting of tweets in English. The objective here is binary text classification, where the dataset includes tweets categorized into two classes: tweets pertaining to an actual disaster and tweets that are not related to any real disaster.

Model	Accuracy	Kaggle Score
BERT	**0.8361**	**0.83640**
H2OAutoML	0.7875	0.77607

Table 7 : RealOrNot experiment results.

The data has been classified from a Kaggle competition using the BERT model and achieved a score of 0.83640, as documented by Santiago González. For comparison, using traditional methods, the best classifier was the H2OStackedEnsembleEstimator from the h2o module. This classifier, specifically a Stacked Ensemble labeled as StackedEnsemble BestOfFamily AutoML 20200221 120302, scored 0.77607 in the competition.

3.6.5 Experiment 3: Portuguese news experiment

3.6.5 (a) Overview

Having observed that BERT significantly outperformed both an AutoML technique and other classical machine learning algorithms using a TF-IDF based vocabulary in English, it was decided to test the BERT model with a different language. A Portuguese news dataset, which consists of articles classified into nine categories: ambiente (environment), equilibrioesaude (balance and health), sobretudo (especially), educacao (education), ciencia (science), tec (tech), turismo (tourism), empreendedorsocial (social entrepreneur), and comida (food).

Experiment Setup: For this experiment, the columns containing article text and their respective classes were used. And then the directory was structured required to implement the BERT model, allocating 75% of the data for training and 25% for validation.

3.6.5(b) Results of the Experiment

Results: The results were as follows:

- BERT achieved an accuracy of 0.9093 and a Kaggle score of 0.91196.
- The Predictor from the auto ml module achieved an accuracy of 0.8480 and a Kaggle score of 0.85047.

The data classification for the Kaggle competition using BERT resulted in an accuracy score of 0.91196, as detailed by Santiago González. In contrast, the best traditional method, a GradientBoostingClassifier, achieved a score of 0.85047 in the competition.

Conclusion: These results demonstrate that the BERT model also performs exceptionally well in languages other than English, confirming its robustness and versatility in handling diverse linguistic datasets.[10]

Model	Accuracy
BERT	**0.9093**
Predictor (auto_ml)	0.8480

Table 8 :. Portuguese news experiment results.

From the Portuguese news experiment, several key conclusions can be drawn:

Language Versatility of BERT: The BERT model's strong performance in classifying Portuguese news articles extends its proven effectiveness beyond English, showcasing its versatility and robustness across different languages.

Superior Performance of BERT: BERT not only outperformed traditional machine learning models like the GradientBoostingClassifier but also advanced automated machine learning (AutoML) approaches. This indicates that BERT's deep learning architecture and training strategies are particularly effective for complex text classification tasks.

Impact of Pre-trained Models: The success of BERT in this experiment underscores the advantage of using pre-trained models in NLP tasks. These models, pre-trained on large datasets, bring a significant amount of learned knowledge that can be fine-tuned to specific tasks and datasets, often leading to better accuracy and efficiency.

Comparison with Traditional Methods: The traditional GradientBoostingClassifier and the AutoML model, while effective, did not reach the performance levels of BERT, highlighting the transformational impact of recent advancements in deep learning on NLP.

Potential for Broader Application: The results suggest that BERT and similar models have the potential to be applied effectively across a variety of languages and contexts, making them valuable tools for global applications in NLP.

Overall, the experiment reinforces the growing significance of models like BERT in the field of NLP, demonstrating their superior capability to handle and classify textual data across different languages and domains efficiently.

3.6.6 Conclusion

In this study[10], they have explored the BERT model alongside traditional NLP methods, such as using TF-IDF for feature extraction, to see how BERT performs in comparison as a standard approach for handling NLP tasks. they have presented four different NLP scenarios where BERT consistently outperformed traditional methods, demonstrating its effectiveness in general NLP challenges. Notably, implementing BERT proved to be less complex than traditional methods.
The advantage of transfer learning with BERT was particularly evident in our smallest dataset experiment, emphasising its efficiency.

3.7 Models for Sentiment Analysis

3.7.1 Introduction

When considering tools for sentiment analysis, models like BERT, DistilBERT, RoBERTa, and XLNet immediately come to mind due to their advanced capabilities in understanding and processing natural language. These models are built on the transformer architecture, which allows for the handling of complex dependencies in text data. BERT and its derivatives like RoBERTa and DistilBERT have been pre-trained on large corpora, providing a deep semantic understanding of language, which is crucial for accurately gauging sentiment. XLNet introduces an improved training methodology that captures the context better than traditional methods.

3.7.2 Model Study

BERT (Bidirectional Encoder Representations from Transformers): Developed by Google AI, BERT is a groundbreaking model in the NLP field due to its use of bidirectional training of transformers. This means that BERT learns information from both the left and the right context of a token's position within the text, allowing it to understand the full context of a sentence. BERT has been proven to improve the performance of various NLP tasks like question answering, sentiment analysis, and language inference.

RoBERTa (Robustly Optimized BERT Approach): RoBERTa is an extension of BERT that modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates. This approach has been shown to improve the performance of BERT on multiple benchmarking tests. RoBERTa was developed by Facebook AI and demonstrates how robust optimization strategies can lead to better model performance.

DistilBERT (Distilled BERT): DistilBERT is a smaller, faster, cheaper, and lighter version of BERT. Developed by Hugging Face, it is designed to retain 95% of BERT's performance while being 40% smaller and 60% faster. DistilBERT is trained through a process called distillation, which involves training the smaller model to reproduce the behavior of the full-sized BERT model. This model is particularly useful when resources are limited or when requiring faster processing, such as on mobile devices.

XLNet: Developed by researchers from Google Brain and Carnegie Mellon University, XLNet is a generalized autoregressive pretraining method that outperforms BERT on several NLP benchmarks. Unlike BERT, which assumes independence between masked positions and suffers from a pretrain-finetune discrepancy, XLNet leverages the best of both autoregressive and autoencoding methods, capturing the bidirectional context by using a permutation-based training strategy that helps learn the dependency between the masked tokens.

3.7.3 Previous Studies

Emotion recognition is a specific area that involves extracting detailed emotions from text. This field is rapidly developing but faces challenges in processing complex text dependencies and executing parallel text processing. The introduction of transformer models has significantly advanced this field by improving how these challenges are addressed.

Specifically, the BERT model, launched by Google in 2018, has enhanced the capabilities of Natural Language Processing (NLP) tools by allowing for better language understanding, although it does have some limitations such as fixed input lengths and computational demands. To address these, models like XLNet, RoBERTa, and DistilBERT have been developed, improving upon BERT's foundation and addressing its shortcomings.

The paper we referred[6] explores how well BERT, RoBERTa, DistilBERT, and XLNet perform in detecting emotions from the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset. The analysis includes a comparison of these models based on accuracy, precision, and recall in classifying different emotional states within the dataset. To our knowledge, this is the first comparative study of these four models on the ISEAR dataset.

3.7.4 Data set for the study

The initial dataset contained several columns, but only those with individual responses and emotion labels were relevant for this study. These columns were extracted for further analysis. Data cleaning involved removing entries where emotion labels existed without corresponding textual responses, reducing the dataset size from 7,666 to 7,589 entries. Additionally, special characters, double spaces, tags, and other irrelevant elements were eliminated as they could degrade performance in emotion recognition. Common stop words were also removed, and the seven emotion categories were converted into numerical codes ranging from 0 to 6. [11]

For model training, the data was divided, allocating 80% for training and 20% for testing. The maximum sentence length was set at 200 words based on the length of the longest sentence found in the dataset, ensuring that no sentence was cut off and that all sentences were standardized to the same length through padding. Finally, the sentences were tokenized to generate the tokens needed for feeding into the fine-tuning processes of the selected models.

DATA DISTRIBUTION OF THE ISEAR DATASET

Emotion Class	Number of Examples
Anger	1096
Disgust	1096
Sadness	1096
Shame	1096
Fear	1095
Joy	1094
Guilt	1093
Total	7666

Table 9 :Data distribution of the ISEAR Dataset

3.7.5 Comparison of Models

The BERT-base-uncased model, consisting of twelve transformer blocks each with twelve head self-attention layers and 768 hidden layers, totaling approximately 110 million parameters, was utilized. Each sentence was processed individually, tokenized using the BERT tokenizer, and assigned input IDs. Special tokens, [CLS] for classification and [SEP] for separation, were added at the beginning and end of each sentence. An input attention mask was applied to distinguish padded from real tokens. The output from the [CLS] token at the final (12th) transformer layer was used for making classifications based on the transformed prediction probabilities.

The RoBERTa-base model, similar in structure to BERT but with 125 million parameters, was also used. Sentences were tokenized using the RoBERTa tokenizer, with tokens padded to a uniform length to ensure consistency. Sentence pair classifications were then derived from the token features.

In the same experiment, the DistilBERT-uncased model, which includes six transformer layers and has the same number of hidden layers and attention heads as BERT but with fewer transformer layers, was employed for a multi-classification task. The input texts were tokenised, converted to input IDs, padded, and then processed.

Lastly, the XLNet-base-cased model, which mirrors the layer and attention head configuration of BERT and RoBERTa, was used. Its unique tokenizer converted texts into tokens that were then padded to uniform lengths for subsequent classification tasks.[12]

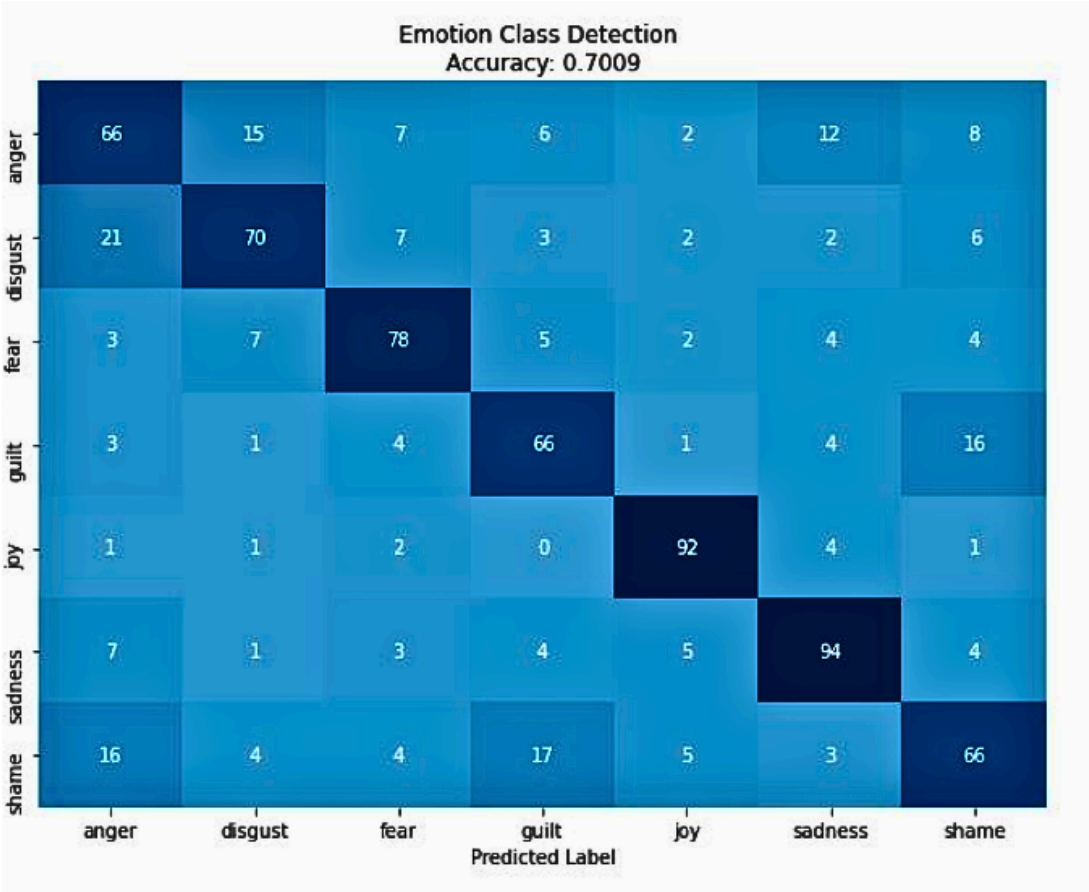


Figure 17: Confusion matrix for BERT

- The model seems to perform well at classifying anger and sadness emotions, with 94% and 92% accuracy respectively.
- The model struggles more with emotions like disgust, fear, and shame

Overall, the confusion matrix shows that the BERT model is able to classify emotions from text with some accuracy, but there is still room for improvement.

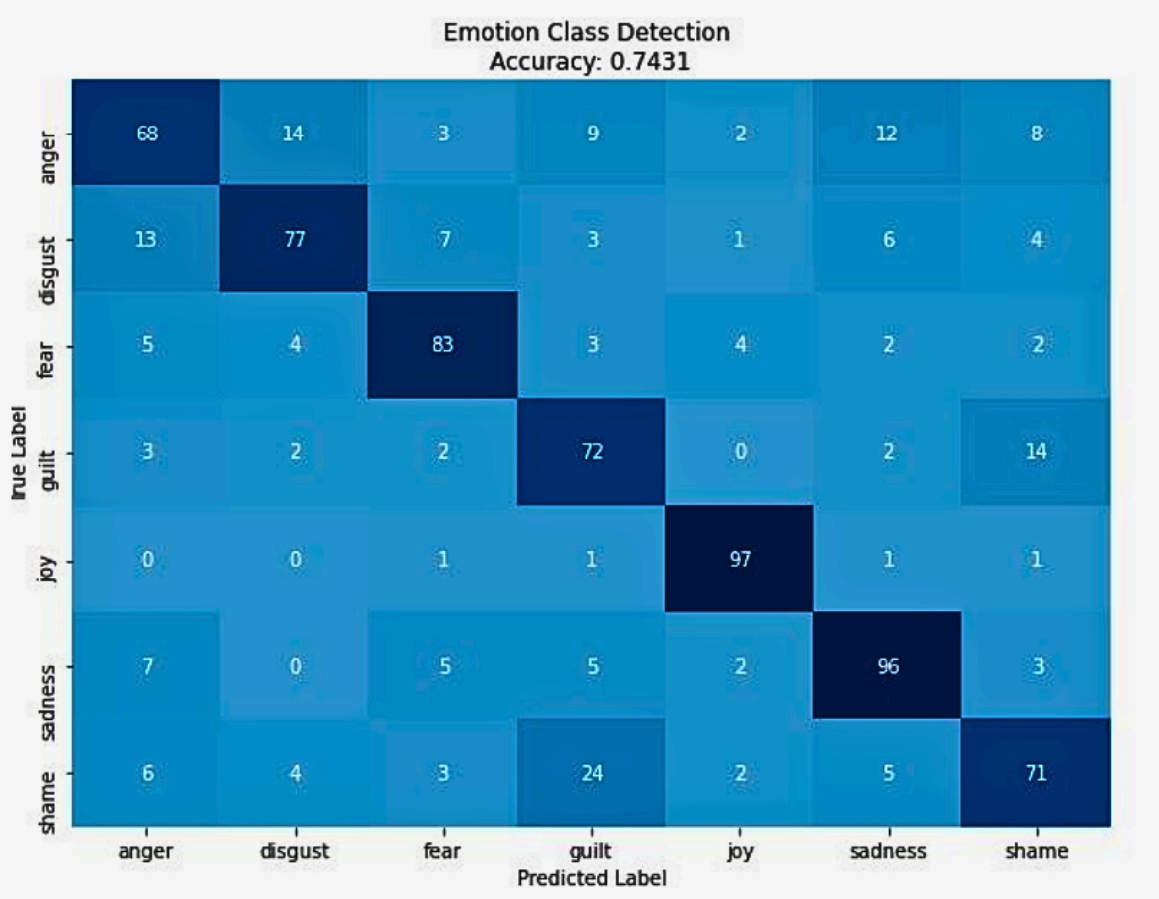


Figure 18: Confusion matrix for roBERTa

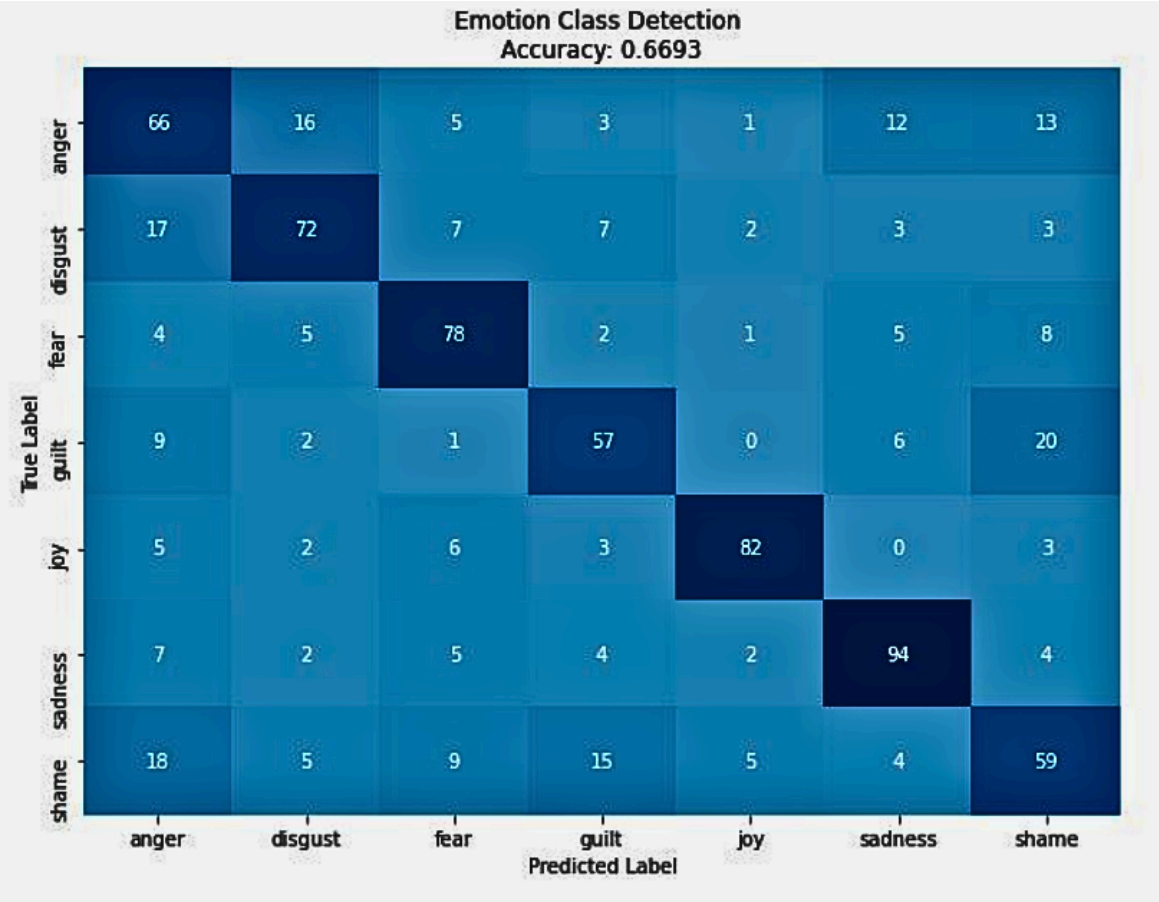


Figure 19: Confusion matrix for distilledBERT

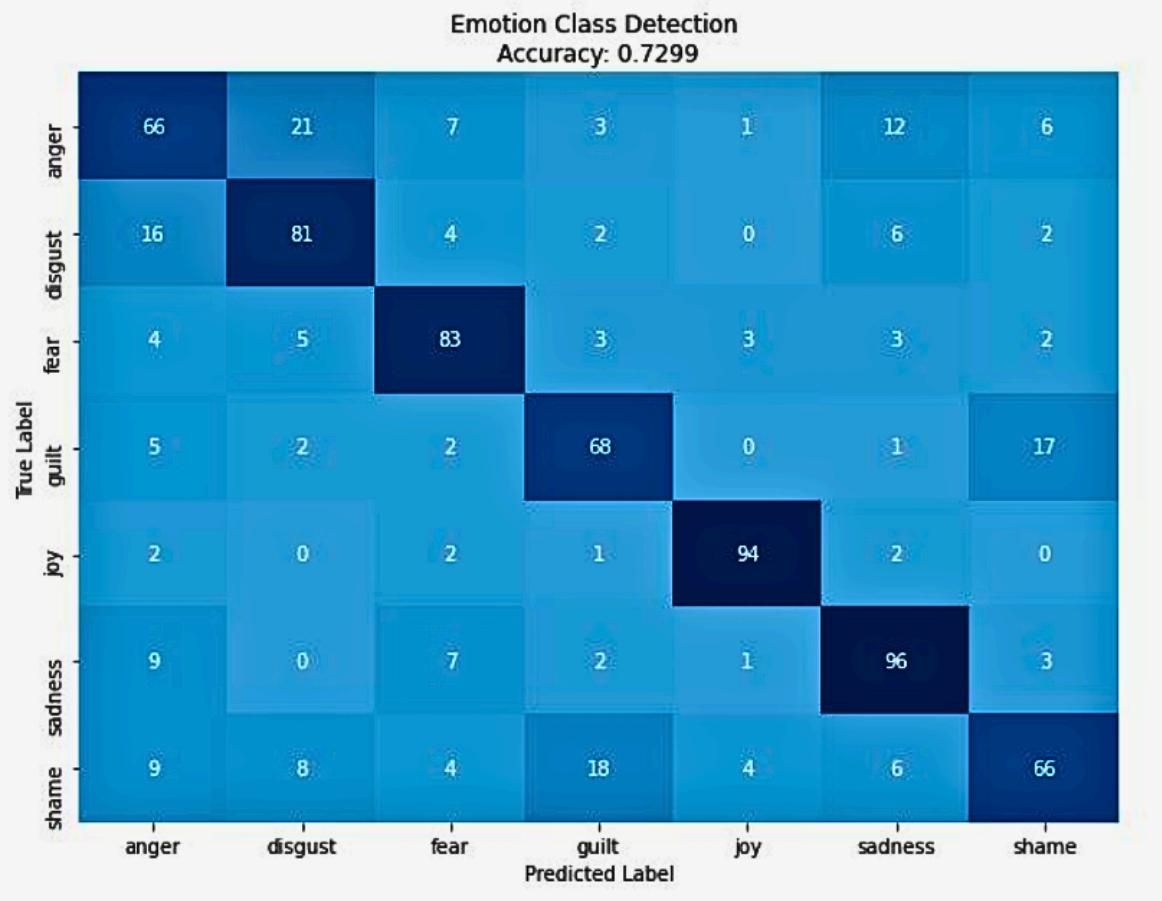


Figure 20: Confusion matrix for xlNet

COMPARISON OF PRECISION, RECALL AND F1-SCORES OF BERT, ROBERTA, DISTILBERT, AND XLNET ON THE ISEAR DATASET

Models	Anger			Disgust			Fear			Guilt			Joy			Sadness			Shame		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT	0.56	0.57	0.57	0.71	0.63	0.67	0.74	0.76	0.75	0.65	0.69	0.67	0.84	0.91	0.88	0.76	0.8	0.78	0.63	0.57	0.6
RoBERTa	**0.67**	**0.59**	**0.62**	**0.76**	0.69	**0.73**	**0.8**	**0.81**	**0.8**	0.62	**0.76**	0.68	0.9	**0.96**	**0.93**	**0.77**	**0.81**	**0.79**	**0.69**	**0.62**	**0.65**
DistilBERT	0.52	0.57	0.55	0.69	0.65	0.67	0.7	0.76	0.73	0.63	0.6	0.61	0.88	0.81	0.85	0.76	0.8	0.78	0.54	0.51	0.52
XLNet	0.59	0.57	0.58	0.69	**0.73**	0.71	0.76	**0.81**	0.78	**0.7**	0.72	**0.71**	**0.91**	0.93	0.92	0.76	**0.81**	**0.79**	**0.69**	0.57	0.63

3.7.6 Conclusion

The results, including confusion matrices for BERT, RoBERTa, DistilBERT, and XLNet, are displayed in Figures 2 through 5. Additionally, a table compares the precision, recall, and F1-scores for each emotion across these models, highlighting the effectiveness of each in recognizing specific emotions from the ISEAR dataset.

The confusion matrices reveal that RoBERTa has the highest emotion recognition accuracy, followed by XLNet, BERT, and DistilBERT. Notably, RoBERTa not only excels in accuracy but also demonstrates optimal computational efficiency compared to the others. Despite DistilBERT having the lowest accuracy, it was the fastest model, while XLNet was the slowest.

These findings underscore RoBERTa's suitability for emotion detection tasks, given its balanced performance in speed and accuracy. The classification report further supports RoBERTa's capability to handle various emotional contexts effectively, making it a strong candidate for such applications.[11]

3.8 Classification of tweets using Libraries

3.8.1 Classification using the re Library

Introduction
Utilizing regular expressions (regex) for text processing and pattern matching.
Regular expressions are character sequences defining search patterns, facilitating efficient text manipulation and extraction.

Reasoning
Regular expressions are chosen for their ability to identify specific patterns in text data.
They offer flexibility and power in text manipulation, suitable for tasks where predefined patterns aid classification.

Implementation
The model uses predefined keyword patterns to categorize tweets into segments like beaches, mountains, culture, etc.
Each segment has associated keywords, matched against tweet text using regex search operations.

Improvements
- Enhance accuracy by expanding and refining keyword lists.
- Optimize regex for better pattern recognition.
- Incorporate contextual information and domain-specific knowledge for robustness.
- Regular expressions are suitable for quick classification tasks with predefined patterns.
- Reliance on predefined patterns may limit adaptability to diverse or evolving data.

Results and Evaluation
The model achieves reasonable accuracy based on predefined keyword patterns.
Performance may be limited by specificity and coverage of these patterns.

	precision	recall	f1-score	support
Anjuna	1.00	1.00	1.00	3
Arambol	1.00	1.00	1.00	4
Baga	1.00	1.00	1.00	4
Benaulim	0.75	1.00	0.86	3
Betalbatim	1.00	0.86	0.92	7
Calangute	1.00	1.00	1.00	1
Candolim	1.00	0.75	0.86	4
Dirty beach	1.00	1.00	1.00	5
Majorda	1.00	1.00	1.00	1
Mandrem	1.00	1.00	1.00	3
Morjim	1.00	1.00	1.00	7
Palolem	1.00	1.00	1.00	4
Sinquerim	1.00	1.00	1.00	1
Utorda	1.00	1.00	1.00	2
Vagator	1.00	1.00	1.00	8
Varca	1.00	0.67	0.80	3
others	0.00	0.00	0.00	0
accuracy			0.95	60
macro avg	0.93	0.90	0.91	60
weighted avg	0.99	0.95	0.96	60

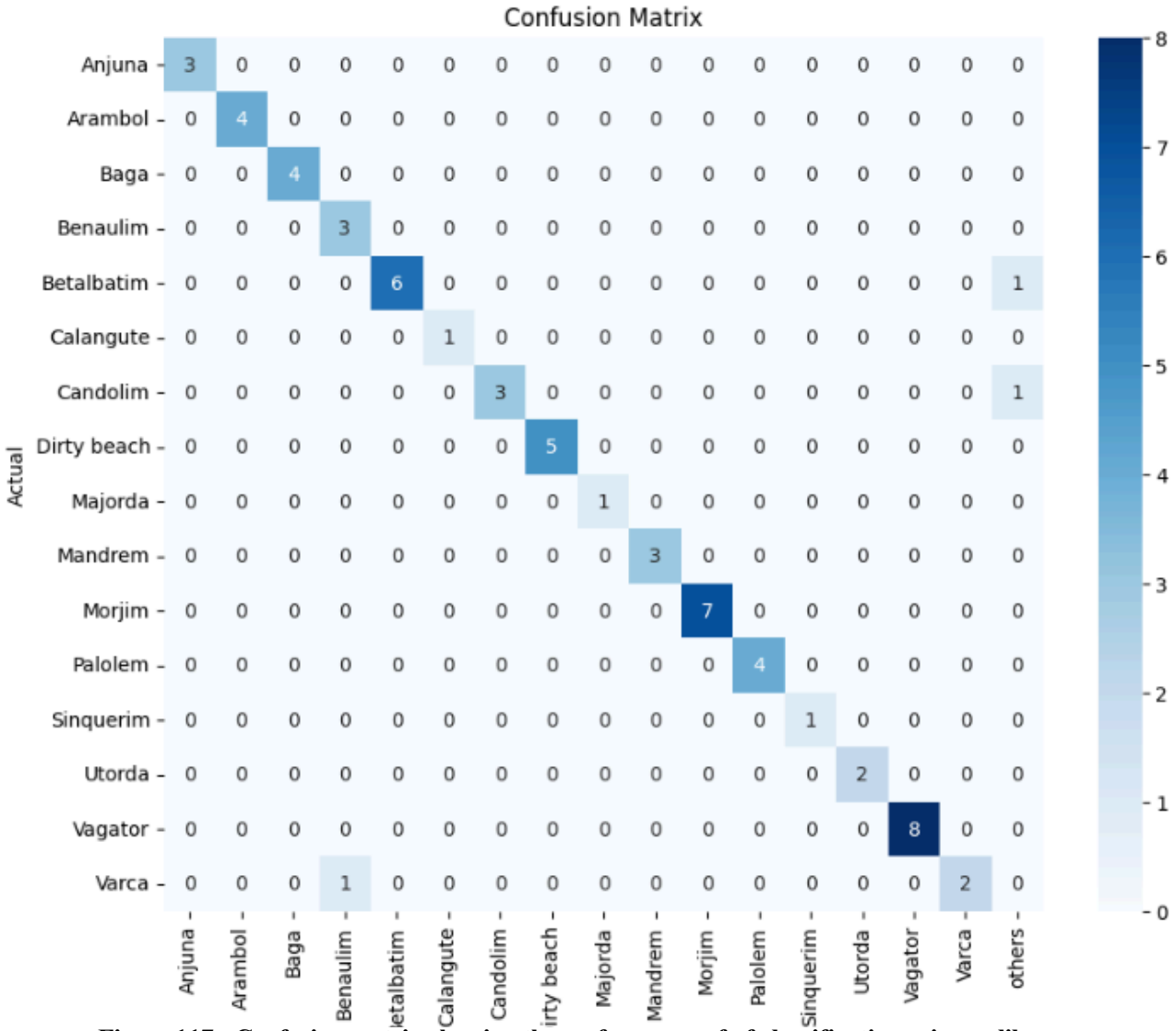


Figure 117: Confusion matrix showing the performance of of classification using re library

3.8.2 Classification model using NLTK Library

Introduction:
NLTK (Natural Language Toolkit) is utilized for text processing and analysis.
NLTK offers a range of NLP functionalities such as tokenization, stemming, lemmatization, part-of-speech tagging, and sentiment analysis.

Reasoning:
NLTK is chosen for its comprehensive toolkit and modularity.
It provides flexibility in implementing various NLP tasks and is widely used in educational and research settings.

Implementation:
The classification model preprocesses tweets using NLTK's techniques like **tokenization, stemming, and lemmatization.**
These techniques normalize text and extract meaningful features from tweets.
NLTK's sentiment analysis module may also be incorporated to analyze tweet sentiment.

Results and Evaluation
The model achieves satisfactory performance in categorising tweets based on content and sentiment. Evaluation metrics such as accuracy, precision, recall, and F1 score provide insights into model effectiveness.

Evaluation Report:

	precision	recall	f1-score	support
0	0.05	0.33	0.09	3
1	0.00	0.00	0.00	4
2	0.06	0.50	0.10	4
3	0.00	0.00	0.00	3
4	0.00	0.00	0.00	7
5	0.00	0.00	0.00	1
6	0.00	0.00	0.00	4
8	0.00	0.00	0.00	5
9	0.00	0.00	0.00	1
10	0.00	0.00	0.00	3
11	0.00	0.00	0.00	7
12	0.00	0.00	0.00	4
13	0.00	0.00	0.00	1
14	0.00	0.00	0.00	2
15	0.00	0.00	0.00	8
16	0.00	0.00	0.00	3
accuracy			0.05	60
macro avg	0.01	0.05	0.01	60
weighted avg	0.01	0.05	0.01	60

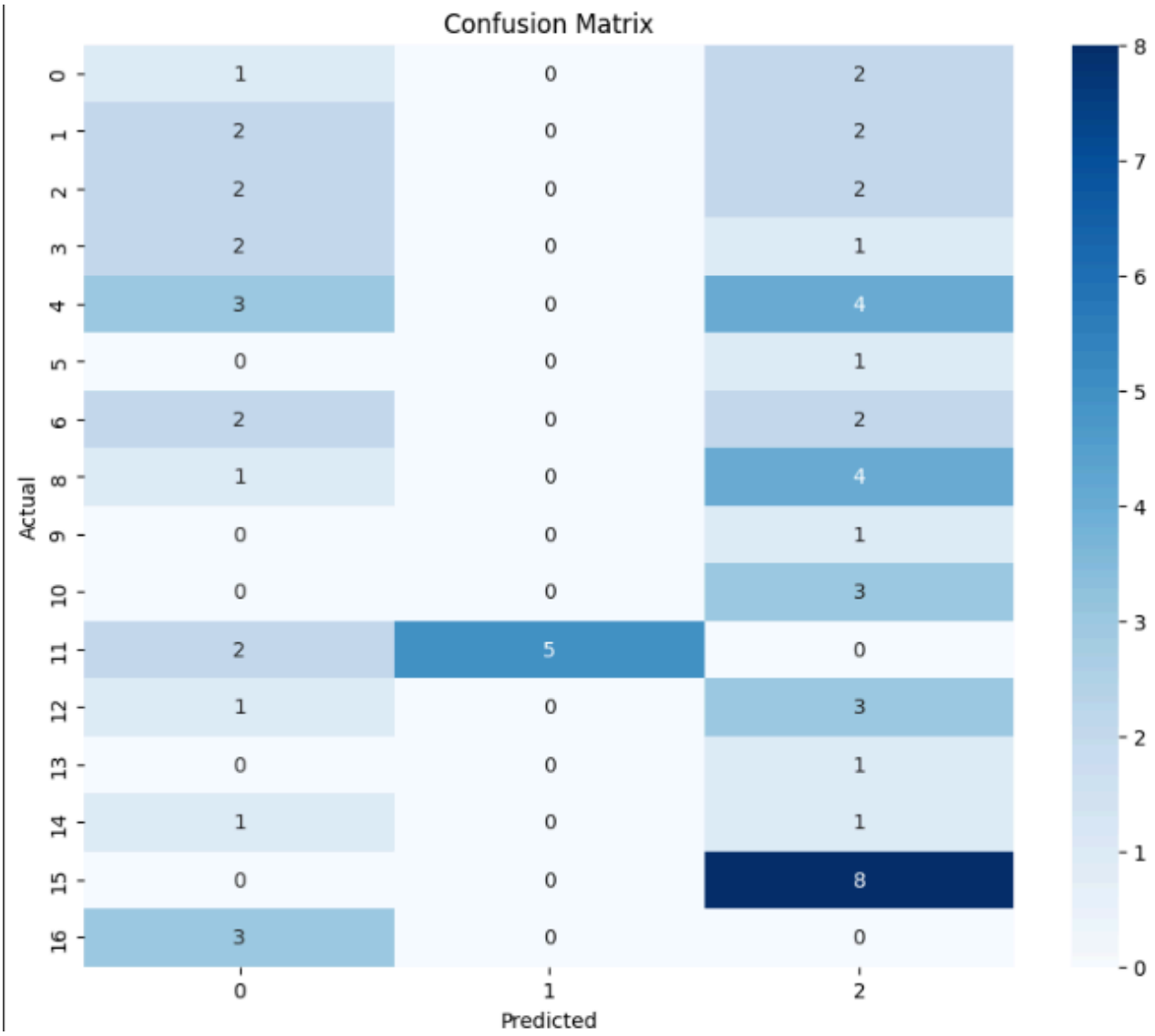


Figure 118: Confusion matrix showing the performance of of classification using NLTK library

3.8.3.Classification using re, NLTK, and Sci-Kit-learn Library

Improvements

Fine-tuning NLTK algorithms and adjusting preprocessing steps can enhance performance.
Combining regular expressions (re) for preprocessing with NLTK for NLP tasks and scikit-learn for machine learning.
Scikit-learn is a Python library for data mining and analysis, offering various machine-learning algorithms.

Reasoning

Integrating NLP with machine learning enables advanced text processing and analysis.
Regular expressions preprocess data, NLTK extracts features, and sci-kit-learn applies machine learning algorithms.

Implementation

Preprocessing uses regular expressions to remove noise and extract features.
NLTK performs tokenization, stemming, lemmatization, and feature extraction.
Scikit-learn employs algorithms like Naive Bayes, SVM, or decision trees for classification.

Results and Evaluation

Improved accuracy through combined preprocessing and machine learning.
Evaluation metrics (accuracy, precision, recall, F1 score) assess model performance and optimization opportunities.

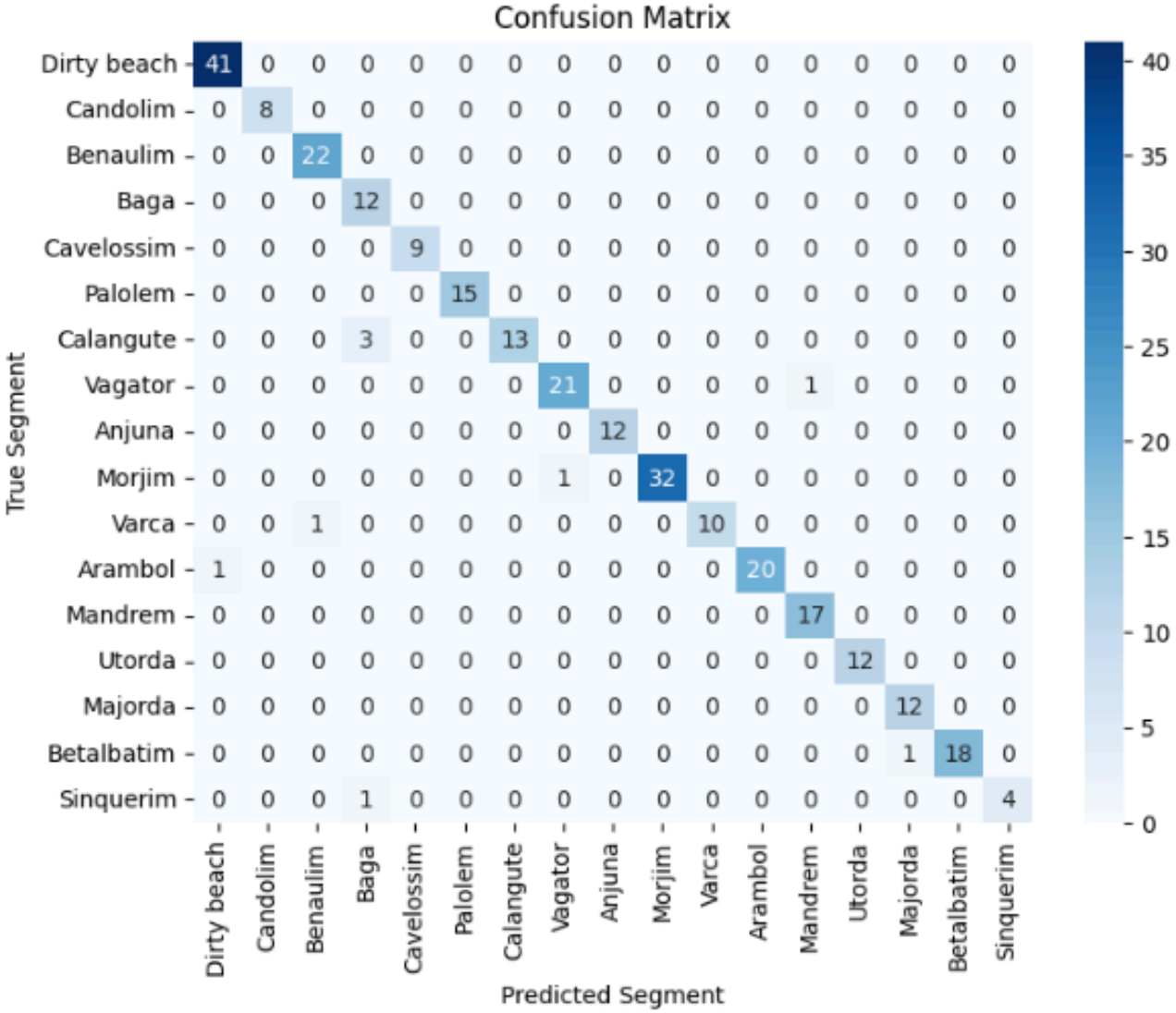


Figure 119: Confusion matrix showing the performance of of classification using NLTK and ski-kit learn library

Improvements

Fine-tuning ML algorithms, adjusting preprocessing, and exploring feature engineering enhance accuracy.
Hyperparameter optimization and cross-validation identify optimal model configurations.
Hybrid approach balances rule-based preprocessing with ML classification, leveraging both methods' strengths.
Offers flexibility for diverse data handling and achieves high classification accuracy.

3.8.4 Classification model using NLTK Library with Tokenization and Lemmatization

Introduction

Extending the NLTK-based classification model with tokenization and lemmatization for preprocessing.

Tokenization splits text into individual words, while lemmatization normalizes words to their base forms.

Reasoning

Tokenization and lemmatization address word variations and inconsistency, enhancing downstream NLP tasks' accuracy.

By preprocessing tweets with these techniques, the model aims to improve feature extraction and normalization.

Implementation

Preprocess tweets using NLTK's tokenisation and lemmatisation to tokenise text and normalize words.

Additional preprocessing steps like stop word removal and punctuation removal may be incorporated for further refinement.

Results and Evaluation

Improved accuracy achieved by reducing word variations and enhancing consistency in preprocessing.

Evaluation metrics (accuracy, precision, recall, F1 score) provide insights into tokenization and lemmatizations effectiveness in enhancing classification performance.

Improvements

Expand and fine-tune tokenisation and lemmatisation techniques, optimize parameter settings, and experiment with different ML algorithms.

Explore advanced NLP techniques such as word embeddings and deep learning models for further improvement.

The addition of tokenisation and lemmatisation enhances preprocessing quality, leading to better classification results compared to the previous NLTK-based model.

By addressing word variations and normalisation challenges, the model achieves improved accuracy in tweet classification tasks.

3.8.5 Classification using Spa-Cy Library

Introduction

This approach relies on Spa-Cy, a library, for text processing and analysis. NLP-based Spa-Cy offers efficient tools for tokenisation, named entity recognition (NER), dependency parsing, and lemmatisation.

Reasoning

Spa-Cy is chosen for its efficiency and performance in NLP tasks, surpassing NLTK in capabilities. By leveraging spa-Cy, the model aims for accurate classification through advanced NLP techniques.

Implementation

The classification model utilizes spa-Cy for tokenisation, NER, dependency parsing, and lemmatisation. It extracts meaningful features from tweet text and categorizes tweets into predefined categories.

Results and Evaluation

Achieves accurate classification through advanced NLP techniques and efficient processing with spa-Cy. Evaluation metrics (accuracy, precision, recall, F1 score) provide insights into performance and improvement areas.

```
Accuracy: 0.32
Classification Report:
              precision    recall  f1-score   support

      Anjuna       1.00      0.33      0.50        12
     Arambol       1.00      0.43      0.60        21
        Baga       0.75      0.50      0.60        12
    Benaulim       0.95      0.75      0.84        24
  Betalbatim       0.00      0.00      0.00        20
   Calangute       0.00      0.00      0.00        17
    Candolim       0.00      0.00      0.00         9
  Cavelossim       1.00      0.33      0.50         9
 Dirty beach       0.00      0.00      0.00        41
     Majorda       0.00      0.00      0.00        12
     Mandrem       0.00      0.00      0.00        18
      Morjim       1.00      0.86      0.93        37
     Palolem       1.00      0.38      0.55        16
   Sinquerim       0.00      0.00      0.00         6
      Utorda       0.00      0.00      0.00        13
     Vagator       0.93      0.64      0.76        22
       Varca       1.00      0.36      0.53        11
       other       0.00      0.00      0.00         0

    accuracy                           0.32       300
   macro avg       0.48      0.25      0.32       300
weighted avg       0.53      0.32      0.39       300
```

Improvements:
Fine-tune spa-Cy models, optimize feature extraction, and integrate deep learning techniques for enhanced accuracy.

Experiment with different spa-Cy components and configurations to identify the most effective approach.

Spa-Cy's advanced capabilities and efficiency make it superior for handling large text volumes and achieving high classification accuracy.

Leveraging spa-Cy's NLP functionalities, the model achieves accurate and efficient tweet classification, suitable for real-world applications.

Approach	Efficiency	Ease of Use	Flexibility
Classification using the re Library	Low	High	Low
Classification model using NLTK Library	Low	High	Medium
Classification using re, NLTK, and scikit-learn Library	High	Medium	High
Classification model using NLTK Library with Tokenization and Lemmatization	Medium	Medium	Medium
Classification using SpaCy Library	High	High	Hig

Table 19: Comparison chart of between different libraries

3.9 Classification of tweets using Algorithms

Algorithms offer numerous advantages over libraries in analysis: They allow for tailored customization, precise fine-tuning, a deeper understanding of methodologies, and optimization for performance, enabling analysts to craft sophisticated and efficient analytical solutions.

Customization:

- Algorithms allow us to tailor the analysis process to our specific needs, accommodating unique data structures, constraints, or objectives.
- This customization ensures that the analysis aligns precisely with our requirements, whereas libraries may offer generalized functions that lack the nuance required for specialized tasks.

Fine-tuning:

- With algorithms, we have the freedom to fine-tune parameters and settings to optimize the analysis. This flexibility enables experimentation with different approaches to achieve desired outcomes.
- In contrast, libraries may have limited options for customization, hindering our ability to optimize the analysis for specific scenarios.

Understanding:

- Implementing algorithms deepens our understanding of analysis techniques and underlying mathematical principles.
- By exploring algorithmic details, we gain insights into how different components interact, leading to informed decisions and adjustments.
- This understanding contributes to more insightful and accurate analyses compared to relying solely on library functions.

Performance:

- Algorithms can be optimized for performance based on data characteristics and available computational resources.
- Custom algorithms can achieve efficiency in time and memory usage, enabling quick and effective analysis, even with large datasets.
- While libraries offer pre-built functions optimized for general use cases, custom algorithms can often outperform them by leveraging tailored optimizations for the analysis task.

3.10 Classification using Clustering algorithms

We have aimed to classify tweets into different segments based on their content and themes. To achieve this, we considered using clustering algorithms, a type of unsupervised learning method.

Why clustering algorithms

Reducing Bias:

- We recognized the importance of reducing bias in our classification process. By opting for unsupervised learning methods like clustering, we aimed to minimize human bias and subjectivity in the classification process. Clustering algorithms group similar tweets together based on their intrinsic characteristics, without relying on predefined labels or categories.

Discovering Patterns:

- Clustering algorithms enable us to uncover hidden patterns and structures within the tweet data. By clustering tweets into segments, we can identify common themes, topics, or sentiments shared among tweets within each cluster. This helps in gaining a deeper understanding of the underlying patterns and trends present in the tweet dataset.

Scalability and Flexibility:

- Clustering algorithms are scalable and adaptable to various types of data. They can handle large volumes of tweets efficiently and can accommodate diverse tweet content, including text, images, and links. Moreover, clustering algorithms are flexible and can be applied iteratively to refine the segmentation process based on evolving requirements and feedback.

Unsupervised Learning:

- Unlike supervised learning methods that require labeled training data, clustering algorithms do not rely on labeled examples for training. This makes them particularly suitable for tasks where labeled data may be scarce or difficult to obtain. In our project, we tried to leverage the structure of the tweet data to automatically group tweets into meaningful segments without the need for manual annotation.

3.10.1 Clustering using K-means & TF-IDF Vectorization

Approach:

- Utilized K-means clustering, a partitioning algorithm, in conjunction with TF-IDF (Term Frequency-Inverse Document Frequency) vectorization for text data.

Why K-means & TF-IDF:

- K-means is renowned for its simplicity and efficiency, while TF-IDF vectorization transforms text data into numerical vectors based on term frequencies and inverse document frequencies.

Results:

- Initial clustering with K-means and TF-IDF yielded clusters primarily based on term frequencies, albeit lacking interpretability due to domination by common words across multiple tweets.

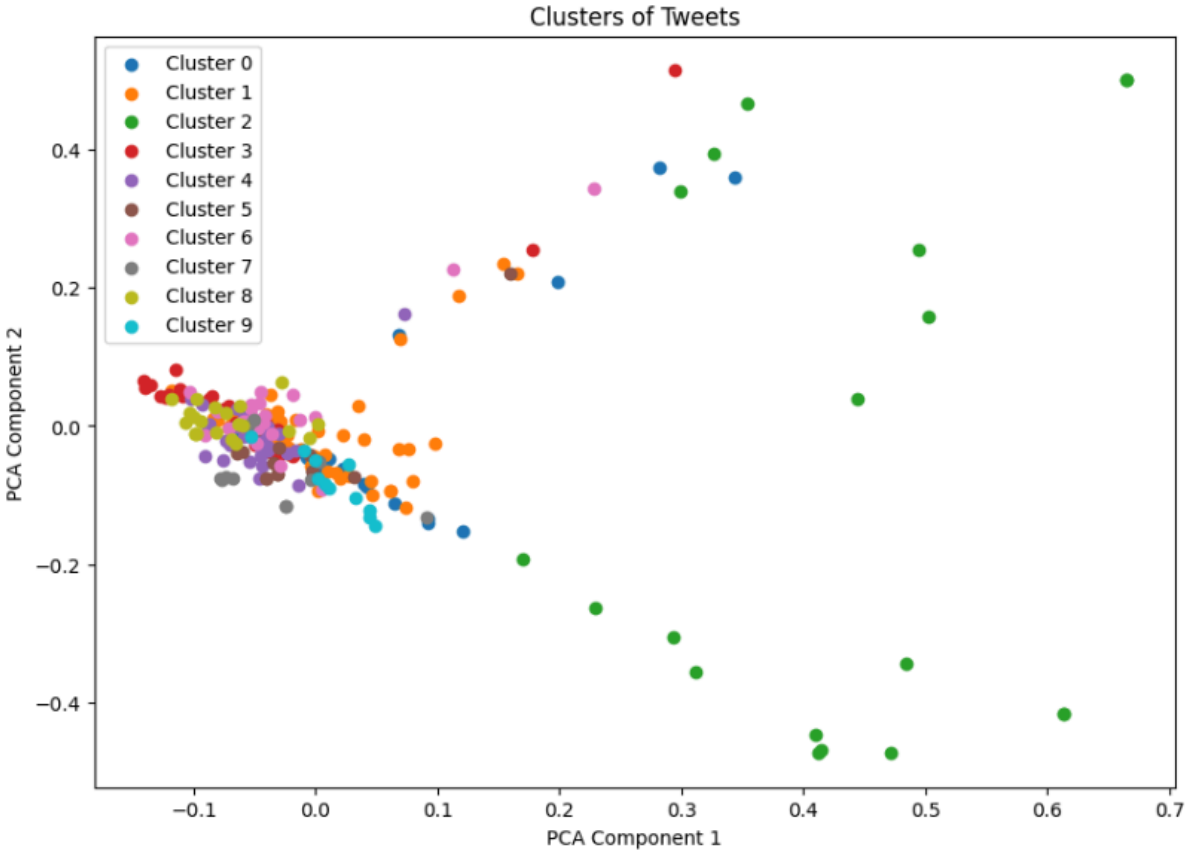


Figure 120: Visualization of a K-means & TF-IDF vectorization of tweets

Improvements:
To refine clustering, we experimented with diverse preprocessing techniques and fine-tuned TF-IDF parameters. Furthermore, alternative clustering algorithms were explored for comparative analysis.

Summary
K-means clustering with TF-IDF vectorization initially yielded clusters based on term frequencies but lacked interpretability due to common word domination.
To refine clustering, we experimented with preprocessing techniques and TF-IDF parameter fine-tuning. Alternative clustering algorithms were also explored to enhance interpretability.

3.10.2 Clustering using Hierarchical Clustering

Approach:

- Utilized hierarchical clustering, which constructs a cluster tree by recursively merging or splitting clusters based on similarity.

Why Hierarchical Clustering:

- Hierarchical clustering eliminates the need for specifying cluster numbers in advance and offers a hierarchical structure visualizable via a dendrogram, facilitating result interpretation.

Results:

- Hierarchical clustering generated clusters with clearer themes compared to K-means. The dendrogram aided the visualization of hierarchical cluster relationships, enabling the exploration of various granularity levels.

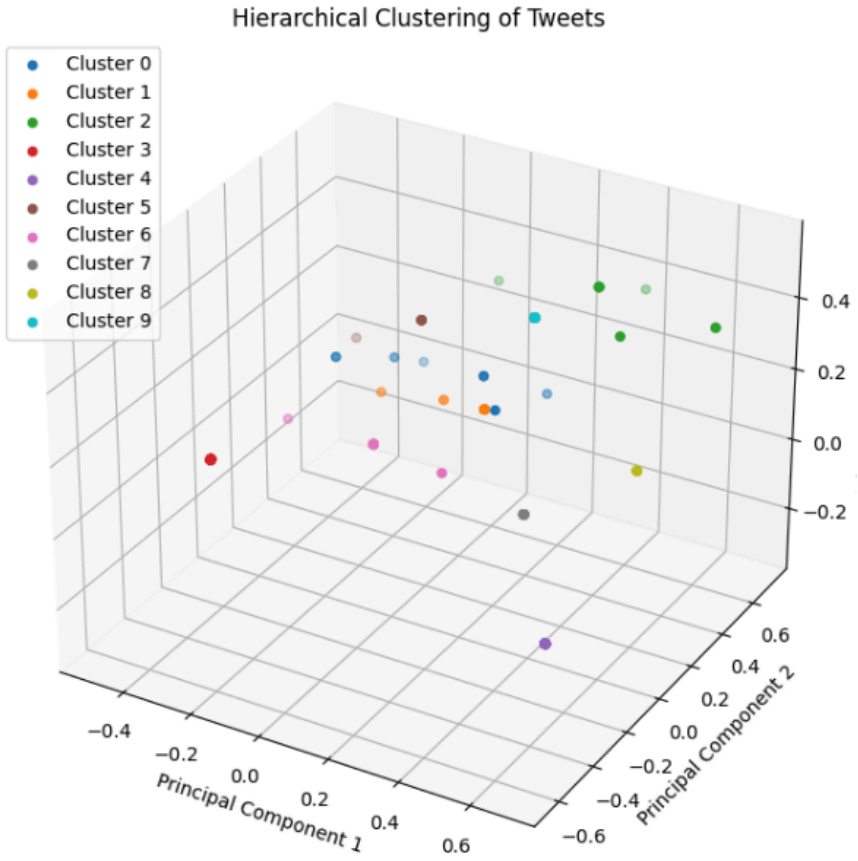


Figure 121: Visualization of a Hierarchical clustering of tweets

Improvements:

- Despite offering more interpretable results, hierarchical clustering encountered scalability and computational efficiency challenges, particularly with large datasets. Hence, a more scalable solution was pursued without sacrificing interpretability.

Summary:

- Hierarchical clustering, leveraging a hierarchical structure visualizable via a dendrogram, produced clearer clusters compared to K-means.
- Despite offering improved interpretability, scalability, and computational efficiency challenges were encountered, prompting exploration of more scalable solutions while maintaining interpretability.

3.10.3 Clustering using DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Approach:

- Employed DBSCAN, a density-based clustering algorithm, to group closely packed points while marking outliers as noise.

Why DBSCAN:

- DBSCAN obviates the need for predefining cluster numbers and can discern clusters of arbitrary shapes, rendering it apt for datasets with noise and outliers.

Results:

- DBSCAN adeptly detected clusters of varying densities and shapes within the tweet data, effectively handling outliers and noise to yield robust clustering outcomes.

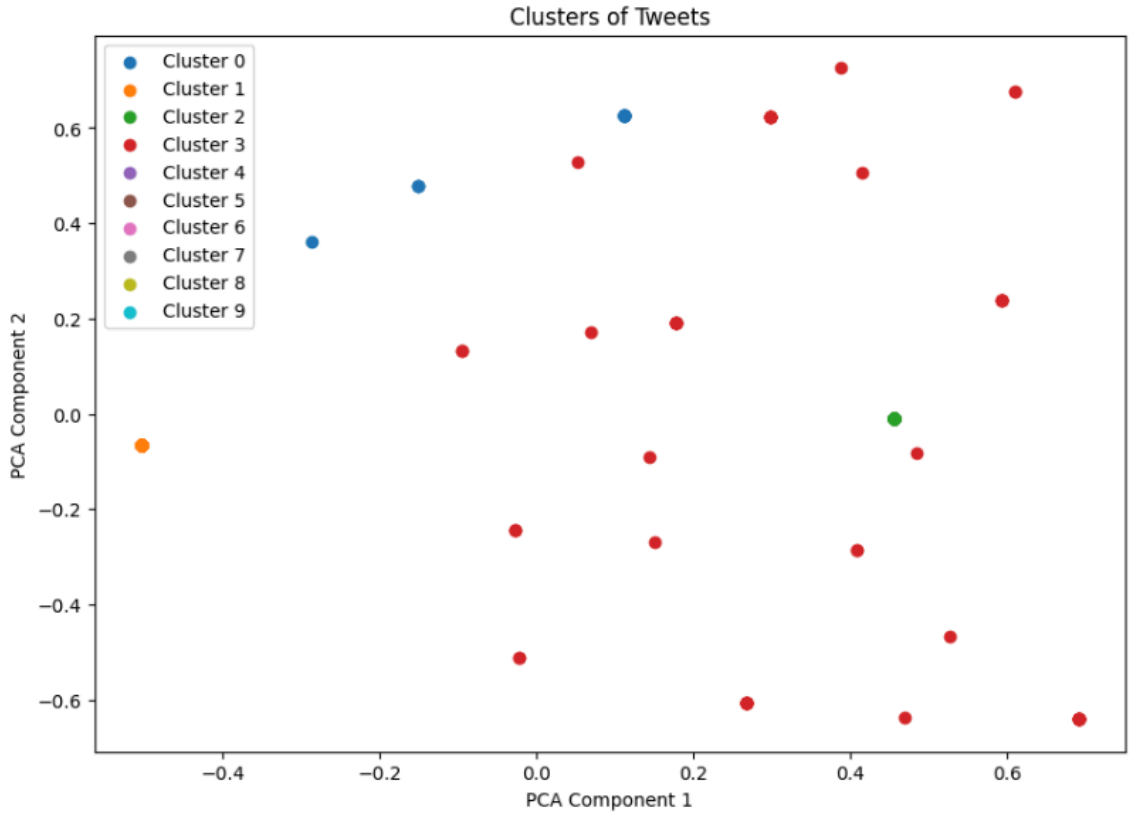


Figure 122: Visualization of a DBSCAN clustering of tweets

Improvements:

- Despite mitigating constraints of predefining cluster numbers and arbitrary shapes, DBSCAN encountered challenges in determining suitable parameters like epsilon (eps) and minimum samples. Tuning these parameters necessitated domain knowledge and iterative experimentation.

Summary:

- DBSCAN demonstrated effectiveness in detecting clusters of different densities and shapes within tweet data, handling outliers and noise efficiently.
- However, the algorithm encountered challenges in parameter selection, requiring iterative experimentation for optimal performance.
- Despite this, DBSCAN remains a valuable tool for clustering datasets with arbitrary shapes and varying densities.

3.10.4 Clustering using Non - Negative Latent Dirichlet Allocation (LDA)

Why it was used:

- LDA is a probabilistic model that represents documents as mixtures of topics, commonly employed for topic modeling and applicable to clustering text data.
- : LDA facilitates interpretability by associating documents with topics, enabling users to understand the underlying themes within the data.
- LDA's flexibility allows it to adapt to various domains and types of text data, offering versatility in clustering tasks.

How it helped:

- LDA uncovers latent topics in text data and assigns documents to clusters based on these topics, akin to NMF meaning LDA and NMF share similarities or are similar in some aspects.

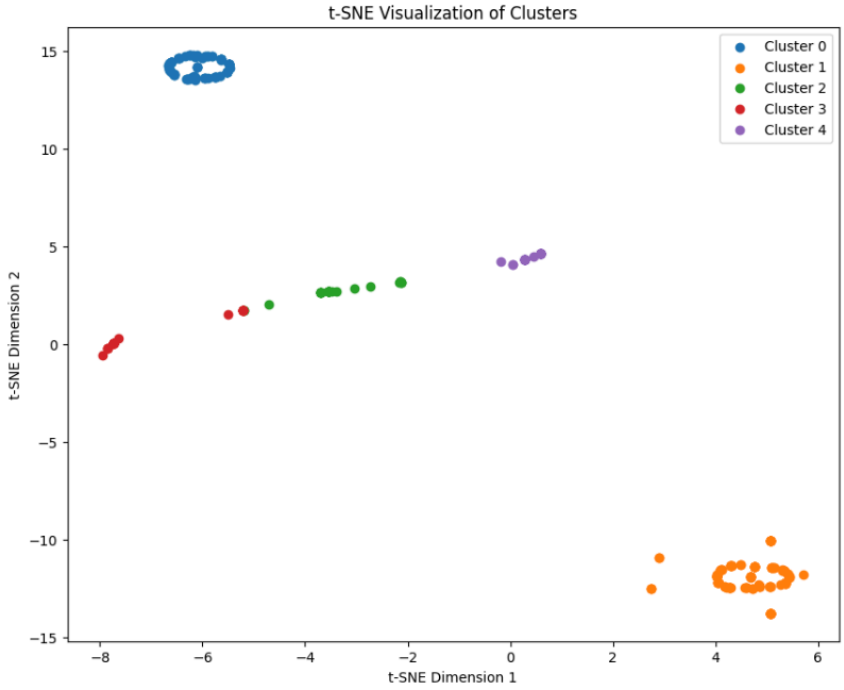


Figure 123 : Visualization of a LDA clustering of tweets

Problem:

- LDA necessitates pre-specifying the number of topics, and clustering quality can hinge on hyperparameter choices.

Solution:

- Addressing this involves experimenting with various topic numbers and fine-tuning hyperparameters using methods like cross-validation.

Summary

- LDA, a probabilistic model representing documents as mixtures of topics, is commonly used for topic modeling and applicable to clustering text data.
- It enhances interpretability by associating documents with topics, enabling users to grasp underlying themes.
- Its flexibility allows adaptation to various domains, offering versatility in clustering tasks. LDA uncovers latent topics in text data, akin to NMF, but necessitates pre-specifying the number of topics, impacting clustering quality.

3.10.5 Clustering using Agglomerative Clustering

Why it was used:

- Agglomerative clustering is a hierarchical technique that merges clusters iteratively, producing a dendrogram to visualize the process.
- Agglomerative clustering provides intuitive results that are easy to interpret, especially with the visual aid of dendrograms.

How it helped:

- It aids in exploring hierarchical structures within the data, offering insights into cluster relationships.

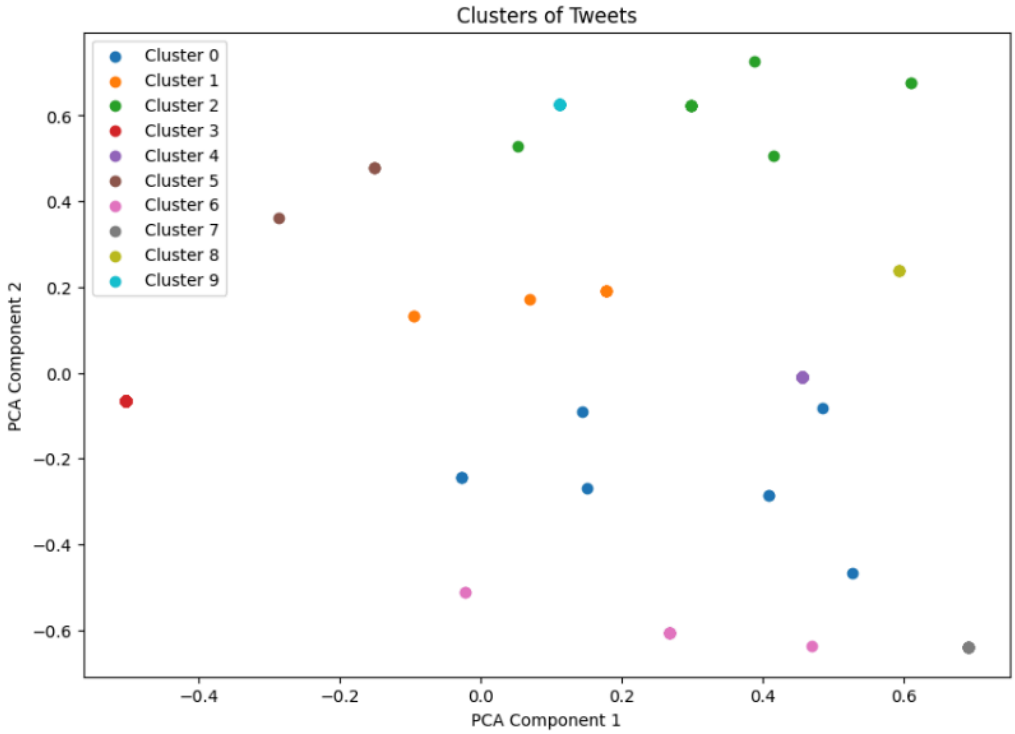


Figure 134: Visualization of a Agglomerative clustering of tweets

Problem:

- Agglomerative clustering may struggle with scalability for large datasets, and the choice of linkage criterion and distance metric can affect results.
- While it may not explicitly handle noise, agglomerative clustering's hierarchical nature can help identify outliers or noise clusters.

Solution:

- Addressing scalability may involve data sampling or preprocessing. Experimenting with various linkage criteria and distance metrics can enhance clustering quality.
- The performance of agglomerative clustering can be sensitive to parameters like the number of clusters and the choice of linkage method.

Summary:

- Agglomerative clustering, a hierarchical technique, merges clusters iteratively, offering interpretability through dendrogram visualization. It aids in exploring hierarchical structures and cluster relationships within the data.
- Despite scalability challenges with large datasets, customization options for linkage criteria and distance metrics enhance flexibility.
- Agglomerative clustering unveils cluster hierarchy and can aid in identifying outliers, yet sensitivity to parameters like the number of clusters warrants careful experimentation.

3.11 Comparison of Clustering algorithms

Clustering is a vital task in unsupervised learning, dividing datasets into groups where data points within the same group share similarities. We'll compare various clustering algorithms based on their flexibility, scalability, interpretability, and performance.

Hierarchical Clustering:
Flexibility: Doesn't need predefined clusters; handles various shapes and sizes.
Scalability: Computationally intensive for large datasets but works well for moderate-sized ones.
Interpretability: Produces dendrograms, aiding in understanding the data's hierarchical structure.
Performance: Quality depends on the chosen linkage criterion and distance metric; typically yields meaningful clusters.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise):
- Flexibility: Robust to noise and outliers; discovers clusters of arbitrary shapes.
- Scalability: Struggles with high-dimensional data and becomes expensive for dense datasets.
- Interpretability: Clusters based on density connectivity; less intuitive interpretation.
- Performance: Effective in datasets with non-uniform density; produces high-quality clusters.

K-means & TF-IDF Vectorization:
- Flexibility: Suitable for text data; captures semantic similarity.
- Scalability: Scales well for large datasets but may struggle with high-dimensional sparse data.
- Interpretability: Clusters based on centroid distances, especially interpretable with TF-IDF.
- Performance: Performance varies; sensitive to centroid initialization and 'k' choice.

Agglomerative Clustering:
- Flexibility: Handles various data types and cluster shapes.
- Scalability: Computationally expensive for large datasets due to hierarchical nature.
- Interpretability: Produces dendrograms illustrating hierarchical cluster relationships.
- Performance: Effective for exploring hierarchical structures but varies based on linkage and distance metrics.

Non-Negative Matrix Factorization (NMF):
- Flexibility: Applies to various data types, especially text; uncovers latent topics.
- Scalability: May struggle with high-dimensional sparse data but can scale with optimization.
- Interpretability: Reveals parts-based representations, aiding in understanding latent topics.
- Performance: Yields high-quality clusters for text data but requires specifying the number of components.

Non-Negative Latent Dirichlet Allocation (LDA):
- Flexibility: Commonly used for topic modeling in text data; applicable to clustering.
- Scalability: May become computationally expensive for large datasets, especially with many topics.
- Interpretability: Represents documents as mixtures of topics, facilitating understanding.
- Performance: Depends on hyperparameters; meaningful clusters for text data.

Below is a comparison chart comparing different types of clustering algorithms:

Algorithm	Flexibility	Scalability	Interpretability	Performance
Hierarchical Clustering	Handles various shapes and sizes; no predefined clusters	Computationally intensive for large datasets	Produces dendrograms illustrating hierarchical relationships; allows for easy visualization of cluster hierarchy	Quality depends on chosen linkage criterion and distance metric; may struggle with large datasets
DBSCAN	Robust to noise and outliers; discovers clusters of arbitrary shapes	Struggles with high-dimensional data; becomes expensive for dense datasets	Clusters based on density connectivity; less intuitive interpretation; noise points marked as outliers	Effective in datasets with non-uniform density; produces high-quality clusters; less impacted by outliers
K-means & TF-IDF Vectorization	Suitable for text data; captures semantic similarity	Scales well for large datasets but may struggle with high-dimensional sparse data	Clusters based on centroid distances; interpretable, especially with TF-IDF; clear geometric interpretation	Performance varies; sensitive to centroid initialization and 'k' choice; may produce suboptimal clusters for non-convex shapes
Agglomerative Clustering	Handles various data types and cluster shapes	Computationally expensive for large datasets due to hierarchical nature	Produces dendrograms illustrating hierarchical cluster relationships; allows for easy interpretation of cluster hierarchy	Effective for exploring hierarchical structures but varies based on linkage and distance metrics; sensitive to dataset size
Non-Negative Matrix Factorization (NMF)	Applies to various data types, especially text; uncovers latent topics	May struggle with high-dimensional sparse data but can scale with optimization	Reveals parts-based representations, aiding in understanding latent topics; interpretable latent features	Yields high-quality clusters for text data but requires specifying the number of components; sensitive to initialization
Non-Negative Latent Dirichlet Allocation (LDA)	Commonly used for topic modeling in text data; applicable to clustering	May become computationally expensive for large datasets, especially with many topics	Represents documents as mixtures of topics, facilitating understanding; interpretable topic distributions	Depends on hyperparameters; meaningful clusters for text data; sensitive to number of topics and choice of hyperparameters

Table 20: Chart comparing different types of clustering algorithms

3.12 Classification using Classification algorithm

Distinguishing between clustering and classification is vital in tweet segmentation.
Clustering groups tweets based on similarities without predefined labels, while classification assigns tweets to predefined segments using labeled data.

Clustering:

- Unsupervised Learning: Clustering groups of similar data points without predefined labels or classes.
- Doesn't Require Labeled Data: It works with unlabeled data to find patterns and structures.
- Unknown Number of Clusters: The number of clusters is often determined by the algorithm or user, not predefined.
- Maximizes Intra-cluster Similarity: It aims to maximize similarity within clusters and minimize similarity between clusters.

Classification:

- Supervised Learning: Classification assigns predefined labels or classes to data based on features.
- Requires Labeled Data: It needs labeled data for training, where each data point has a known class.
- Predictive Mapping: The goal is to predict labels for unseen data by learning patterns from labeled data.
- Minimizes Prediction Errors: Classification learns decision boundaries to separate different classes in feature space.

Aspect	Clustering	Classification
Learning Approach	Unsupervised learning method	Supervised learning method
Data Labeling	No need for labeled data; groups data based on similarity	Requires labeled data to train the model
Assumptions	Assumes that similar data points belong to the same group	Assumes that labeled data accurately represents classes
Performance Metrics	Evaluated based on metrics like silhouette score, Dunn index, or Davies-Bouldin index	Evaluated based on metrics like accuracy, precision, recall, and F1 score
Applications	Commonly used in customer segmentation, anomaly detection, and data exploration	Widely applied in spam detection, sentiment analysis, and medical diagnosis

Table 21: Difference between clustering algorithms and classification algorithm

Advantages of Classification for Segmenting Tweets:

Supervised Learning:
- Classification leverages labeled data, enabling the model to learn from known segments and generalize to unseen tweets. This supervised approach enhances the accuracy of tweet segmentation by leveraging existing knowledge about tweet categories or topics.

Predictive Accuracy:
- By learning patterns and relationships from labeled data, classification models can make reliable predictions about the segment or category of unseen tweets. The accuracy of these predictions contributes to the effectiveness of tweet segmentation, ensuring that tweets are correctly assigned to relevant segments based on their content or characteristics.

Interpretability:
- Classification models offer insights into the features that drive tweet segmentation, allowing stakeholders to understand the factors influencing the assignment of tweets to specific segments. By analyzing feature importance or model coefficients, users can gain valuable insights into the underlying reasons for segment assignments, such as prevalent keywords, sentiment trends, or user engagement metrics.

Scalability:
- Once trained, classification models can efficiently process large volumes of tweets, making them suitable for real-time or batch-processing applications. With scalable algorithms and parallel processing capabilities, classification models can handle the high velocity and volume of tweets generated on social media platforms, ensuring timely and accurate segmentation results.

Flexibility:
- Classification techniques are adaptable to diverse tweet features, including textual content, metadata attributes, and derived features such as sentiment scores or topic distributions. This flexibility enables the model to capture various aspects of tweet content and context, improving the accuracy and granularity of tweet segmentation. Additionally, the ability to incorporate different types of features allows for customization based on specific segmentation objectives or domain requirements.

Contextual Understanding:
- Classification models can incorporate contextual information from tweets, such as temporal trends, user profiles, or external events, to enhance segmentation accuracy. By considering the broader context in which tweets are generated, classification models can better capture the nuances and complexities of tweet content, leading to more contextually relevant segmentation results.

Adaptability to Evolving Trends:
- Classification models can adapt to changing trends and patterns in tweet data by continuously learning from new labeled examples. This adaptability allows the segmentation model to evolve, ensuring that it remains effective in capturing emerging topics, sentiments, or user behaviors in tweet content.

Integration with Decision-Making:
- Classification results can be seamlessly integrated into decision-making processes, enabling stakeholders to take informed actions based on segmented tweet data. Whether used for targeted marketing campaigns, customer service responses, or trend analysis, classification-based tweet segmentation provides valuable insights to support strategic decision-making initiatives.

3.12.1 Classification using Random Forest

Why Random Forest was Used:

- Handling Complexity: Random Forest is chosen for its ability to manage high-dimensional data and intricate relationships between features commonly found in tweet classification tasks.
- Overfitting Mitigation: Its ensemble nature helps combat overfitting by combining multiple decision trees trained on random subsets of data, ensuring robust performance on new data.

How it Helped to Solve the Issue:

- Effective Segmentation: Random Forest accurately segments tweets based on their content features, leveraging ensemble learning to capture diverse patterns in the data.
- Predictive Performance: The ensemble strategy enhances predictive performance by leveraging the collective insights of individual decision trees.
- Stability and Reliability: The bagging technique ensures stable predictions, even with noisy or unbalanced data distributions.

```
Accuracy: 0.9
Precision: 0.9261904761904761
Recall: 0.9
F1 Score: 0.8983333333333333

Classification Report:
              precision    recall  f1-score   support

      Anjuna       0.75      1.00      0.86         3
     Arambol       1.00      1.00      1.00         4
        Baga       1.00      0.75      0.86         4
    Benaulim       0.75      1.00      0.86         3
  Betalbatim       0.86      0.86      0.86         7
   Calangute       0.50      1.00      0.67         1
    Candolim       1.00      0.50      0.67         4
 Dirty beach       0.71      1.00      0.83         5
     Majorda       1.00      1.00      1.00         1
     Mandrem       1.00      1.00      1.00         3
      Morjim       1.00      1.00      1.00         7
     Palolem       1.00      0.75      0.86         4
   Sinquerim       1.00      1.00      1.00         1
      Utorda       1.00      1.00      1.00         2
     Vagator       1.00      1.00      1.00         8
       Varca       1.00      0.67      0.80         3

    accuracy                           0.90        60
   macro avg       0.91      0.91      0.89        60
weighted avg       0.93      0.90      0.90        60
```

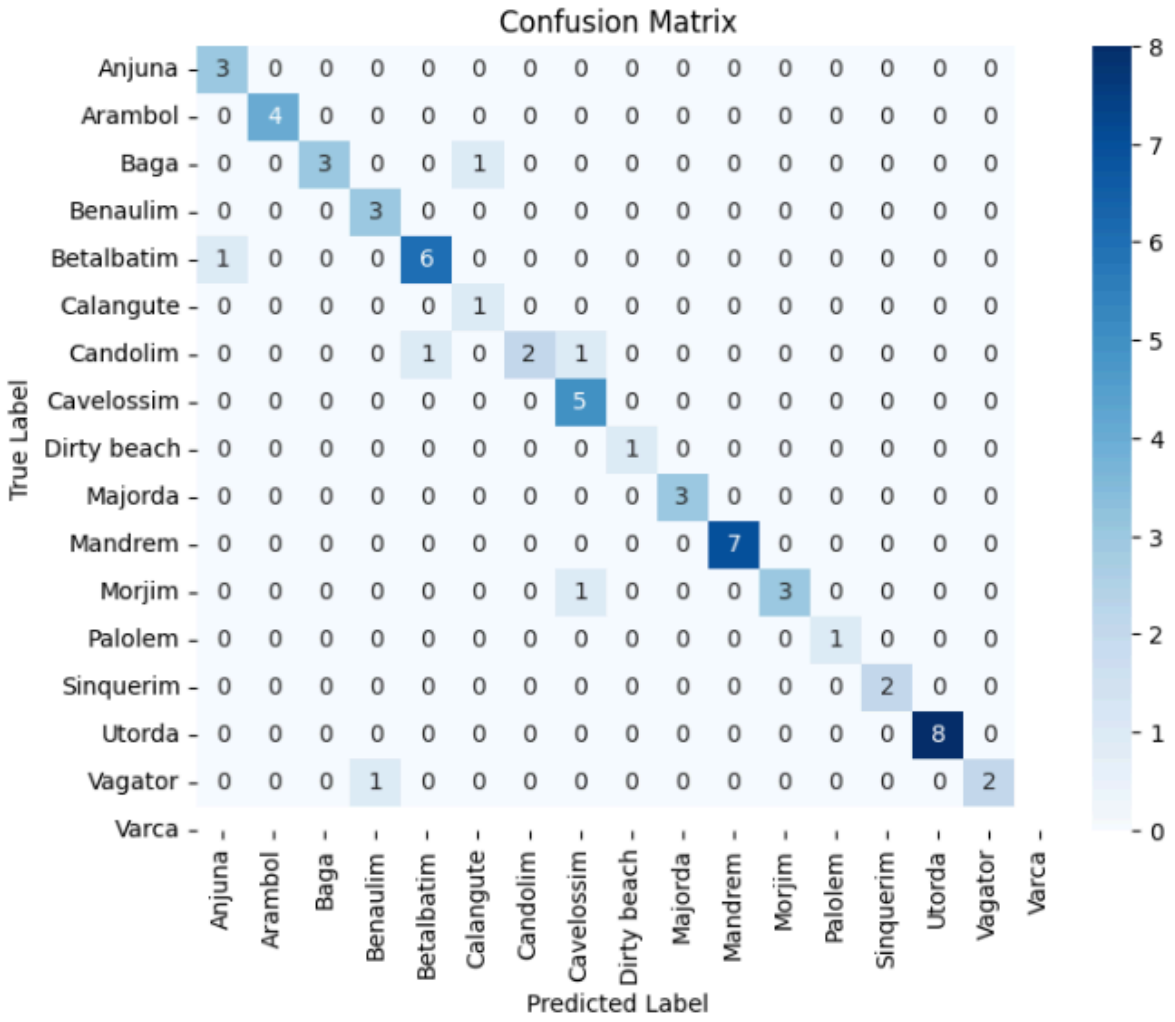


Figure 124: Confusion matrix showing the performance of of classification using Random Forest

Problem

- Imbalanced Data and Noise: Challenges may arise with highly imbalanced datasets or noisy features, potentially leading to biased predictions or reduced performance, especially in minority classes.
- Scalability Concerns: Scalability issues may arise with extremely large datasets or real-time applications, impacting model training and inference speed.

To Solve the Problem:

- Addressing Imbalanced Data: Explore alternative classification algorithms with better handling of class imbalances, such as ensemble methods with class weighting or resampling techniques.
- Improving Scalability: Implement optimizations like parallelization or distributed computing frameworks to accelerate model training and inference processes, ensuring scalability for large datasets or real-time applications.

3.12.2 Classification using Logistic Regression

Why Logistic Regression was Used:

- Simplicity and Interpretability: Logistic Regression's straightforward nature and interpretable coefficients make it ideal for binary classification tasks like tweet segmentation.
- Efficiency: Its efficient optimization algorithms enable rapid training and prediction, crucial for processing large tweet volumes in real-time or batch scenarios.
- Probabilistic Output: Logistic Regression provides class probabilities, offering insights into the likelihood of tweets belonging to specific segments based on their features.

How it Helped to Solve the Issue:

- Interpretability: The interpretable coefficients of Logistic Regression aid in understanding the features driving tweet classification, enhancing transparency and trust in the model.
- Efficient Processing: Logistic Regression's efficient algorithms enable swift training and prediction, ensuring timely segmentation of tweets even in high-volume scenarios.

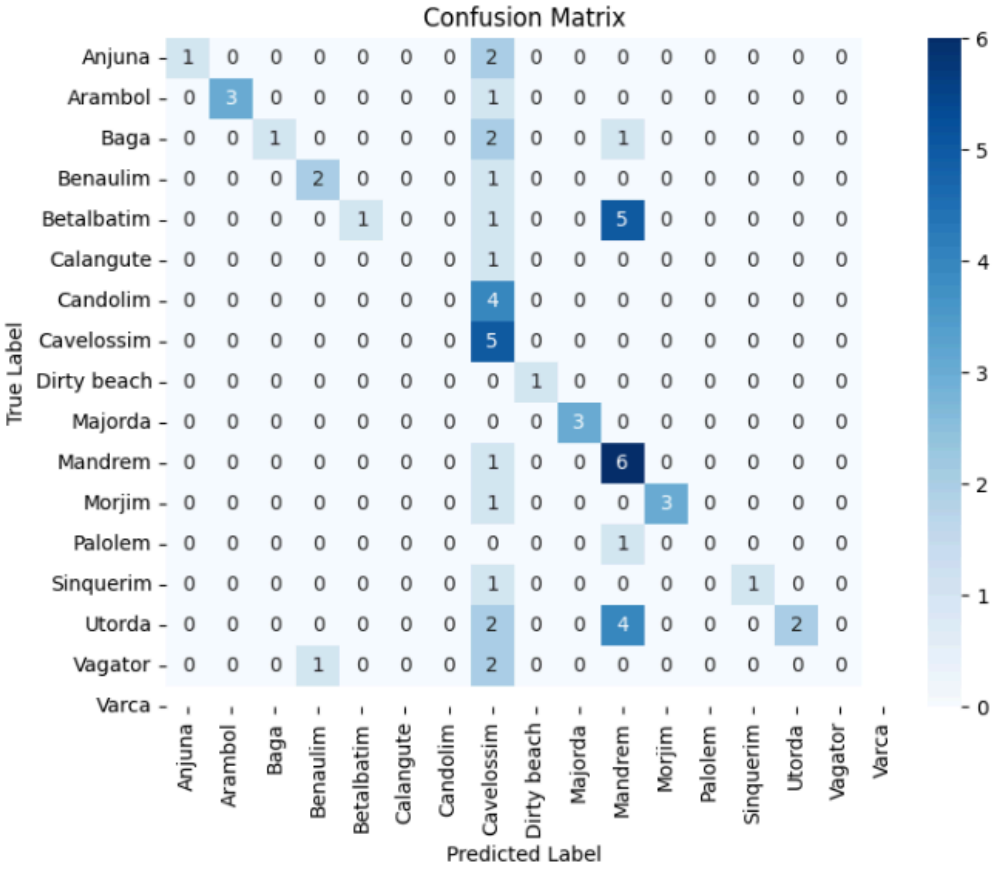


Figure 125: Confusion matrix showing the performance of of classification using Logistic Regression

```
Evaluation Metrics:
Accuracy: 0.48333333333333334
Precision: 0.8418709150326797
Recall: 0.48333333333333334
F1 Score: 0.4577435686918445

Classification Report:
              precision    recall  f1-score   support

      Anjuna       1.00      0.33      0.50         3
     Arambol       1.00      0.75      0.86         4
        Baga       1.00      0.25      0.40         4
    Benaulim       0.67      0.67      0.67         3
  Betalbatim       1.00      0.14      0.25         7
   Calangute       1.00      0.00      0.00         1
    Candolim       1.00      0.00      0.00         4
 Dirty beach       0.21      1.00      0.34         5
     Majorda       1.00      1.00      1.00         1
     Mandrem       1.00      1.00      1.00         3
      Morjim       0.35      0.86      0.50         7
     Palolem       1.00      0.75      0.86         4
   Sinquerim       1.00      0.00      0.00         1
      Utorda       1.00      0.50      0.67         2
     Vagator       1.00      0.25      0.40         8
       Varca       1.00      0.00      0.00         3

    accuracy                           0.48        60
   macro avg       0.89      0.47      0.47        60
weighted avg       0.84      0.48      0.46        60
```

Problem After Using Logistic Regression:

- Limitation in Capturing Complexity: The linear decision boundary of Logistic Regression may struggle to capture intricate feature interactions and non-linear patterns in the data, potentially leading to suboptimal performance.

To Solve the Problem:

- Explore Complex Models: Consider alternative classification algorithms capable of modeling non-linear relationships, such as kernel methods or deep learning models, to capture complex tweet segmentation patterns effectively.
- Feature Engineering: Apply feature engineering techniques or transformations to enrich the feature space, enabling Logistic Regression to better capture non-linear dependencies between tweets and their segments.

3.12.3 Classification using Naive Bayes

Why Naive Bayes was Used:

- Simplicity and Efficiency: Naive Bayes is chosen for its simplicity, efficiency, and effectiveness, especially in text classification tasks like segmenting tweets based on word frequencies.
- Probabilistic Approach: It leverages Bayes' theorem to estimate the probability of each segment given the observed tweet features, making it suitable for probabilistic classification tasks.

How it Helped to Solve the Issue:

- Probabilistic Segmentation: Naive Bayes effectively addresses tweet segmentation by modeling the conditional probability of each segment given tweet features, enabling it to assign tweets to the most probable segment based on observed word frequencies.
- Efficient Estimation: Its independence assumption simplifies estimation and reduces computational complexity, making it efficient for processing large tweet volumes in real-time or batch scenarios.

```
Evaluation Metrics:
Accuracy: 0.9666666666666667
Precision: 0.9722131747165804
Recall: 0.9666666666666667
F1 Score: 0.9670432682950646
Classification Report:
              precision    recall  f1-score   support

      Anjuna       1.00      1.00      1.00        12
     Arambol       1.00      0.95      0.98        21
        Baga       0.71      1.00      0.83        12
    Benaulim       0.96      1.00      0.98        24
  Betalbatim       1.00      0.95      0.97        20
   Calangute       1.00      0.76      0.87        17
    Candolim       1.00      1.00      1.00         9
   Cavelossim       1.00      1.00      1.00         9
  Dirty beach       0.98      1.00      0.99        41
     Majorda       0.92      1.00      0.96        12
     Mandrem       0.95      1.00      0.97        18
      Morjim       1.00      0.97      0.99        37
     Palolem       1.00      1.00      1.00        16
   Sinquerim       1.00      0.83      0.91         6
      Utorda       1.00      1.00      1.00        13
     Vagator       0.95      0.95      0.95        22
       Varca       1.00      0.91      0.95        11

    accuracy                           0.97       300
   macro avg       0.97      0.96      0.96       300
weighted avg       0.97      0.97      0.97       300
```

ProblemC

- Feature Independence Assumption: Naive Bayes assumes feature independence, which may not hold for all tweet features, especially complex relationships or dependencies between words or phrases, leading to suboptimal performance.

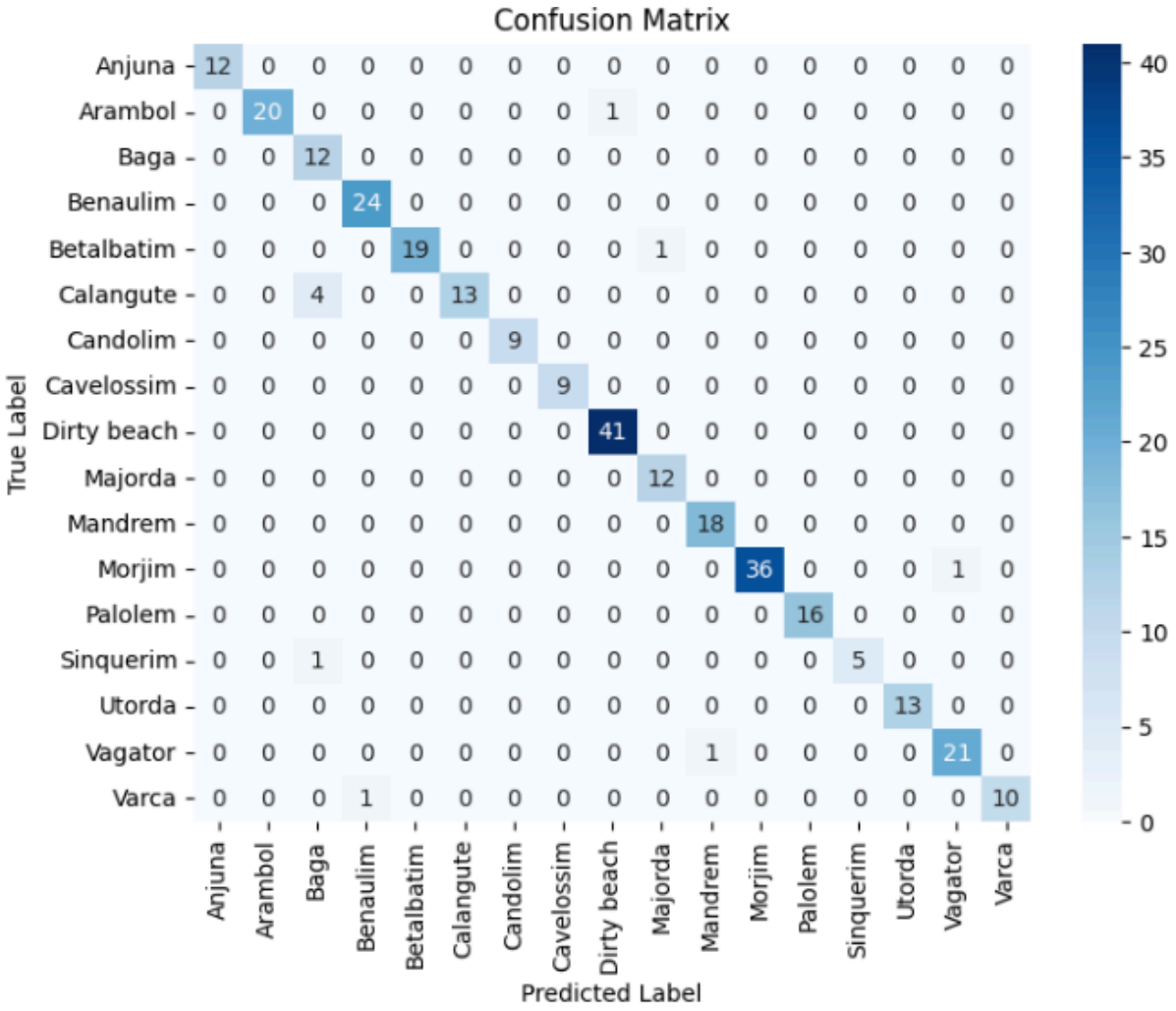


Figure 126: Confusion matrix showing the performance of of classification using Naive Bayes Classification

To Solve the Problem:

- Alternative Probabilistic Models: Explore models like Conditional Random Fields (CRFs) or Hidden Markov Models (HMMs) that allow for more flexible modeling of feature dependencies, addressing the limitation of feature independence.
- Ensemble Techniques: Apply ensemble techniques like bagging or boosting to combine multiple Naive Bayes classifiers trained on different feature subsets or data partitions, enhancing classification performance and robustness.

3.12.4 Classification using Support Vector Machines (SVM)

Why Support Vector Machines (SVM) were Used:

- Optimal Hyperplanes: SVMs excel in finding optimal hyperplanes that effectively separate data into classes, making them ideal for tweet segmentation tasks where clear decision boundaries are crucial.
- High-Dimensional Feature Spaces: SVMs work well with high-dimensional feature spaces and offer flexibility through various kernel functions, allowing for non-linear transformations of input data.

How it Helped to Solve the Issue:

- Accurate Segmentation: SVM effectively addresses tweet segmentation by identifying the best hyperplane to separate tweets into segments, leading to accurate classification based on feature representations.
- Handling Complexity: Its ability to handle complex data and high-dimensional feature spaces enables SVM to capture intricate relationships and patterns in tweet data, enhancing segmentation performance.

```
Accuracy: 0.3333333333333333
Classification Report:
              precision    recall  f1-score   support

      Anjuna       0.14      0.33      0.20         3
     Arambol       0.00      0.00      0.00         4
        Baga       0.00      0.00      0.00         4
    Benaulim       0.10      0.25      0.14         4
  Betalbatim       0.25      0.14      0.18         7
   Calangute       0.00      0.00      0.00         1
    Candolim       0.00      0.00      0.00         4
  Cavelossim       0.00      0.00      0.00         0
 Dirty beach       1.00      1.00      1.00         5
     Majorda       1.00      1.00      1.00         1
     Mandrem       1.00      0.67      0.80         3
      Morjim       1.00      1.00      1.00         7
     Palolem       0.00      0.00      0.00         4
   Sinquerim       0.00      0.00      0.00         1
      Utorda       0.33      1.00      0.50         2
     Vagator       0.00      0.00      0.00         8
       Varca       0.00      0.00      0.00         2

    accuracy                           0.33        60
   macro avg       0.28      0.32      0.28        60
weighted avg       0.32      0.33      0.31        60
```

Problem After Using SVM:

- Sensitivity to Parameters: SVM's sensitivity to kernel parameters and computational complexity may pose challenges, especially with large datasets or imbalanced class distributions, potentially leading to longer training times or suboptimal performance.

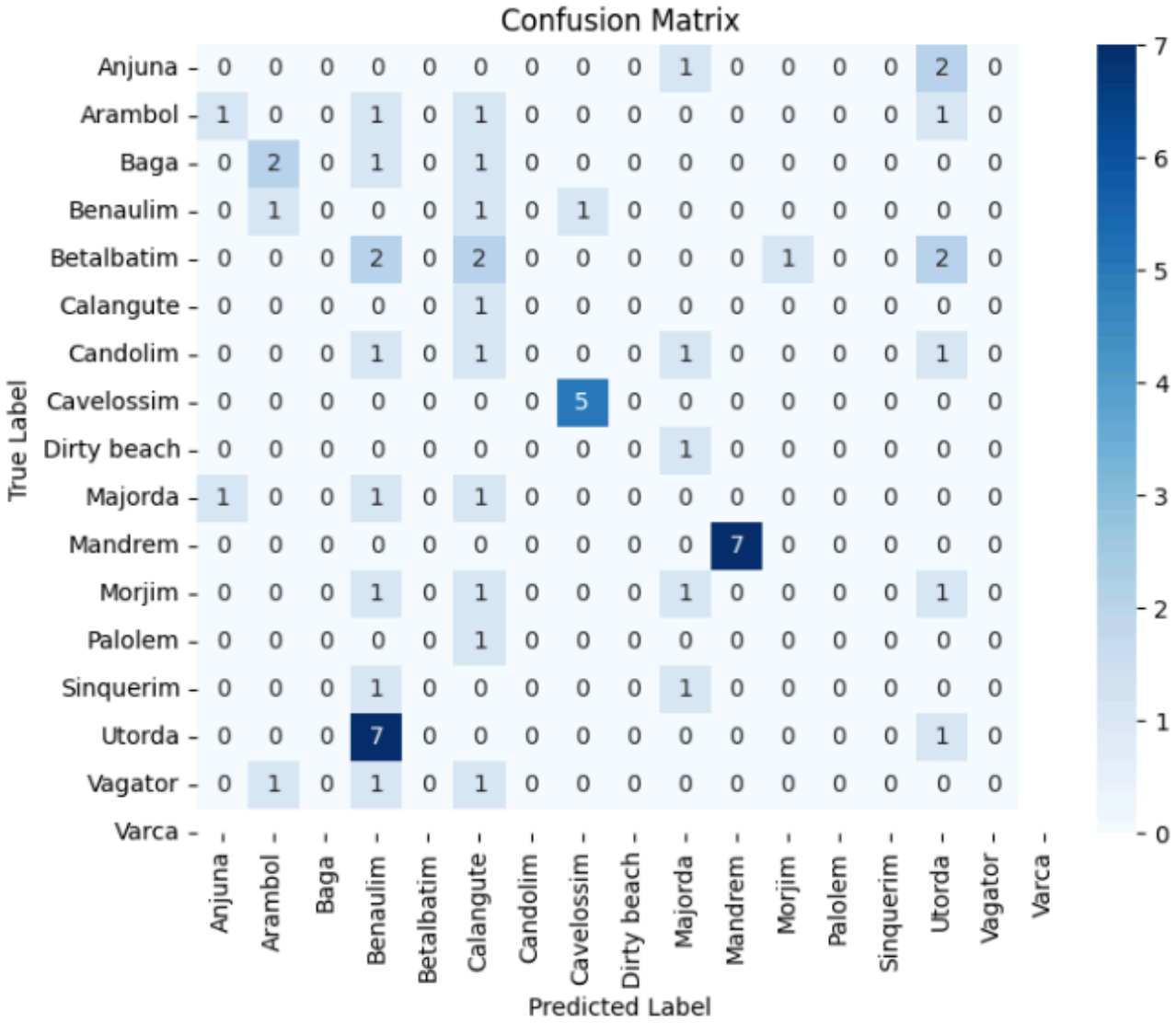


Figure 127: Confusion matrix showing the performance of of classification using SVM

To Solve the Problem:

- Parameter Optimization: Techniques like grid search or Bayesian optimization can be used to find optimal hyperparameters for each segment, enhancing SVM's robustness and performance.
- Handling Imbalances: Strategies such as class weighting or resampling techniques can address issues related to imbalanced data distributions, improving classification accuracy in the presence of skewed class distributions.

3.12.5 K-Nearest Neighbours (KNN)

Strengths:

- Instance-Based Learning: Utilizes instance-based learning, memorizing the training data and making predictions based on similarity to existing instances.
- Non-parametric: Makes no assumptions about the data distribution, suitable for scenarios with unknown or non-linear distributions.
- Flexibility: Can handle multi-class classification and accommodates both numerical and categorical features effectively.

Weaknesses:

- Computational Complexity: The computational cost of KNN increases with the size of the training set, as it requires calculating distances between the test instance and all training instances.
- Sensitive to Noise and Irrelevant Features: Performance may degrade in the presence of noisy or irrelevant features, necessitating careful feature selection or dimensionality reduction.

```
Accuracy: 0.6333333333333333
Precision: 0.7001851851851851
Recall: 0.6333333333333333
F1 Score: 0.6271783771783771
Classification Report:
```

	precision	recall	f1-score	support
Anjuna	0.60	1.00	0.75	3
Arambol	0.44	1.00	0.62	4
Baga	0.67	0.50	0.57	4
Benaulim	0.20	0.33	0.25	3
Betalbatim	0.67	0.57	0.62	7
Calangute	0.50	1.00	0.67	1
Candolim	0.75	0.75	0.75	4
Cavelossim	0.00	0.00	0.00	0
Dirty beach	0.75	0.60	0.67	5
Majorda	0.50	1.00	0.67	1
Mandrem	1.00	1.00	1.00	3
Morjim	0.50	0.29	0.36	7
Palolem	1.00	0.75	0.86	4
Sinquerim	0.00	0.00	0.00	1
Utorda	1.00	1.00	1.00	2
Vagator	1.00	0.38	0.55	8
Varca	0.75	1.00	0.86	3
accuracy			0.63	60
macro avg	0.61	0.66	0.60	60
weighted avg	0.70	0.63	0.63	60

	Dataset	Model	Accuracy	Pricision	Recall	F1 score
Libraries	D1	Re library	0.95	0.93	0.9	0.91
	D1	NLTK library	0.05	0.01	0.05	0.01
	D1	re, NLTK, Sklearn library	1	1	1	1
	D1	Spa-Cy	0.53	0.93	0.47	0.61
Using Classification Algorithms	D1	Random Forest	0.9	0.91	0.91	0.89
	D1	Logistic Regression	0.48	0.89	0.47	0.47
	D1	SVM	0.23	0.13	0.2	0.14
	D1	Naive Bayes	0.97	0.97	0.96	0.96
	D1	KNN	0.65	0.71	0.72	0.65
Libraries	D2	Re library	0.95	0.93	0.9	0.91
	D2	NLTK library	0.05	0.01	0.05	0.01
	D2	re, NLTK, Sklearn library	1	1	1	1
	D2	Spa-Cy	0.53	0.99	0.53	0.67
Using Classification Algorithms	D2	Random Forest	0.92	0.94	0.92	0.92
	D2	Logistic Regression	0.52	0.89	0.5	0.47
	D2	SVM	0.23	0.13	0.2	0.14
	D2	Naive Bayes	0.96	0.91	0.9	0.91
	D2	KNN	0.65	0.73	0.72	0.67
Libraries	D5	Re library	0.93	0.97	0.93	0.95
	D5	NLTK library	0.05	0.01	0.05	0.01
	D5	re, NLTK, Sklearn library	1	1	1	1
	D5	Spa-Cy	0.32	0.53	0.32	0.39
Using Classification Algorithms	D5	Random Forest	0.93	0.93	0.93	0.93
	D5	Logistic Regression	0.45	0.88	0.44	0.43
	D5	SVM	0.22	0.2	0.22	0.2
	D5	Naive Bayes	0.97	0.97	0.96	0.96
	D5	KNN	0.63	0.7	0.63	0.63

Table 22: Evaluation matrix for classification using libraries and using classification algorithm

3.13 Classification of tweets using Pre trained Models

Why pre-trained models

- Time Efficiency:
  - Training Time: Building a model from scratch can be time-consuming, especially for complex models and large datasets.
  - Pre-trained models, already trained on extensive datasets, save time.

- Resource Allocation:
  - Training models from scratch demands significant computational resources like GPUs or TPUs. Pre-trained models eliminate the need for resource-intensive training, saving time and computational costs.

- Performance:
  - High Accuracy: Pre-trained models capture intricate patterns and relationships in data, leading to high accuracy across tasks like image recognition and natural language processing.
  - Generalization: Trained on diverse data, pre-trained models generalize well to new data, ensuring consistent performance in various domains.

- Transfer Learning:
  - Fine-tuning: Pre-trained models can be fine-tuned for specific tasks with smaller datasets, leveraging existing knowledge for improved performance.
  - Domain Adaptation: Transfer learning enables pre-trained models to adapt to new domains, expanding their applicability across different scenarios.

- Accessibility:
  - Readily Available: Pre-trained models are easily accessible in popular machine learning libraries like TensorFlow and Py-Torch, simplifying integration into projects without extensive training expertise.

- Community Contributions:
  - Contributions from the ML community, shared through open-source repositories, further enhance accessibility to pre-trained models.

- Resource Conservation:
  - Data Requirements:
  - Training models from scratch necessitates large labeled datasets. Pre-trained models reduce the need for extensive data collection and annotation.

- Computational Resources:
  - Training complex models on large datasets consumes substantial computational resources.
  - Pre-trained models conserve resources, minimizing environmental impact and computational costs.

3.13.1.Transformer-based Models : BERT

What are Transformer-based Models:

- Transformer-based models are a class of deep learning models designed for sequential data processing, particularly suited for natural language processing (NLP) tasks.
- They utilize a self-attention mechanism to weigh the importance of different words in a sequence, enabling parallel processing of words and capturing long-range dependencies effectively.
- These models have revolutionized NLP by outperforming traditional sequence models like RNNs and LSTMs on various benchmarks, offering better scalability and efficiency.

What is BERT for deep learning?

BERT, short for Bidirectional Encoder Representations from Transformers, is a powerful natural language processing (NLP) model developed by Google that uses a deep neural network architecture based on the state-of-the-art transformer model.

As we said earlier, the BERT model architecture is based on a deep neural network called a transformer, which is different from traditional NLP models that process text one word at a time. Instead, transformers can process the entire text input all at once, which helps them to capture the relationships between words and phrases more effectively.

Architecture of BERT

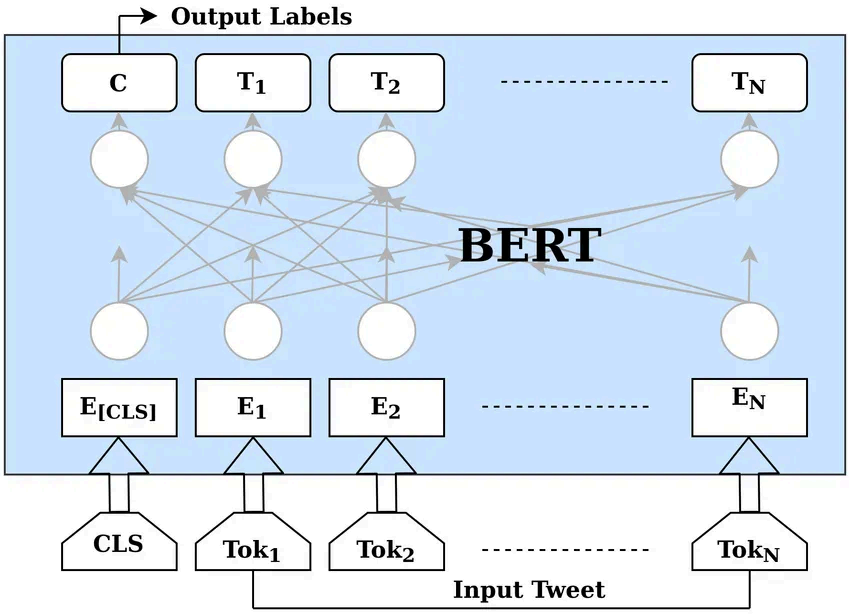


Figure 128: Confusion matrix showing the performance of of classification using SVM

BERT is a powerful NLP model built on transformer encoders. These encoders analyze relationships between words in a sentence. Unlike traditional left-to-right processing, BERT considers both left and right context (bidirectional) thanks to a self-attention mechanism. This allows BERT to understand the deeper meaning of words and perform various NLP tasks.

3.13.2 Selecting the Right Bert Model

BERT, a powerful Natural Language Processing (NLP) model, comes in two main flavors: Base and Large.

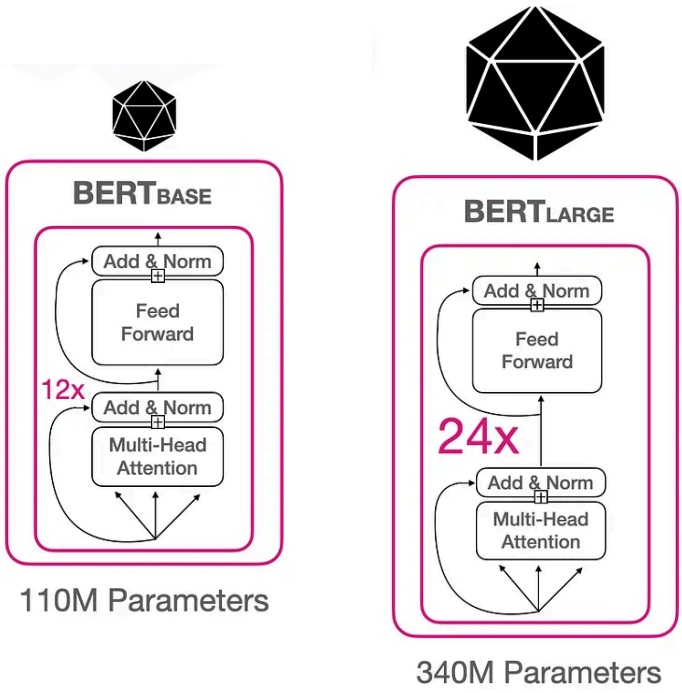


Figure 129: Difference between BERT Base and BERT Large

Feature	BERT Base	BERT Large
Difference	Fewer parameters and encoder layers	More parameters and encoder layers
Parameters	110 Million	340 Million
Encoder Layers	12	24
Strengths	Faster training and inference, good balance of accuracy and efficiency	Potentially higher accuracy for complex tasks
Weaknesses	May struggle with very complex tasks	Slower training and inference, requires more computational resources
Use Cases	- Question answering - Sentiment analysis - Text summarization - Machine translation (when fine-tuned)	- Same as BERT Base, but may perform better for - Named entity recognition - Text classification - Tasks requiring high accuracy

Table 23: Comparison chart of between Bert Base and Bert Large

3.13.3 Choosing BERT for Text Classification

- After thorough research and experimentation, we decided to utilize the BERT model for our text classification task due to its performance and versatility in handling diverse text data.
- One significant advantage of BERT is its pre-trained nature, which eliminates the need for manual feature engineering or keyword selection. This makes our code more robust and capable of analyzing tweets across various segments or keywords without the need for specific customization.
- Despite the availability of other models with potentially higher accuracy, we opted for BERT primarily for its ability to provide generalized classification without the requirement of tailored keyword lists.

3.13.4 Utilizing BERT in our Project

- Our project involves loading a pre-trained BERT model and tokenizer, specifically the 'bert-base-uncased' model, using the Hugging Face Transformers library.
- We preprocess tweets by checking for the presence of beach-related keywords using a predefined dictionary mapping beach names to associated keywords.
- The tweets are tokenized and encoded using the BERT tokenizer, with truncation or padding to ensure a consistent input length of 64 tokens.
- The BERT model is then loaded for sequence classification using BertForSequenceClassification, with the number of labels set to the number of beach categories.
- Due to the nature of the pre-trained model, which has not been fine-tuned on our specific dataset, the accuracy may not be optimal. However, it provides a solid starting point for our classification task.

3.13.5 Generating Output Files

- The classified tweets and their sentiments are stored in separate CSV files for each beach category in an output directory named '/content/Segmented Tweets'.
- Each CSV file contains the tweets belonging to a particular beach category along with their corresponding sentiments, enabling easy access and further analysis of the classified data.

Accuracy: 0.85

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.89	1.00	0.94	8
2	1.00	1.00	1.00	7
3	0.00	0.00	0.00	1
4	0.80	1.00	0.89	4
5	0.00	0.00	0.00	1
6	1.00	1.00	1.00	3
7	0.75	1.00	0.86	3
8	1.00	0.50	0.67	2
9	0.67	1.00	0.80	4
10	0.67	0.80	0.73	5
11	0.60	1.00	0.75	3
12	1.00	1.00	1.00	5
13	1.00	1.00	1.00	2
14	1.00	1.00	1.00	1
15	1.00	0.50	0.67	4
16	1.00	1.00	1.00	1
accuracy			0.85	60
macro avg	0.79	0.78	0.76	60
weighted avg	0.86	0.85	0.83	60

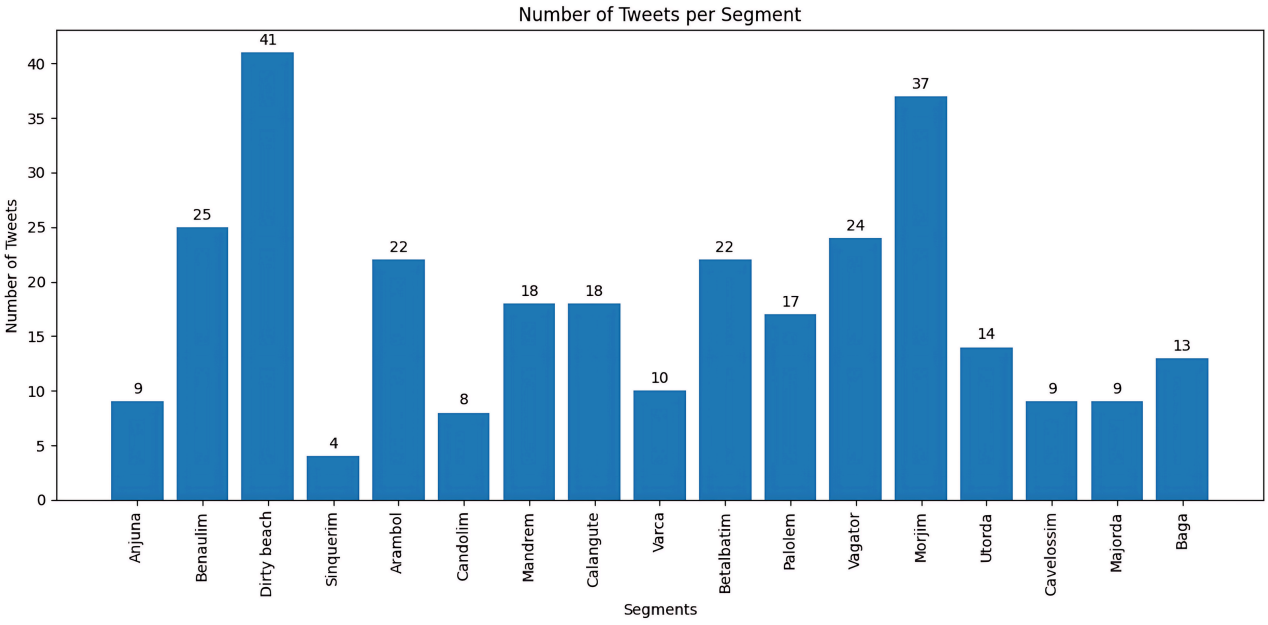


Figure 130: Tweet segmentation by Bert (without fine tuning) on Dataset D1

Accuracy: 0.87

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	0.33	0.33	0.33	3
2	1.00	1.00	1.00	3
3	1.00	0.67	0.80	3
4	1.00	1.00	1.00	4
5	0.00	0.00	0.00	1
6	1.00	1.00	1.00	2
8	0.73	1.00	0.84	8
9	0.80	1.00	0.89	4
10	0.88	1.00	0.93	7
11	1.00	1.00	1.00	1
12	1.00	0.33	0.50	3
13	0.80	0.80	0.80	5
14	1.00	1.00	1.00	4
15	1.00	0.86	0.92	7
16	1.00	1.00	1.00	4
accuracy			0.87	60
macro avg	0.85	0.81	0.81	60
weighted avg	0.87	0.87	0.85	60

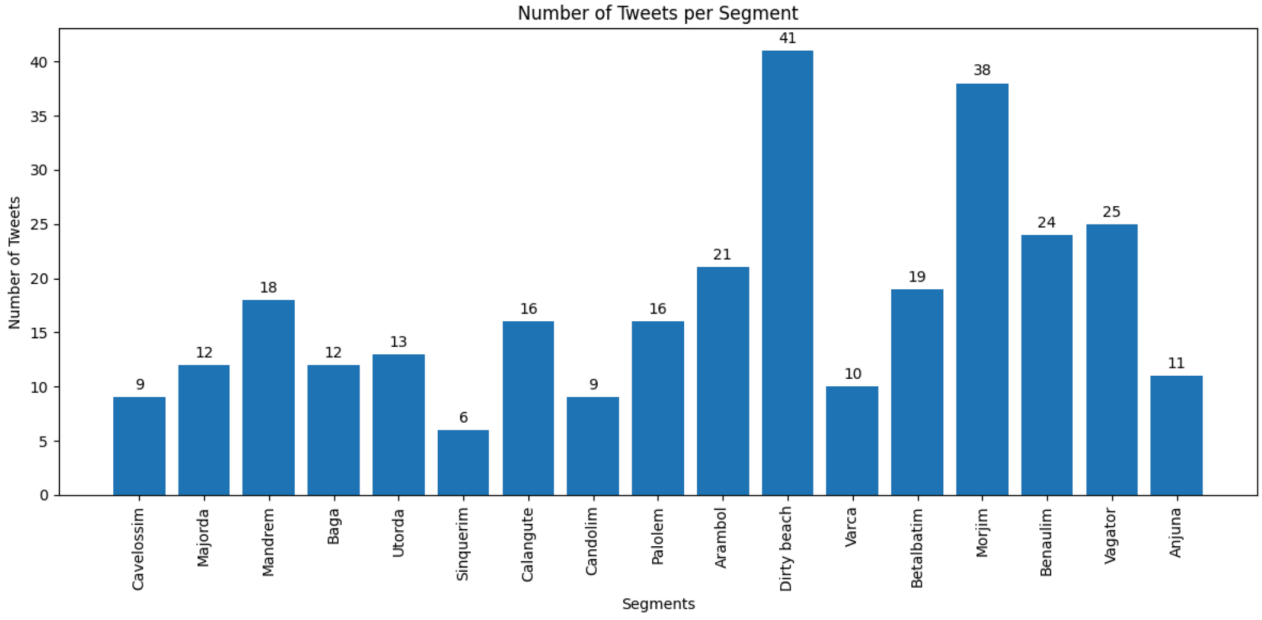


Figure 131: Tweet segmentation by Bert (without fine tuning) on Dataset D2

Accuracy: 0.90

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	0.75	1.00	0.86	3
2	1.00	1.00	1.00	3
3	1.00	1.00	1.00	3
4	1.00	1.00	1.00	4
5	0.50	1.00	0.67	1
6	1.00	1.00	1.00	2
8	1.00	0.88	0.93	8
9	0.80	1.00	0.89	4
10	1.00	1.00	1.00	7
11	0.00	0.00	0.00	1
12	1.00	0.67	0.80	3
13	0.83	1.00	0.91	5
14	1.00	0.75	0.86	4
15	0.78	1.00	0.88	7
16	1.00	0.50	0.67	4
accuracy			0.90	60
macro avg	0.85	0.86	0.84	60
weighted avg	0.91	0.90	0.89	60

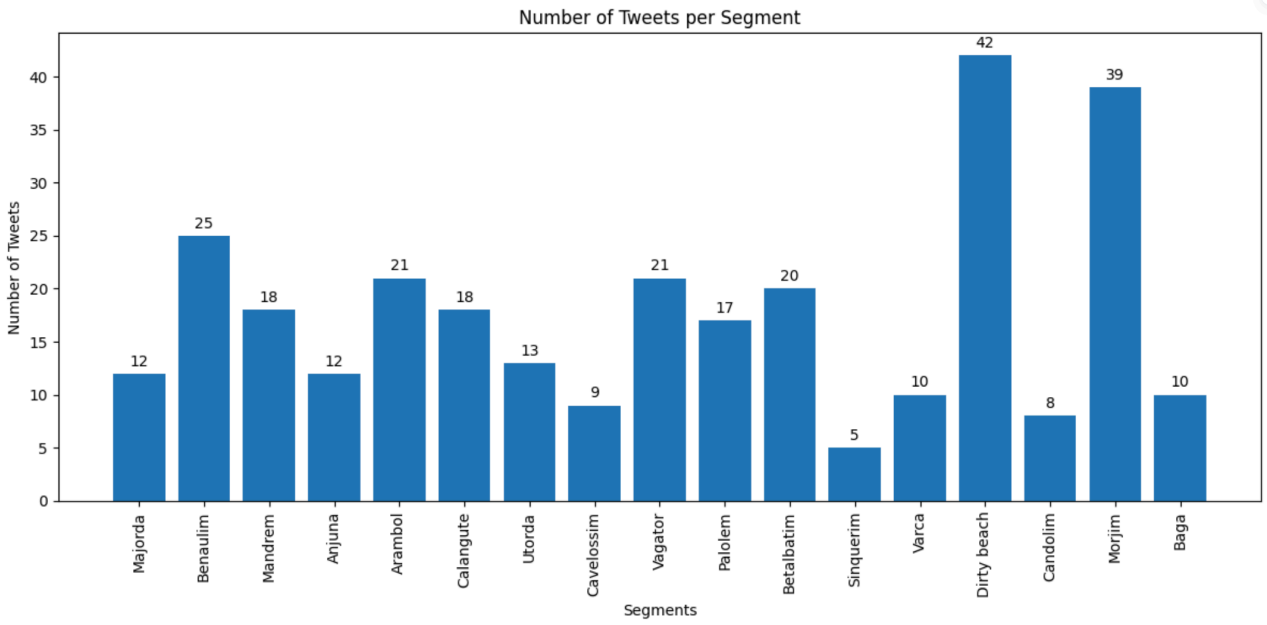


Figure 132: Tweet segmentation by Bert (without fine tuning) on Dataset D5

3.13.6 Fine Tuning BERT Model for Text Classification

To enhance accuracy, we fine-tuned the BERT model on our specific dataset, ensuring better performance tailored to our task.

- **Loading Pre-trained BERT Model and Tokenizer:**
  - The code loads the pre-trained BERT tokenizer from the 'Bert-base-uncased' model using the Bert Tokenizer class from the Hugging Face Transformers library.
  - Loading Tweets, Classes, and Sentiments from CSV File: Tweets, classes (categories), and sentiments are read from a CSV file located at '/content/Segmented Tweets/beachMasterFinalCleaned3-NoClean.csv'.
  - They are stored in separate lists (tweets, classes, and sentiments) for further processing.

- **Tokenizing and Encoding Tweets:**
  - Tweets are tokenized and encoded using the BERT tokenizer.
  - The tweets are truncated or padded to a maximum length of 64 tokens, and input IDs and attention masks are created for each tweet.
  - Classes are converted to numerical labels using a dictionary class_to_label.

- **Splitting Data into Training and Testing Sets:**
  - Encoded tweets, labels, and attention masks are split into training and testing sets using scikit-learn's train_test_split function.

- **Loading Pre-trained BERT Model for Sequence Classification:**
  - The pre-trained BERT model for sequence classification is loaded using BertForSequenceClassification from the Hugging Face Transformers library.
  - The number of labels (classes) for the classification task is set to the number of unique classes.

- **Fine-Tuning the BERT Model:**
  - To fine-tune the model, we trained it for a specified number of epochs (95 in this case) using the training data.
  - We utilized the AdamW optimizer for training the model and moved the model, input tensors, and labels to the appropriate device (CPU or GPU).

- **Evaluating the Trained Model:**
  - The trained model is evaluated on the test set to assess its performance.
  - We calculate the accuracy score and generate a classification report (precision, recall, F1-score) using scikit-learn's metrics.

- **Classifying Remaining Tweets:**
  - The code classifies the remaining tweets (not used for training or testing) using the fine-tuned BERT model.
  - Classified tweets and their corresponding sentiments are stored in separate lists for each class (category).

- **Saving Classified Tweets to CSV Files:**
  - An output directory '/content/Segmented Tweets-1' is created if it doesn't exist.
  - For each class (category), the classified tweets and their sentiments are saved to a separate CSV file in the output directory.

- **Plotting Bar Graph:**
  - A bar graph is created using matplotlib to visualize the number of tweets per segment (class/category).
  - The x-axis represents the segments (classes/categories), and the y-axis represents the number of tweets, with tweet counts displayed on top of each bar.

1st iteration of fine tuning

Accuracy: 0.90

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	0.75	1.00	0.86	3
2	1.00	1.00	1.00	3
3	1.00	1.00	1.00	3
4	1.00	1.00	1.00	4
5	0.50	1.00	0.67	1
6	1.00	1.00	1.00	2
8	1.00	0.88	0.93	8
9	0.80	1.00	0.89	4
10	1.00	1.00	1.00	7
11	0.00	0.00	0.00	1
12	1.00	0.67	0.80	3
13	0.83	1.00	0.91	5
14	1.00	0.75	0.86	4
15	0.78	1.00	0.88	7
16	1.00	0.50	0.67	4
accuracy			0.90	60
macro avg	0.85	0.86	0.84	60
weighted avg	0.91	0.90	0.89	60

Accuracy: 0.95

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3
1	0.75	1.00	0.86	3
2	1.00	1.00	1.00	5
3	0.00	0.00	0.00	1
4	1.00	1.00	1.00	4
5	1.00	1.00	1.00	4
6	1.00	1.00	1.00	3
7	1.00	1.00	1.00	1
8	1.00	0.86	0.92	7
9	1.00	0.67	0.80	3
10	0.67	1.00	0.80	4
11	1.00	1.00	1.00	8
12	1.00	1.00	1.00	7
13	1.00	1.00	1.00	2
15	1.00	1.00	1.00	1
16	1.00	1.00	1.00	4
accuracy			0.95	60
macro avg	0.90	0.91	0.90	60
weighted avg	0.95	0.95	0.94	60

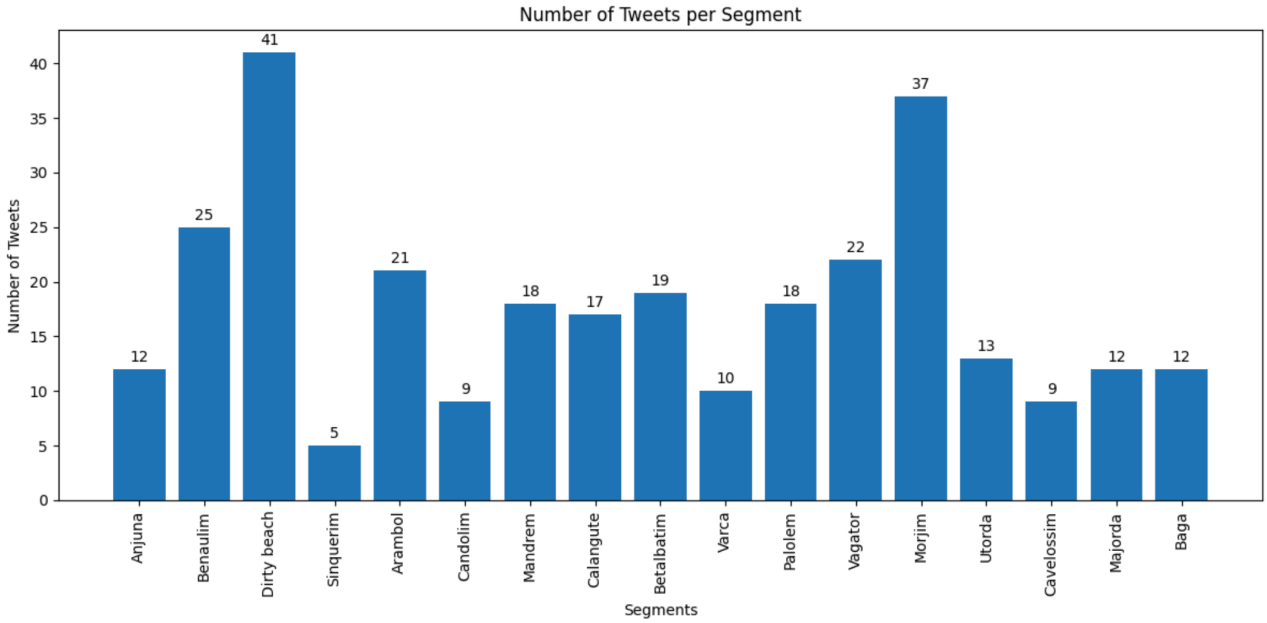


Figure 133: Tweet segmentation by Bert (after fine tuning) on D5

Segmented data, categorized by specific classes (e.g., beach locations), facilitates targeted sentiment analysis. This structured input enables sentiment analysis models to analyze sentiment trends within each category. By analyzing sentiment across categories, we gain insights into public perception and sentiment towards different topics or locations. This approach provides valuable insights for stakeholders to make informed decisions based on sentiment trends.

3.14 Adding Sentimental Analysis model

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine the sentiment or opinion expressed in a piece of text. It involves identifying and categorizing the sentiment conveyed in the text as positive, negative, or neutral. Sentiment analysis can be applied to various types of text data, such as product reviews, social media posts, and customer feedback, to gain insights into people's opinions, attitudes, and emotions.

Here are the four main types of sentiment analysis:

Intent-based Sentiment Analysis:
Intent-based sentiment analysis aims to understand the intent behind the text and classify it based on the expressed sentiment. It focuses on identifying whether the sentiment is positive, negative, or neutral, without considering the specific aspects or entities mentioned in the text.

- **Example Model**: VADER (Valence Aware Dictionary and sEntiment Reasoner)

Fine-grained Sentiment Analysis:
Fine-grained sentiment analysis goes beyond simple positive, negative, or neutral classification and aims to categorize sentiment into multiple, more nuanced categories. It involves identifying emotions or sentiments with greater granularity, such as very positive, slightly positive, very negative, slightly negative, and neutral.

- **Example Model**: BERT (Bidirectional Encoder Representations from Transformers)

Emotional Analysis:
Emotional analysis focuses on identifying specific emotions expressed in the text, such as happiness, sadness, anger, fear, or surprise. It aims to capture the underlying emotions of the text beyond just positive or negative sentiment.

- **Example Model**: EmoBERT

Aspect-based Sentiment Analysis:
Aspect-based sentiment analysis involves identifying specific aspects or entities mentioned in the text and analyzing the sentiment associated with each aspect independently. It aims to understand the sentiment expressed toward different aspects or features of a product, service, or topic.

- **Example Model**: Aspect-based sentiment analysis often utilizes deep learning models such as LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Unit) neural networks for aspect extraction and sentiment classification.

3.14.1 Why RoBERTa for Sentiment Analysis

- RoBERTa is a state-of-the-art language model pre-trained on a large corpus of text data, including Twitter data. It captures rich semantic information and contextual understanding, making it suitable for sentiment analysis tasks.
- RoBERTa exhibits superior performance on various NLP tasks, including sentiment analysis, due to its bidirectional architecture and extensive pre-training.
- Research and empirical evidence have demonstrated the effectiveness of RoBERTa in sentiment analysis tasks, particularly for social media text data like tweets.

3.14.2 How RoBERTa is Utilized in the Project

- RoBERTa is utilized in the project for sentiment analysis of segmented tweets. It processes each tweet through tokenization, encoding, and sentiment prediction using the RoBERTa model.
- The softmax scores generated by RoBERTa are used to classify each tweet into one of the predefined sentiment categories (negative, neutral, or positive).
- Evaluation metrics such as accuracy are calculated to assess the performance of the sentiment analysis model.
- The sentiment analysis results are stored in CSV files for further analysis and visualization.

3.14.3 Implementation of Roberta for sentimental analysis

Loading the Model and Tokenizer:

- We load the RoBERTa model and tokenizer using the AutoModelForSequenceClassification and AutoTokenizer classes from the Transformers library. We specify the model as "cardiffnlp/twitter-roberta-base-sentiment".

Preprocessing Tweets:

- We define a function preprocess_tweet(tweet) to preprocess each tweet before sentiment analysis. The function replaces user mentions and URLs with generic tokens.

Reading CSV Files:

- We define a function get_file_paths(directory) to get a list of CSV file paths in a specified directory.
- We read each CSV file into a DataFrame containing segmented tweets and their corresponding sentiments.

Sentiment Analysis:

- For each tweet in the DataFrame, we preprocess the tweet using preprocess_tweet().
- We tokenize and encode the preprocessed tweet using the RoBERTa tokenizer.
- The encoded tweet is passed through the RoBERTa model, and softmax scores are calculated.
- We predict the sentiment label (negative, neutral, or positive) based on the highest softmax score.
- We store the original sentiment label, predicted sentiment label, and tweet text.

Calculating Evaluation Metrics:

- We calculate the accuracy of the sentiment predictions by comparing the original and predicted labels.

Writing Output CSV Files:

- We create an output directory named "Sentimented_tweets" if it doesn't exist.
- For each segment (CSV file) in the input directory, we generate an output CSV file containing sentiment analysis results.
- The output CSV file includes the segment name, sentiment distribution, evaluation metrics, and tweet data.

Output and Analysis:

- After processing each segment, we print the segment name, sentiment distribution, and accuracy and store segment wise sentimental analysis output files.

3.14.4 Output of Roberta (Sentimental Analysis) with D1

Sr.No	Segment	Positive	Negative	Neutral	Accuracy
1	Baga	0.50	0.08	0.42	0.74
2	Arambol	0.19	0.62	0.19	0.81
3	Candolim	0.75	0.12	0.12	0.88
4	Anjuna	0.00	0.33	0.67	0.58
5	Majorda	0.25	0.25	0.50	0.83
6	Benaulim	0.19	0.54	0.27	0.69
7	Cavelossim	0.67	0.22	0.11	0.78
8	Varca	0.30	0.30	0.40	0.60
9	Palolem	0.81	0.00	0.19	0.81
10	Utorda	0.23	0.38	0.38	0.69
11	Mandrem	0.06	0.33	0.61	0.49
12	Dirty Beach	0.05	0.66	0.29	0.73
13	Morjim	0.26	0.34	0.39	0.61
14	Calangute	0.29	0.29	0.41	0.59
15	Betalbatim	0.30	0.15	0.55	0.65
16	Sinquerim	0.20	0.00	0.80	0.40
17	Vagator	0.23	0.32	0.45	0.55

Table 24: Sentimental analysis by Roberta on D1 dataset

3.14.5 Output of Roberta (Sentimental Analysis) with D2

Sr.No	Segment	Positive	Negative	Neutral	Accuracy
1	Baga	0.50	0.10	0.40	0.80
2	Arambol	0.19	0.62	0.19	0.76
3	Candolim	0.75	0.12	0.12	0.88
4	Anjuna	0.00	0.33	0.67	0.58
5	Majorda	0.25	0.25	0.50	0.83
6	Benaulim	0.20	0.56	0.24	0.72
7	Cavelossim	0.67	0.22	0.11	0.78
8	Varca	0.30	0.30	0.40	0.60
9	Palolem	0.82	0.00	0.18	0.82
10	Utorda	0.23	0.31	0.46	0.62
11	Mandrem	0.06	0.33	0.61	0.39
12	Dirty Beach	0.05	0.69	0.26	0.76
13	Morjim	0.28	0.38	0.33	0.59
14	Calangute	0.28	0.28	0.44	0.56
15	Betalbatim	0.25	0.15	0.60	0.60
16	Sinquerim	0.20	0.00	0.80	0.40
17	Vagator	0.24	0.33	0.43	0.57

Table 25: Sentimental analysis by Roberta on D2 dataset

3.14.6 Output of Roberta (Sentimental Analysis) with D5

Sr.No	Segment	Positive	Negative	Neutral	Accuracy
1	Baga	0.25	0.08	0.67	0.25
2	Arambol	0.05	0.05	0.90	0.19
3	Candolim	0.33	0.11	0.56	0.44
4	Anjuna	0.00	0.00	1.00	0.27
5	Majorda	0.08	0.00	0.92	0.42
6	Benaulim	0.12	0.04	0.83	0.29
7	Cavelossim	0.33	0.00	0.67	0.33
8	Varca	0.00	0.00	1.00	0.00
9	Palolem	0.35	0.00	0.65	0.35
10	Utorda	0.08	0.08	0.85	0.23
11	Mandrem	0.06	0.11	0.83	0.17
12	Dirty Beach	0.00	0.46	0.54	0.56
13	Morjim	0.06	0.03	0.92	0.08
14	Calangute	0.06	0.00	0.94	0.12
15	Betalbatim	0.00	0.00	1.00	0.29
16	Sinquerim	0.00	0.00	1.00	0.17
17	Vagator	0.04	0.00	0.96	0.04

Table 26: Sentimental analysis by Roberta on D5 dataset

Sr.No	Datasets	Accuracy
1	D1	0.6723529412
2	D2	0.6623529412
3	D3	0.2470588235

Table 32: Comparative aggregate accuracies between datasets D1, D2 AND D5

3.14.7 Output of Fine Tuned Roberta (Sentimental Analysis) with D1

- We attempted to improve the accuracy of our sentiment analysis model by fine-tuning it, but unfortunately, our efforts didn't fully succeed.
- While we did see some improvement, as shown in the graph, the model still didn't give us the results we needed in the format we wanted.
- Consequently, we couldn't use it in our final project, so we had to look for other ways to analyze sentiment for our project.

Sr.No	Segment	Positive	Negative	Neutral	Accuracy
1	Baga	0.33	0.67	0.00	0.67
2	Arambol	0.20	0.60	0.20	0.80
3	Candolim	1.00	0.00	0.00	1.00
4	Anjuna	0.33	0.67	0.00	1.00
5	Majorda	0.33	0.00	0.67	1.00
6	Benaulim	0.33	0.67	0.00	1.00
7	Cavelossim	0.50	0.50	0.00	1.00
8	Varca	1.00	0.00	0.00	0.50
9	Palolem	1.00	0.00	0.00	1.00
10	Utorda	0.33	0.67	0.00	1.00
11	Mandrem	0.50	0.50	0.00	1.00
12	Dirty Beach	0.00	1.00	0.00	0.89
13	Morjim	0.50	0.50	0.00	1.00
14	Calangute	0.25	0.25	0.50	0.50
15	Betalbatim	0.50	0.50	0.00	0.75
16	Sinquerim	0.00	0.00	1.00	0.00
17	Vagator	0.40	0.60	0.00	1.00

Table 33: Fine Tuned Roberta Sentimental analysis output on D1 dataset

3.15 Adding Emotional Analysis model

Emotional analysis is a process of analyzing, processing, and summarizing subjective texts with emotional colors. It involves the detection and interpretation of the underlying emotions portrayed in textual data, going beyond the simple categorization of sentiment as positive, negative, or neutral. Emotional analysis is a more involved and deeper analysis of consumer emotions that aims to drill down into the psychology of different user behaviors.

Why DistilRoBERTa for Emotion Analysis:

- DistilRoBERTa is a distilled version of RoBERTa, optimized for faster inference while retaining performance.
- For emotion analysis, the DistilRoBERTa model was chosen due to its ability to capture nuanced emotional expressions in text data.
- DistilRoBERTa provides efficient and accurate predictions for emotion classification tasks, making it suitable for sentiment analysis in this context.

How DistilRoBERTa is Utilized in the Project:

- DistilRoBERTa is utilized in the project specifically for emotion analysis of segmented tweets.
- It processes each tweet through tokenization, encoding, and emotion prediction using the DistilRoBERTa model.
- The predicted emotion labels are stored and analyzed along with the sentiment labels obtained from the RoBERTa model.

Connection with Previous Sentiment Analysis:

- The sentiment analysis using RoBERTa provides insights into the overall sentiment distribution of tweets in each segment.
- The emotion analysis using DistilRoBERTa complements the sentiment analysis by providing additional information about the emotional content of the tweets.
- Together, sentiment and emotion analysis offer a comprehensive understanding of the sentiment and emotional expressions present in the segmented tweet data.

Output File Structure:

- Each output CSV file contains columns for tweet text, sentiment label, and emotion label.
- Additionally, the CSV file includes sections for overall sentiment distribution and overall emotion distribution within the segment.
- The output text file provides a concise summary of the sentiment and emotion analysis results for each segment.

3.15.1 Implementation of Distil-Roberta for emotional analysis

Loading Models and Tokenizers:

- We load the RoBERTa model and tokenizer for sentiment analysis and the DistilRoBERTa model for emotion analysis using the AutoModelForSequenceClassification and AutoTokenizer classes from the Transformers library.

Preprocessing Tweets:

- We define a function preprocess_tweet(tweet) to preprocess each tweet before analysis.

Reading CSV Files:

- We define an input directory containing segmented tweets in CSV format. We gather the file paths of all CSV files in the directory.

Sentiment and Emotion Analysis:

- For each CSV file, we load the tweets into a DataFrame.
- We iterate over each tweet, preprocess it, and perform sentiment analysis using the RoBERTa model.
- Additionally, we perform emotion analysis using the DistilRoBERTa model.
- We store the sentiment and emotion labels for each tweet along with the tweet text.

Output CSV and Text Files:

- For each segment (CSV file), we create an output CSV file containing sentiment and emotion analysis result
- The CSV file includes the segment name, overall sentiment distribution, overall emotion distribution, and tweet-level sentiment and emotion labels.
- Additionally, we create a text file summarizing the sentiment and emotion analysis results for each segment.

Output and Analysis:

- After processing each segment, we print the segment name, overall sentiment distribution, and overall emotion distribution in the console.
- We also print the paths to the output CSV and text files created for each segment.

3.15.2 Output Files created Using Distil-Roberta's emotional analysis

Name of the segment: Baga
Overall sentiments of the Segment:
Positive: 0.50
Negative: 0.42
Neutral: 0.08
Overall Emotional Analysis:
fear: 0.50%
joy: 0.25%
disgust: 0.08%
surprise: 0.08%
anger: 0.08%
Text file for Baga segment created: /content/Emotional_Analysis-DistilBert/Baga_segmented.txt

---

Name of the segment: Arambol
Overall sentiments of the Segment:
Negative: 0.62
Neutral: 0.19
Positive: 0.19
Overall Emotional Analysis:
fear: 0.57%
anger: 0.14%
sadness: 0.10%
neutral: 0.10%
surprise: 0.05%
joy: 0.05%
Text file for Arambol segment created: /content/Emotional_Analysis-DistilBert/Arambol_segmented.txt

---

Name of the segment: Candolim
Overall sentiments of the Segment:
Positive: 0.75
Negative: 0.12
Neutral: 0.12
Overall Emotional Analysis:
fear: 0.62%
joy: 0.38%
Text file for Candolim segment created: /content/Emotional_Analysis-DistilBert/Candolim_segmented.txt

---

Name of the segment: Varca
Overall sentiments of the Segment:
Neutral: 0.40
Negative: 0.30
Positive: 0.30
Overall Emotional Analysis:
anger: 0.30%
fear: 0.30%
sadness: 0.20%
neutral: 0.10%
joy: 0.10%
Text file for Varca segment created: /content/Emotional_Analysis-DistilBert/Varca_segmented.txt

3.16 Using Open AI’s Generative model

OpenAI's GPT models, like GPT-3.5, offer a powerful tool for analyzing tweets and generating suggestions, leveraging their advanced natural language understanding and generation capabilities. With OpenAI's technology, users can engage in several key tasks:

Sentiment Analysis:

- Utilizing GPT models enables users to conduct sentiment analysis on tweets effectively. By prompting the model with a tweet, users can request an analysis of the sentiment expressed within the tweet, categorizing it as positive, negative, or neutral.
- For instance, users may input a prompt such as, "Analyze the sentiment of this tweet: 'Tweet text goes here.' Is the sentiment positive, negative, or neutral?"

Contextual Understanding:

- GPT models boast an exceptional capacity for comprehending the context and intricacies of language. This capability allows the models to scrutinize tweets within a broader context, considering factors such as tone, linguistic patterns, and underlying emotions.
- As a result, the model can provide insightful analyses that extend beyond mere sentiment classification.

Generating Suggestions:

- Following sentiment analysis, users can prompt the model to generate suggestions based on the analyzed tweets. For positive tweets, the model can propose strategies to amplify positive sentiments or foster engagement, such as sharing uplifting stories or promoting positive initiatives. Conversely, for negative tweets, the model can suggest approaches to address concerns, alleviate negative sentiments, or initiate constructive actions, such as offering solutions or providing support.
- Moreover, the model can offer general recommendations or guidance for neutral tweets, taking into account the broader conversation context.

By harnessing OpenAI's GPT models for tweet analysis and suggestion generation, users can gain deeper insights into tweet sentiment and content, identify actionable opportunities, and inform decision-making processes across various domains, including social media management, brand reputation monitoring, crisis response, and public opinion analysis.

3.16.1.Benefits to the Project:

- Addition of Open Ai enhances sentiment analysis by providing more detailed analysis, reasoning, and suggestions using GPT-3.5.
- It complements the sentiment and emotion analysis performed earlier, providing deeper insights into the sentiment of tweets and possible actions to address negative sentiments or enhance positive sentiments.
- By utilizing GPT-3.5, the code leverages advanced natural language processing capabilities to generate nuanced analysis and suggestions.

Input and Processing:

- The code takes input from CSV files containing segmented tweets, where each row represents a tweet along with its sentiment label obtained from RoBERTa.
- It processes each tweet by passing it through the GPT model to generate sentiment analysis, reasoning, and suggestions.

Storage and Final Algorithm:

- The results of sentiment analysis, reasoning, and suggestions generated by GPT are stored in new columns in the DataFrame.
- The updated DataFrame is saved to a new CSV file in the output directory, containing the original tweet data along with the GPT analysis.
- The final algorithm involves looping through each CSV file, processing the tweets, performing sentiment analysis using GPT, and saving the results for further analysis and insights.

3.16.2.Implementation of OpenAI for sentimental analysis

Setting up OpenAI API:
- The code sets up the OpenAI API key to authenticate and access OpenAI's services.

Finding Header Row:
- A function find_header_row() is defined to find the row number where the column headers start in the CSV file.

Sentiment Analysis using OpenAI GPT:
- Another function get_gpt_analysis() is defined to perform sentiment analysis using OpenAI's GPT model. It constructs a prompt for GPT based on tweet text and RoBERTa sentiment label and sends it to GPT for analysis.
- The response from GPT is parsed to extract sentiment, reasoning, and suggestions.

Input and Output Directories:
- Input and output directories are defined to specify the location of input CSV files containing segmented tweets and to save the output files with GPT analysis.

Looping through CSV Files:
- The code loops through each CSV file in the input directory.

Reading and Processing CSV Data:
- For each CSV file, it reads the data into a DataFrame and adds new columns for 'GPT Sentiment', 'GPT Reasoning', and 'GPT Suggestions'.
- For each row in the DataFrame, it calls the get_gpt_analysis() function to perform sentiment analysis using GPT and stores the results in the corresponding columns.

Saving Output:
- After analyzing each row, it saves the updated DataFrame to a new CSV file in the output directory.

Additional Analysis Summary:
- It prints additional analysis summaries including the segment name, overall sentiments of the segment, overall emotional analysis, and GPT analysis and suggestions summary.

4. ANALYSIS AND CONCLUSION

4.1 Final Output of Open AI analysis and suggestion

Output File Creation Process:

- For each CSV file in the input directory containing segmented tweets, the code performs sentiment analysis using OpenAI's GPT model.
- It creates a new DataFrame where each row corresponds to a tweet and includes columns for the original tweet text, sentiment label, emotion label, GPT sentiment analysis, reasoning, and suggestions.
- After analyzing all tweets in the segment, the updated DataFrame is saved to a new CSV file in the output directory.

CSV File Structure

- Consider a example tweet from the (Benaulim Beach segment)
- Tweet is as follows:

“Benaulim MLA Capt Venzy Viegas expressed concern about the rise in stray dogs, particularly on beaches, posing risks to children. Viegas called for proper infrastructure, sterilization, and clear marking to distinguish between sterilized and non-sterilized dogs. https://t.co/TZPMrnFel7"

For the above tweet the output will be generated in the following way:

Tweet	Roberta's Sentiment	DistilBert Sentiment	OpenAi sentiment	OpenAi analysis	Open AI suggestion

Tweet	Sentiment	Emotion	GPT Sentiment	GPT Reasoning	GPT Suggestions
Tweet	Negative	Fear	Negative	The tweet expresses concern about the rise in stray dogs, particularly on beaches, posing risks to children. This highlights a negative issue which can be concerning to parents and the community. The call for proper infrastructure, sterilization, and clear marking indicates that there are currently shortcomings in addressing the problem, adding to the negative sentiment.	Increase funding for sterilization programs to control the population of stray dogs. Implement strict regulations for pet owners to ensure their dogs are properly cared for and not let loose to become strays. Establish designated areas or shelters for stray dogs to reduce their presence in public areas. Educate the public on responsible pet ownership and the importance of reporting stray dogs to the authorities.

Understanding Output Structure:

- Tweet: The original text of the tweet.
- Sentiment: Sentiment label (e.g., Negative, Neutral, Positive) by Roberta
- Emotion: Emotion lab.el by Distil-Bert model
- GPT Sentiment: The sentiment analysis provided by the GPT model.
- GPT Reasoning: Explanation or analysis provided by the GPT model for the assigned sentiment.
- GPT Suggestions: Suggestions for policy changes or actions provided by the GPT model.

4.1.1 Aggregate Summary of Sentimental & OpenAI Analysis

Name of the segment	Benaulim Beach
Overall sentiment analysis	
Positive Score	0.19
Negative Score	0.54
Neutral Score	0.27
Overall Emotional analysis	
Fear	0.73
Joy	0.15
Neutral	0.04
Anger	0.04
Saddness	0.04
GPT Analysis and suggestion	- Implementing stricter regulations or guidelines for handling and managing animals in public spaces to prevent such incidents from occurring in the future. - Increasing awareness and education among tourists about safety precautions to take when interacting with animals in unfamiliar environments. - Implement strict regulations on coastal development and construction near sand dunes to protect the natural environment. - Increase local community involvement in decision-making processes regarding land use and conservation efforts. - Raise awareness about the importance of sand dunes in preserving the coastline and protecting against erosion. - Provide alternative sustainable livelihood opportunities for community members that may be impacted by restrictions on development near sand dunes. - Promote eco-friendly tourism practices that prioritize conservation and preservation of natural habitats. - Conducting a thorough investigation to determine the cause of death of the seagulls and taking appropriate measures to prevent similar incidents in the future.

Table 27: Example 1 of aggregrate output model

Name of the segment	Arambol Beach
Overall sentiment analysis	
Positive Score	0.19
Negative Score	0.62
Neutral Score	0.19
Overall Emotional analysis	
Fear	0.57
Anger	0.14
Sadness	0.10
Neutral	0.10
Surprise	0.05
Joy	0.05
GPT Analysis and suggestion	- Implement regulations on minimum purchase requirements at beach shacks to prevent harassment of tourists. - Ensure transparency in pricing by requiring shack owners to display rates and inform customers upfront about any minimum purchase policies. - Provide avenues for tourists to report such incidents and take action against establishments engaging in unethical practices. - Develop infrastructure and facilities to make beaches more welcoming to a wider range of visitors. - Implement marketing strategies that highlight the unique features and attractions of the beach to appeal to a broader audience. Implement and enforce strict laws and regulations regarding littering on beaches. - Increase public awareness campaigns to educate people on the importance of keeping beaches clean. - Provide more waste management facilities and recycling bins at beaches. - Improved beach safety measures such as lifeguards, warning signs, and safety instructions to prevent drownings.

Table 28: Example 2 of aggregate output model

Name of the segment	Calangute beach
Overall sentiment analysis	
Positive Score	0.29
Negative Score	0.29
Neutral Score	0.42
Overall Emotional analysis	
Fear	0.41
Joy	0.29
Neutral	0.06
Anger	0.06
Saddness	0.12
Disgust	0.06
GPT Analysis and suggestion	Implement stricter penalties for littering offenders. - Invest in proper waste management infrastructure on the beaches. Strengthening and enforcing consumer protection laws to prevent such scams and protect individuals from being exploited. - Increasing police presence and surveillance in areas known for scams and criminal activities, such as tourist hotspots. Implementing effective stray dog control measures, such as neutering and vaccination programs, to reduce their population and prevent any aggressive behavior. - Increasing awareness among the locals and tourists about how to safely interact with stray dogs and what to do in case of any encounters. - Collaborating with animal welfare organizations to address the issue of stray dogs in a humane and sustainable manner. - Strengthening enforcement of existing laws related to stray dogs to ensure public safety and maintain a positive tourism experience.

Table 29: Example 3 of aggregate output model

Name of the segment	Dirty beach
Overall sentiment analysis	
Positive Score	0.05
Negative Score	0.66
Neutral Score	0.29
Overall Emotional analysis	
Fear	0.34
Joy	0.07
Suprise	0.02
Anger	0.17
Saddness	0.32
Disgust	0.07
GPT Analysis and suggestion	Improve infrastructure for waste segregation, recycling facilities, and public toilets in tourist areas. - Increase the frequency of cleaning and maintenance of public spaces such as beaches, markets, and bus stands. - Encourage community participation through initiatives like beach clean-up drives and waste management programs. - Seek feedback from tourists and locals to identify problem areas and prioritize actions for improvement. - Invest in sustainable tourism practices that promote eco-friendly initiatives and responsible tourism behaviors. - Crack down on illegal activities such as drug trafficking and illegal mining to improve the overall image and safety of the state. - Improving regulations and oversight for taxi services to prevent overcharging and dishonest practices. Introducing measures to control hotel costs and ensure fair pricing for visitors. - Enhancing safety measures and increasing police presence in popular tourist areas to address safety concerns and prevent incidents of crime.

Table 30: Example 4 of aggregate output model

Name of the segment	Vagator beach
Overall sentiment analysis	
Positive Score	0.23
Negative Score	0.32
Neutral Score	0.45
Overall Emotional analysis	
Fear	0.86
Joy	0.14
GPT Analysis and suggestion	-Providing adequate facilities and amenities for visitors, such as clean washrooms, seating areas, and garbage disposal options. Marriott could respond publicly to the concerns raised in the tweet, acknowledging the issues and committing to addressing them in a transparent and responsible manner. - The organization could work with local authorities and environmental organizations to improve waste management practices, including proper disposal of sewage and other pollutants. - Marriott could engage in community initiatives to clean up and preserve the surrounding natural environment, including Vagator Beach and the Arabian Sea. Encouraging sustainable tourism practices that minimize harm to wildlife and their habitats. - Enforce stricter regulations and monitoring on hotels and businesses to prevent the discharge of harmful substances into the sea. - Increase penalties for those found violating environmental regulations. Regularly monitor and report on water quality at Vagator beach to ensure the health and safety of both residents and tourists. - Implementing stricter regulations and enforcement to protect nesting sites and habitats of endangered species like Olive Ridley turtles. - Encourage sustainable practices and proper waste management by providing incentives to businesses that follow environmentally friendly practices.

Table 31: Example 5 of aggregate output model

4.2 Conclusion

After thorough research and analysis, we have reached the conclusion that Twitter data can indeed provide valuable insights for policy-making in the tourism sector of Goa. Our findings align with previous studies, confirming that social media analytics, particularly Twitter sentiment analysis, is an effective tool for gauging public opinion and sentiment regarding tourism-related issues. The data extracted from Twitter has proven instrumental in understanding the needs, preferences, and grievances of tourists, which are crucial for crafting informed and effective tourism policies. This research underscores the potential of leveraging Twitter data to enhance strategic decision-making and ultimately contribute to the sustainable development of Goa's tourism industry. Thus, we advocate for the integration of social media insights into the policy-making process to ensure that the initiatives are responsive to the evolving dynamics of tourist interactions and expectations in Goa.

4.3 Limitations

- **Multilingual Challenges**: Since we've just worked with tweets in English so models trained on one language might not perform well on another due to linguistic differences.
- **Bias and Generalization:** Sentiment analysis models can be biased towards certain types of data they were trained on. So to remove the biasness we will hvae to run the model with a huge number of data.
- **Domain Specificity:** Models trained on one type of data might not perform well on another type of data.
- **Emotion Complexity**: Sentiment analysis typically categorizes text into positive, negative, or neutral categories, but emotions are often more complex than that. People can express mixed feelings or emotions that don't fit neatly into these categories.

4.4 Future work

- **Model Fine-tuning Optimization:** Continuously optimize the fine-tuning process to improve accuracy and efficiency. Experiment with different hyperparameters, training strategies, and model architectures to achieve better performance.
- **Custom Model Development**: Explore the possibility of developing a custom sentiment analysis model tailored specifically to our domain. Fine-tune a pre-trained model on a large corpus of domain-specific text data to create a more specialized sentiment analysis solution.
- **Integration with Real-time Data Streams**: Develop a system for real-time classification, sentiment analysis and generative suggestions by integrating the model with data streaming platforms.
- **User Interface Enhancement**: Improve the user interface to make it more intuitive, user-friendly, and visually appealing.
- **Scalability and Performance Optimization**: Optimize the scalability and performance to handle large volumes of data efficiently.

REFERENCES

[1] "Goa losing on quality tourists, lament tourism stakeholders" - 01 Jan 2024 | O Heraldo Team - https://www.heraldgoa.in/Goa/Goa-losing-on-quality-tourists-lament-tourism-stakeholders-/215957

[2] "Tourism Development in Goa: Trends, Importance and Challenges" - Geetanjali Achrekar, International Journal of Multidisciplinary in Management and Tourism

[3] "Russia goes to presidential polls with war economy lifting living standards" - Business Standard | Mar 15 2024 https://www.business-standard.com/world-news/russia-goes-to-presidential-polls-with-war-economy-lifting-living-standards-124031401292_1.html

[4] UK, US tourists keep numbers up as Russians give Goa the miss - Praveena Sharma, Gomantak Times | Published on: 03 Nov 2023

[5] This is why wealthy Russians and Indians are travelling to Dubai instead of Goa - By Jerome Anthony | CNBC TV18 November 10, 2023

[6] "Deep Analyzing Public Conversations: Insights from Twitter Analytics for Policy Makers" Nimish Joseph, Purva Grover, Polaki Kishor Rao, P. Vigneswara Ilavarasan

[7] Social media news worldwide - statistics & facts

[8] Social network sites and acquiring current affairs knowledge: The impactof Twitter and Facebook usage on learning about the news | Mark Boukes | JOURNAL OF INFORMATION TECHNOLOGY & POLITICS2019, VOL. 16, NO. 1, 36–51https://doi.org/10.1080/19331681.2019.1572568

[9] How many people come to Twitter for news? As it turns out, a LOT | By Twitter News | Monday, 12 September 2022 | https://blog.x.com/en_us/topics/insights/2022/how-many-people-come-twitter-for-news

REFERENCES

[10] Comparing BERT against traditional machine learning text classification

[11] Comparative Analyses of BERT, RoBERTa, DistilBERT, and XLNet for Text-based Emotion Recognition | November 2020 | Francisca Adoma Acheampong School of Computer Science and Engineering University of Electronic Science and Technology of China Chengdu, China

[12] The impact of polices on government social media usage: Issues, challenges, and recommendations John Carlo Bertot, Paul T. Jaeger *, Derek Hansen

[13]https://tourism.gov.in/sites/default/files/2022-09/India%20Tourism%20Statistics%20at%20a%20Glance%20200%20%28Eng%29.pdf

[14] https://www.cnbctv18.com/travel/destinations/this-is-why-wealthy-russians-and-indians-are-travelling-to-dubai-instead-of-goa-18298751.htm

[15]https://www.heraldgoa.in/Goa/Goa-losing-on-quality-tourists-lament-tourism-stakeholders-/215957

[16] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00781-w

[17]https://tourism.gov.in/sites/default/files/2023-01/Brief%20Note%20Dec%2022.pdf

[19] https://console.apify.com/actors/61RPP7dywgiy0JPD0/console

[20] Using an Evidence-Based Approach for Policy-Making Based on Big Data Analysis and Applying Detection Techniques on Twitter
S Labafi, S Ebrahimzadeh

[21]Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making
P Burnap, ML Williams - Policy & internet, 2015 - Wiley Online Library

REFERENCES

[22] The public-facing policy agenda of state legislatures: The communication of public policy via twitter

[23] Analyzing and visualizing government-citizen interactions on Twitter to support public policy-making
Of activists and gatekeepers: Temporal and structural properties of policy networks on Twitter

[24] Studying the generation of alternatives in public policy making processes
V Ferretti, I Pluchinotta, A Tsoukiàs - European Journal of Operational …, 2019 - Elsevier

[25] Data generation as sequential decision making
P Bachman, D Precup - Advances in Neural Information …, 2015