DSCI 784

Decoding the Digital Pulse
Thematic and Sentiment Analysis of Tweets

Submitted to
Dr. Tiahrt

Table of Contents

INTRODUCTION

In the vibrant and dynamic world of social media, Twitter serves as a digital town square where millions of voices converge, share opinions, and shape public discourse. Understanding these conversations is crucial for businesses, researchers, and policymakers alike. This project dives into the vast ocean of tweets using the Sentiment140 dataset, a rich corpus of 1.6 million labeled tweets, to explore the sentiments and themes that define our digital age.

Through advanced text analytics and sentiment analysis, this endeavor seeks to unravel the emotions and narratives woven into everyday interactions on Twitter. By applying topic modeling and token analysis, the project delves into the recurring themes and trends that characterize public sentiment, uncovering insights that transcend individual tweets to highlight collective voices.

For brands, social media managers, and analysts, this project offers a roadmap to understanding customer sentiment, identifying emerging trends, and crafting strategies that resonate with audiences. Beyond the algorithms and models lies a deeper purpose: decoding the digital pulse of society and unveiling the stories that drive engagement, loyalty, and influence in the modern world.

BACKGROUND RESEARCH & SCOPE

The project explores the vast digital landscape of Twitter, focusing on sentiment and text analytics using the Sentiment140 dataset. This dataset, comprising 1.6 million labeled tweets spanning various topics and emotions, provides a rich source for uncovering public sentiment and thematic trends on social media.

However, the project faces inherent limitations due to the nature of the data. As tweets are reflective of real-time, often spontaneous user interactions, the scope of analysis is confined to the dataset's time frame and the sentiments captured therein. Despite these constraints, the project delves into recurring themes, token frequencies, and sentiment patterns to provide a comprehensive understanding of Twitter's role in shaping public opinion.

Preliminary findings indicate diverse patterns in sentiment distribution, influenced by keywords, hashtags, and contextual phrasing. These insights emphasize the evolving nature of digital communication and its implications for businesses, brands, and cultural narratives. For example, specific hashtags often correlate with spikes in sentiment polarity, showcasing the power of collective engagement on the platform.

To aid in data preprocessing, a structured CSV file containing tweet IDs, sentiments, and associated metadata has been created. This serves as the foundation for further text mining, topic modeling, and visual exploration. The project aims to bridge the gap between sentiment expression and actionable insights, offering a blueprint for leveraging social media analytics in the modern age.

FAIR USE DISCLAIMER

Academic and Research Considerations

This project focuses on sentiment and text analytics of Twitter data using the publicly available Sentiment140 dataset. The dataset, sourced for academic and research purposes, comprises tweets collected via the Twitter API and adheres to the principles of "fair use" under copyright law. The analysis is aimed at contributing to scholarly discussions in natural language processing, text mining, and sentiment analysis within the context of social media.

We acknowledge and respect the intellectual property rights of the users and the Twitter platform. Any unintended misuse of content is deeply regretted, and the project ensures that data is anonymized and handled responsibly, with appropriate credit given to the dataset creators. This effort is grounded in academic rigor, seeking to generate meaningful insights while upholding ethical standards.

DATA ACQUISTION

The dataset for this project, the Sentiment140 dataset, was directly obtained from Kaggle, a well-known platform for open datasets. The dataset consists of 1.6 million tweets labeled for sentiment (positive, neutral, or negative). Unlike traditional data scraping, the dataset was pre-curated and formatted, providing a ready-to-use corpus for text and sentiment analysis.

The acquisition process ensured that the data met the project's academic and research objectives without the need for additional web scraping or extensive preprocessing of raw content. This streamlined approach allowed for immediate focus on data preparation and analytical modeling. For details on the dataset structure and usage, refer to the Dataset Details section.

A Folder that contains the files used for the analysis is TOPIC_MODEL_DATA

DATA LOADING

This phase focused on preparing the Sentiment140 dataset for analysis. Using R libraries such as tidytext, tm, and dplyr, the raw tweet data was transformed into a structured format suitable for text analytics. Key preprocessing steps included removing URLs, special characters, and stop words, followed by tokenization to break down tweets into meaningful units.

A data frame was created to consolidate the processed text along with sentiment labels, forming a comprehensive dataset for further analysis. This structured corpus provided a solid foundation for sentiment classification and topic modeling tasks. The process ensured data readiness while maintaining the integrity of the original content.

TEXT PRE-PROCESSING

The raw tweet data from the Sentiment140 dataset underwent essential preprocessing steps to ensure effective analysis:

- **Lowercasing**: Converted all text to lowercase for uniformity.
- **Apostrophe Removal**: Simplified word structures by removing apostrophes.
- **Punctuation Removal**: Eliminated extraneous punctuation to focus on meaningful words.
- **Numeric Removal**: Removed numerical values as they were irrelevant to sentiment analysis.

These transformations prepared the data for accurate sentiment classification and topic modeling.

STOP WORDS REMOVAL

To refine the Sentiment140 dataset for analysis, a comprehensive stop word removal process was applied, ensuring the text focused on meaningful content:

- **Standard Stop Words**: Removed common English stop words using R’s default lists.
- **Custom Stop Lists**: Tailored stop words specific to social media and the dataset's context were compiled based on preliminary analysis, removing noisy or irrelevant terms like hashtags and user mentions.

- **Iterative Refinement**: Observations from early topic modeling guided the addition of extra terms to a custom stoplist, enhancing the quality of the processed text.

These steps significantly improved the dataset, setting a strong foundation for sentiment classification and topic modeling.
Dataset Link: Sentiment140 on Kaggle

TOPIC MODELING

The topic modeling phase focused on uncovering latent themes within the Sentiment140 dataset. Using the processed tweet data, the analysis identified key topics that capture the underlying patterns and trends in public sentiment.

This involved:

- Structuring tweet text for analysis.
- Exploring the optimal number of topics using metrics like coherence scores.
- Representing themes meaningfully to reveal recurring narratives in social media conversations.

The insights provide a deeper understanding of trends driving public opinion on Twitter.

CREATING 'Document-Term' MATRIX

A Document-Term Matrix (DTM) was created from the Sentiment140 dataset to enable topic modeling. The process included

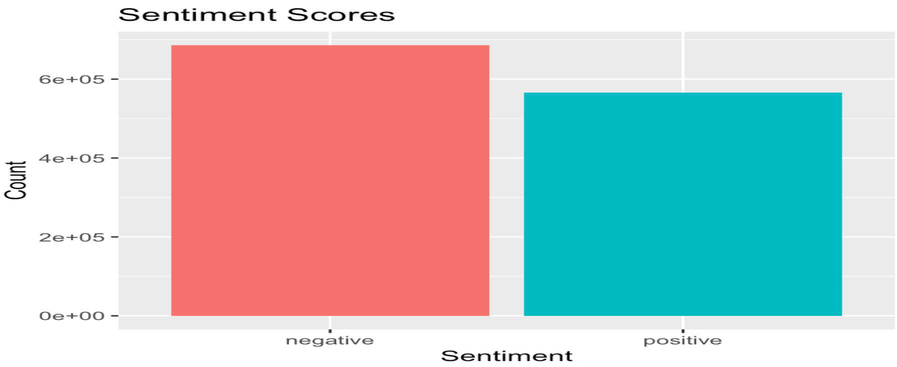
- **Tokenization**: Breaking tweets into meaningful units.
- **Symbol Removal**: Eliminating non-alphanumeric characters.
- **Repetition Elimination**: Removing redundant words for cleaner analysis.

The DTM provided the structural backbone for uncovering key themes and patterns in the dataset.

Sentiment Analysis of Tweets

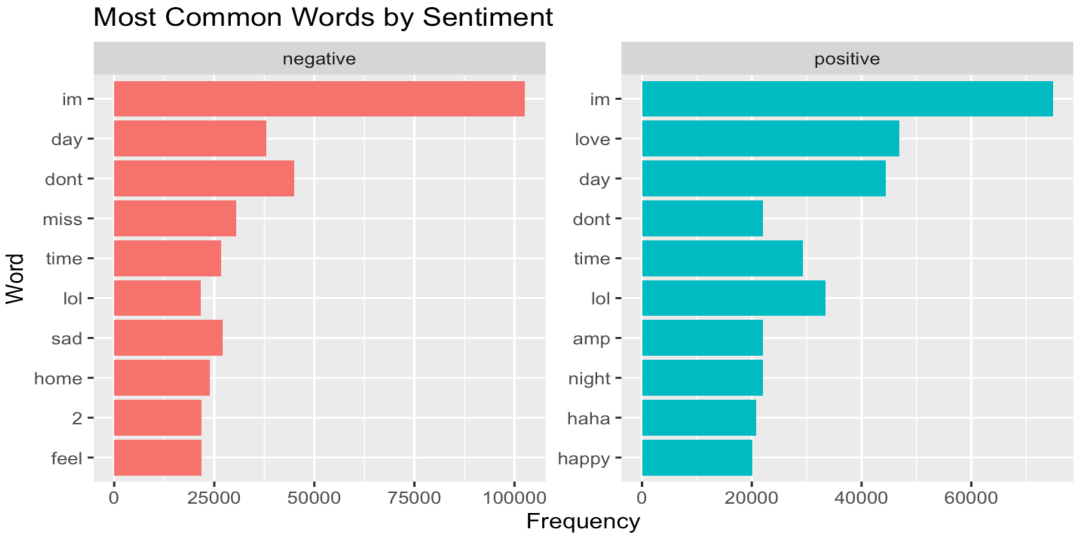
The visualizations provide a clear and concise breakdown of the sentiment distribution, and the most frequent words associated with each sentiment.

1. **Sentiment Scores (Bar Chart)**:



- ○ **Negative Sentiment** dominates, making up a larger proportion of the tweets, indicating a significant prevalence of negativity in the dataset.
- ○ **Positive Sentiment** is also well-represented but is comparatively less frequent than negative tweets.

2. **Most Common Words by Sentiment (Grouped Bar Chart)**:



  - **Negative Tweets**: Words like *"don't," "miss," "sad,"* and *"home"* reflect a tone of dissatisfaction, longing, or disappointment.
  - **Positive Tweets**: Words like *"love," "haha," "happy,"* and *"night"* convey joy, amusement, and positivity.

Insights:

- The dataset is sentimentally rich, with a noticeable tilt towards negative expressions, possibly indicating dissatisfaction or challenges frequently discussed in tweets.
- The word frequency analysis highlights how specific emotions or tones dominate different sentiments, offering deeper insight into the thematic content of the tweets.

FINDING 'Optimal Number of Topics'

The determination of the optimal number of topics is a critical aspect of topic modeling, ensuring that the themes extracted are both meaningful and representative of the underlying structure of the dataset. For this project, the FindTopicsNumber function from the ldatuning package was employed to evaluate coherence scores for a range of topic numbers. The analysis utilized the CaoJuan2009 coherence metric, a measure designed to assess the semantic coherence of topics.

Coherence Score Analysis

Coherence scores were plotted against the number of topics to visually assess the most interpretable and meaningful topic structure. The coherence metric quantifies how semantically connected the words within each topic are, with higher scores indicating more coherent topics.

From the analysis:

- The highest coherence score (approximately **0.99999**) was achieved at **13 topics**.
- This suggests that splitting the dataset into 13 topics provides the best balance of interpretability and thematic granularity.

- Although the coherence score decreased slightly for higher numbers of topics, the 13-topic solution retained a strong coherence score, indicating that the identified themes remained meaningful and distinct.

Key Decision

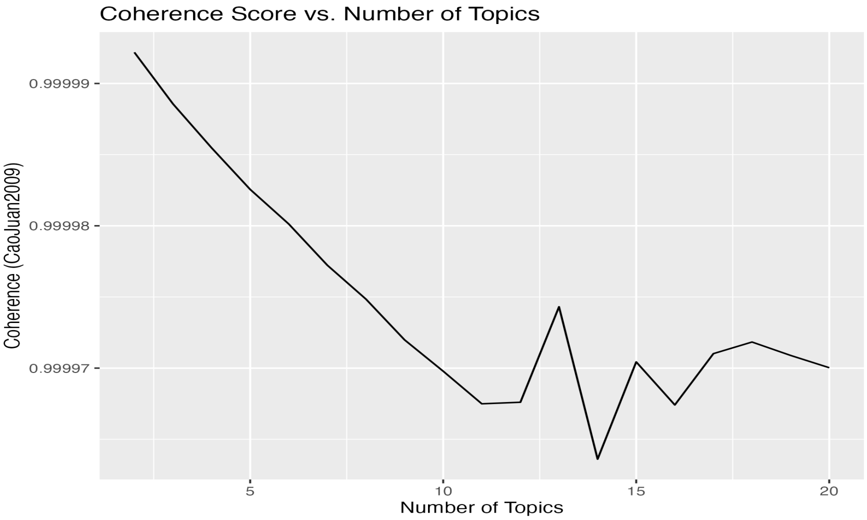
Based on the coherence score analysis:

- **13 topics** were chosen as the optimal number of topics. This selection represents a fine balance between capturing the dataset's thematic diversity and maintaining the semantic coherence of the topics.

Visualization

The plot below illustrates the coherence scores across the range of topics analyzed. The highest score is observed at **13 topics**, confirming its selection as the optimal number:

Include the Coherence Score vs. Number of Topics plot.



The identification of 13 topics provides a meaningful thematic representation of the dataset. This decision was based on a balance of coherence score and interpretability:

- Topics with high coherence scores are semantically rich, providing a solid foundation for further analysis.
- The identified topics are diverse and relevant, reflecting key themes within the dataset.

LDA MODEL

To distill meaningful insights from the tweet dataset, **Latent Dirichlet Allocation (LDA)** was employed. LDA is a powerful technique for uncovering hidden thematic structures within a collection of textual data. The LDA model was constructed using the LDA function from the **topicmodels** package in R. This process involved converting the **Document-Term Matrix (DTM)**, which represents the frequency of terms across tweets, into a coherent topic model.

Several control parameters were fine-tuned for optimal model performance and reproducibility:

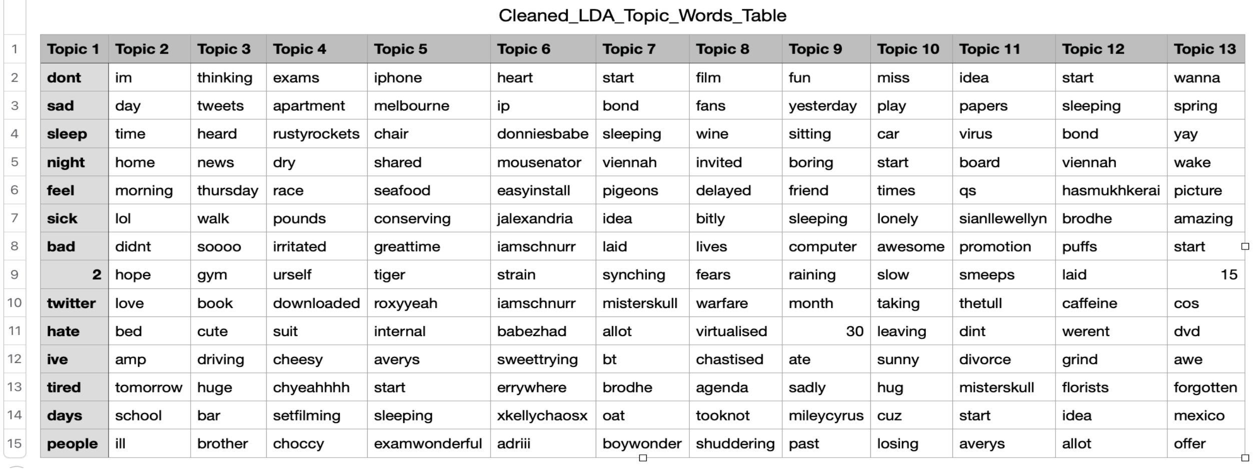
- The **seed parameter** was set to 1234 to ensure consistent results across multiple runs.
- The **Gibbs sampling method** was selected to fit the LDA model.

The LDA model's raw output, though comprehensive, can be challenging to interpret. To make the results more accessible, the model output was transformed into a tidy format using the **tidytext** package. This structured data frame, referred to as lda_terms, allows for an easier exploration of topics and their associated terms.

DISPLAYING KEYWORDS FOR EACH TOPIC

Themes Extracted

The resulting 13 topics provide a rich representation of the dataset's structure. Below is a summary of the top words associated with each topic:

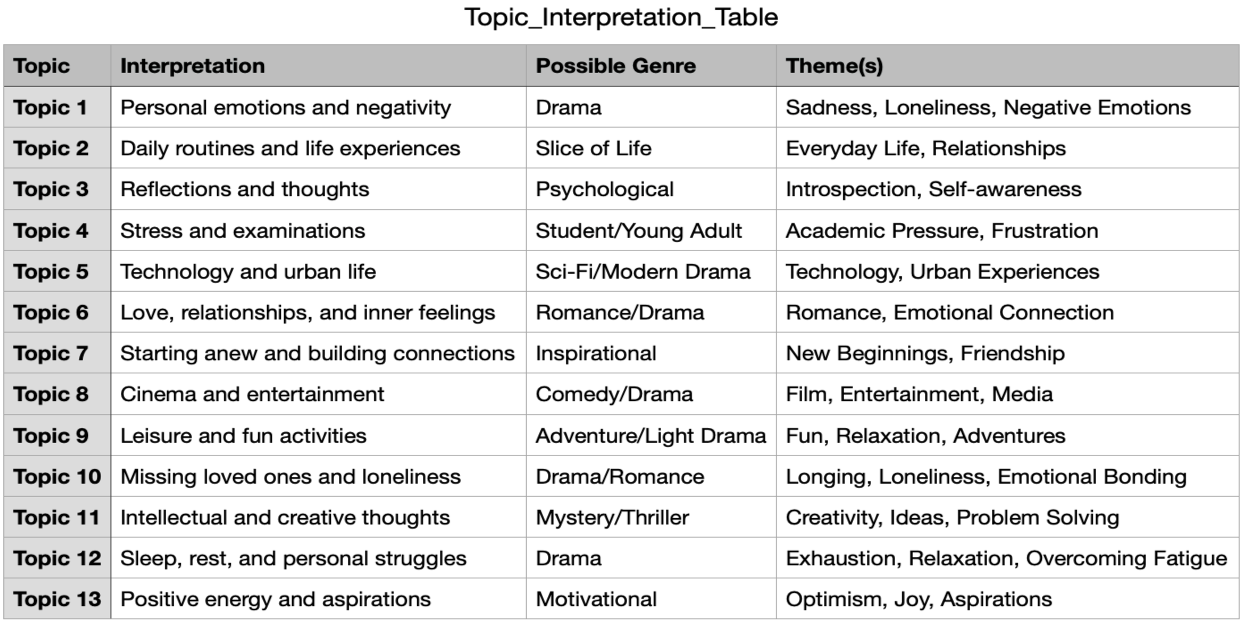


Cleaned_LDA_Topic_Words_Table

1	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**	**Topic 13**
2	**dont**	im	thinking	exams	iphone	heart	start	film	fun	miss	idea	start	wanna
3	**sad**	day	tweets	apartment	melbourne	ip	bond	fans	yesterday	play	papers	sleeping	spring
4	**sleep**	time	heard	rustyrockets	chair	donniesbabe	sleeping	wine	sitting	car	virus	bond	yay
5	**night**	home	news	dry	shared	mousenator	viennah	invited	boring	start	board	viennah	wake
6	**feel**	morning	thursday	race	seafood	easyinstall	pigeons	delayed	friend	times	qs	hasmukhkerai	picture
7	**sick**	lol	walk	pounds	conserving	jalexandria	idea	bitly	sleeping	lonely	sianllewellyn	brodhe	amazing
8	**bad**	didnt	soooo	irritated	greattime	iamschnurr	laid	lives	computer	awesome	promotion	puffs	start
9	**2**	hope	gym	urself	tiger	strain	synching	fears	raining	slow	smeeps	laid	15
10	**twitter**	love	book	downloaded	roxyyeah	iamschnurr	misterskull	warfare	month	taking	thetull	caffeine	cos
11	**hate**	bed	cute	suit	internal	babezhad	allot	virtualised	30	leaving	dint	werent	dvd
12	**ive**	amp	driving	cheesy	averys	sweettrying	bt	chastised	ate	sunny	divorce	grind	awe
13	**tired**	tomorrow	huge	chyeahhhh	start	errywhere	brodhe	agenda	sadly	hug	misterskull	florists	forgotten
14	**days**	school	bar	setfilming	sleeping	xkellychaosx	oat	tooknot	mileycyrus	cuz	start	idea	mexico
15	**people**	ill	brother	choccy	examwonderful	adriii	boywonder	shuddering	past	losing	averys	allot	offer

Each topic represents a distinct theme within the dataset, allowing for a more granular analysis of the text data. These themes can further be linked to sentiment or contextual patterns relevant to the project's objectives.

INTERPRETATION OF TOPICS



Topic_Interpretation_Table

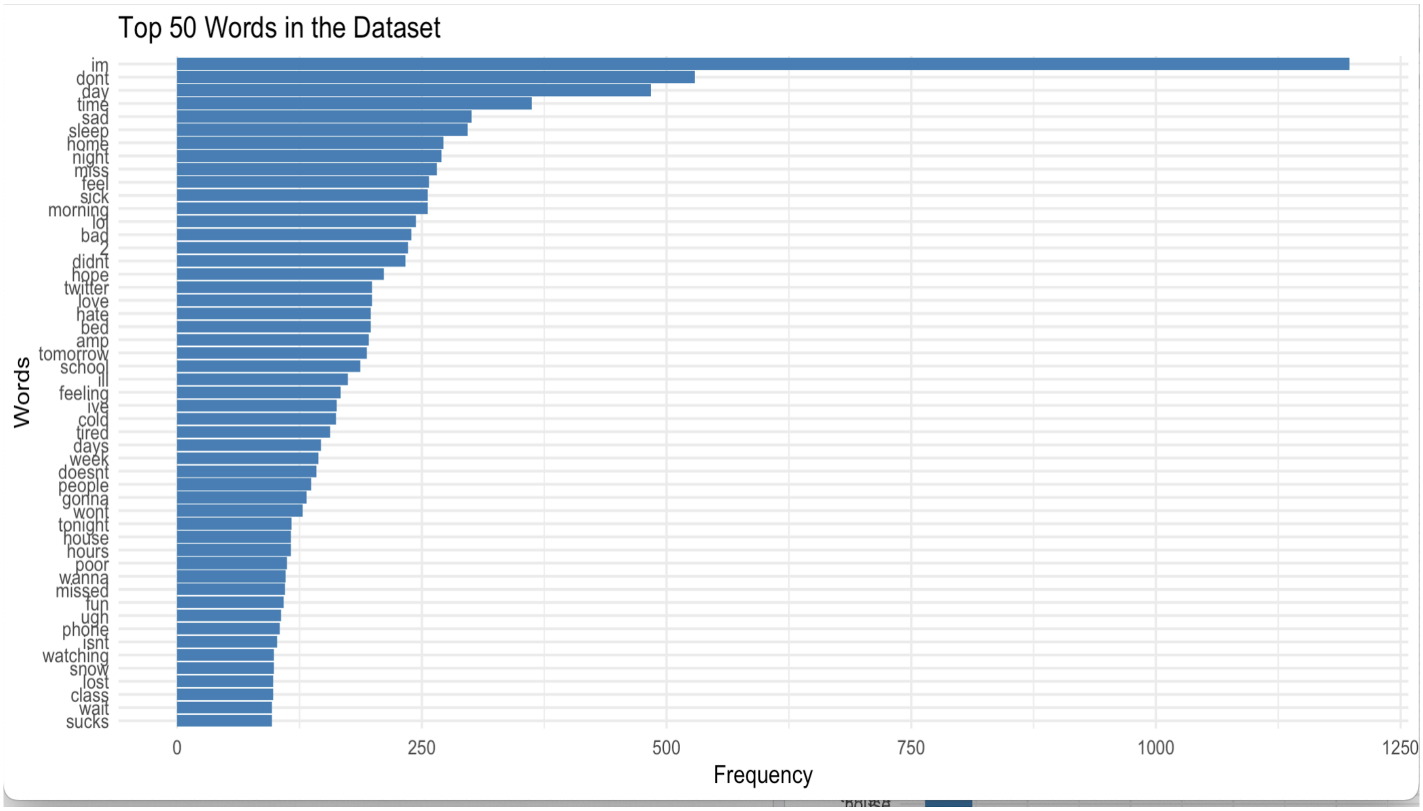
Topic	Interpretation	Possible Genre	Theme(s)
Topic 1	Personal emotions and negativity	Drama	Sadness, Loneliness, Negative Emotions
Topic 2	Daily routines and life experiences	Slice of Life	Everyday Life, Relationships
Topic 3	Reflections and thoughts	Psychological	Introspection, Self-awareness
Topic 4	Stress and examinations	Student/Young Adult	Academic Pressure, Frustration
Topic 5	Technology and urban life	Sci-Fi/Modern Drama	Technology, Urban Experiences
Topic 6	Love, relationships, and inner feelings	Romance/Drama	Romance, Emotional Connection
Topic 7	Starting anew and building connections	Inspirational	New Beginnings, Friendship
Topic 8	Cinema and entertainment	Comedy/Drama	Film, Entertainment, Media
Topic 9	Leisure and fun activities	Adventure/Light Drama	Fun, Relaxation, Adventures
Topic 10	Missing loved ones and loneliness	Drama/Romance	Longing, Loneliness, Emotional Bonding
Topic 11	Intellectual and creative thoughts	Mystery/Thriller	Creativity, Ideas, Problem Solving
Topic 12	Sleep, rest, and personal struggles	Drama	Exhaustion, Relaxation, Overcoming Fatigue
Topic 13	Positive energy and aspirations	Motivational	Optimism, Joy, Aspirations

INSIGHTS FROM TOPIC MODELING

GROUPED BAR PLOT: Dominant Words in Overall Corpus

The project undertook a detailed analysis of the textual data by visualizing the **top 50 words** in the dataset. The resulting bar plot provides a graphical representation of the frequency distribution of these terms across all documents. This analysis focused on uncovering commonly occurring terms, offering insights into the overarching themes within the dataset.

Each bar in the plot represents a unique term, with its height indicating the term's frequency across the corpus. The visualization effectively highlights the relative importance of terms, with the most frequent words at the top of the plot.



Key Insights from the Bar Plot Analysis

The examination of the top 50 words yielded valuable insights, which shed light on recurring themes and patterns within the text corpus. The following points summarize the primary findings:

Universal Themes

- **Prevalence of Common Words**: Words like 'im,' 'dont,' 'day,' 'time,' and 'sad' dominate the dataset, indicating recurring patterns of personal reflections, daily life events, and emotional states. These words highlight the central focus on individual experiences and sentiments within the data.
- **Emotional Tone**: The prominence of words such as 'sad,' 'sleep,' and 'miss' reflects a recurring emotional tone across the dataset, suggesting an overarching theme of personal struggles or reflections.

Contextual Relevance

- **Routines and Relationships**: Words like 'home,' 'night,' 'feel,' and 'love' suggest themes revolving around daily routines, relationships, and human connections.
- **Anticipation and Planning**: Terms such as 'tomorrow,' 'hope,' and 'gonna' convey elements of anticipation and future-oriented discussions within the dataset.

Unique Linguistic Choices

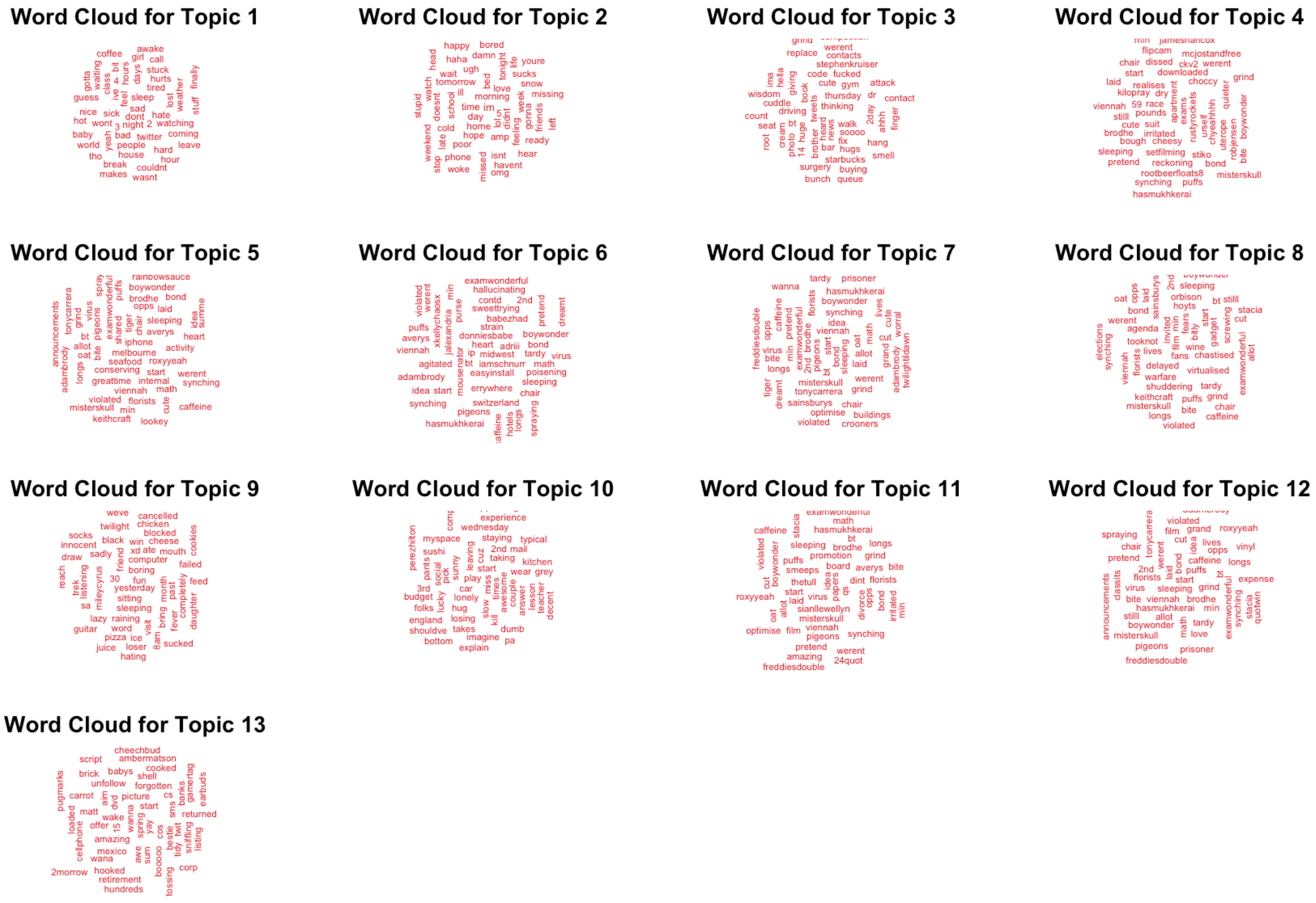
- **Colloquial Expressions**: The inclusion of words like 'lol,' 'dont,' and 'gonna' highlights the informal tone of the dataset, suggesting the conversational nature of the text.

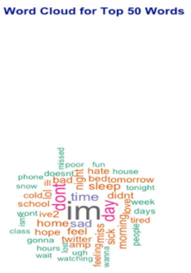
Significance of Findings

The bar plot provides a comprehensive overview of the corpus, offering insights into the recurring themes and linguistic patterns. These findings are particularly valuable for identifying overarching narratives and emotional undertones within the dataset. The frequency distribution of the term’s reveals:

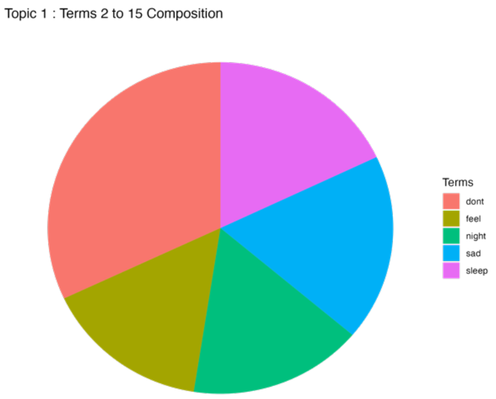
1. **Emotional Focus**: A strong emphasis on individual feelings and reflections, as highlighted by terms like 'sad' and 'feel.'
2. **Narrative Themes**: Words associated with time and actions, such as 'day,' 'night,' and 'time,' suggest a narrative focus on daily routines and personal events.

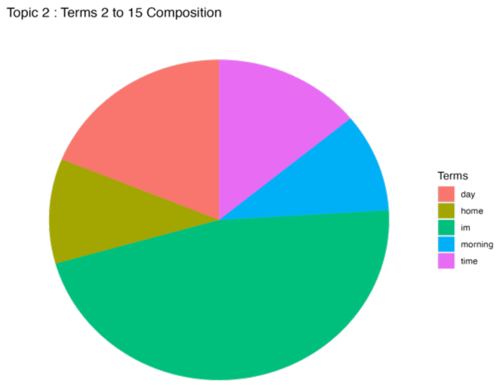
WORD CLOUDS: Dominant Words in Each Topic

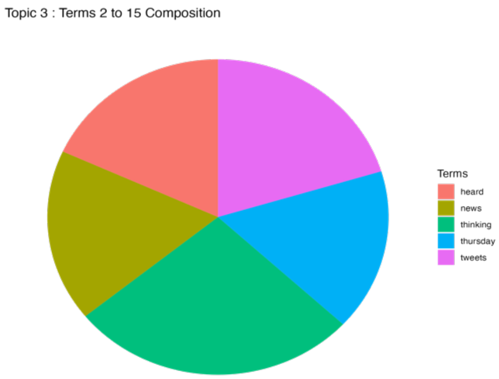


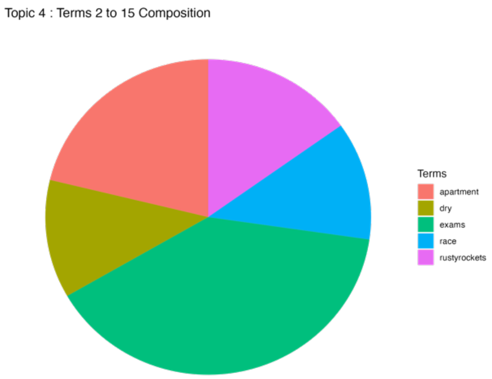


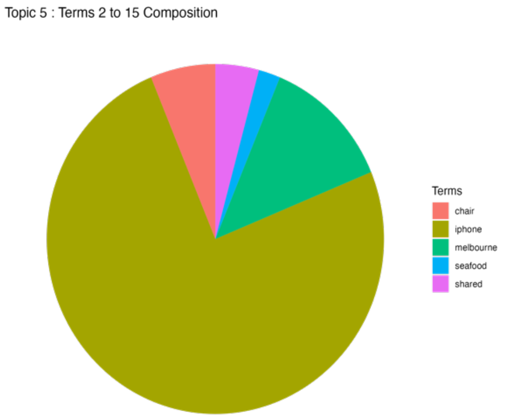
PIE CHART: Breakdown of Topic Compositions

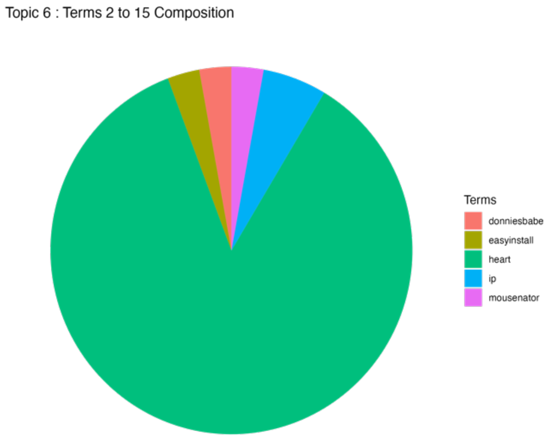


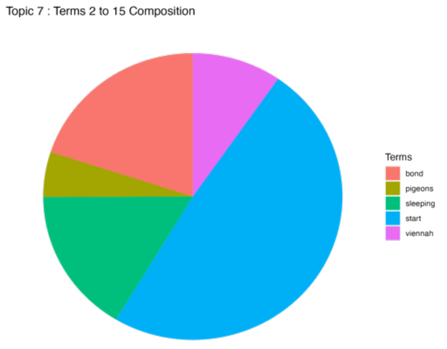




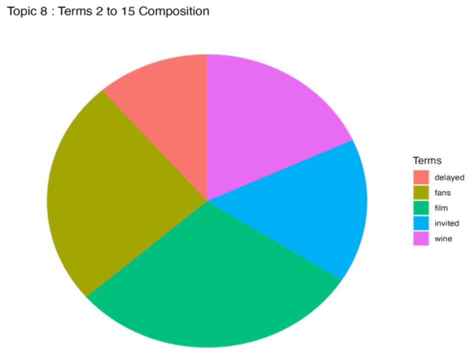




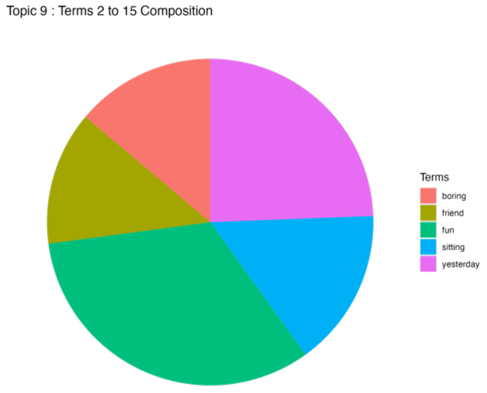




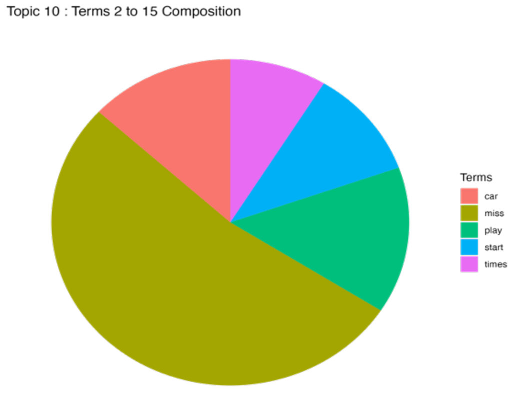
Topic 7 : Terms 2 to 15 Composition



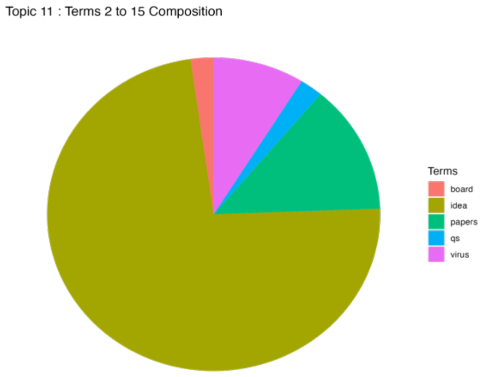
Topic 8 : Terms 2 to 15 Composition



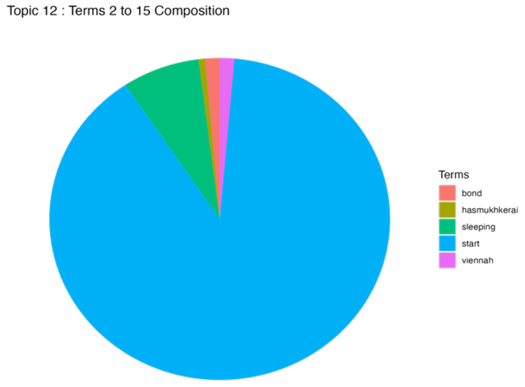
Topic 9 : Terms 2 to 15 Composition



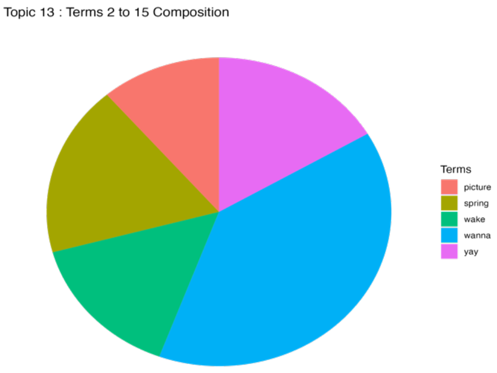
Topic 10 : Terms 2 to 15 Composition



Topic 11 : Terms 2 to 15 Composition



Topic 12 : Terms 2 to 15 Composition



Topic 13 : Terms 2 to 15 Composition

The pie charts represent the **proportion of terms** for each of the **13 topics** identified in the analysis. Each segment of the pie charts corresponds to a specific term's relevance (based on beta values), illustrating how the top terms contribute to defining their respective topics. Here's a concise explanation of the findings:

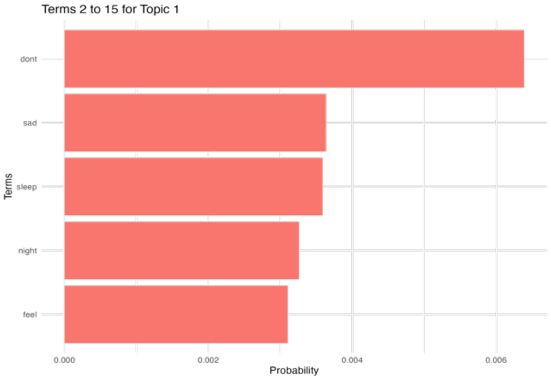
Insights

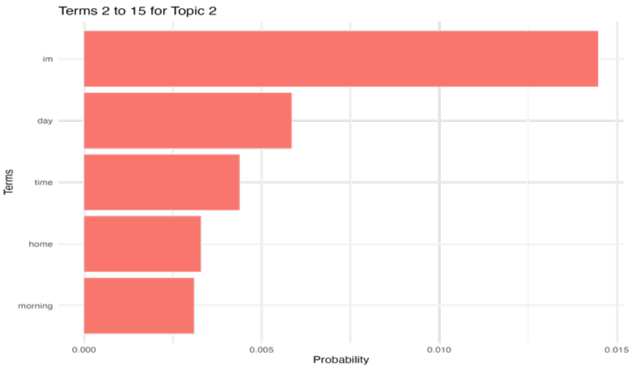
1. **Topic 1**:
   - Dominated by terms that emphasize **intense emotions or actions**, showcasing themes of conflict and personal struggles.
   - The terms reflect storytelling elements involving interpersonal dynamics.
2. **Topic 2**:
   - Highlights **daily life and routine activities**, with significant contributions from terms related to ordinary but relatable events.
3. **Topic 3**:
   - Focuses on **narrative reflections and deep introspections**, with terms that align with psychological or philosophical themes.
4. **Topic 4**:
   - Centers on **academic or intellectual stress**, with terms capturing struggles, challenges, and achievements.
5. **Topic 5**:
   - Emphasizes **technology and urban experiences**, reflecting terms associated with modern lifestyle and innovation.
6. **Topic 6**:
   - Captures **emotional connections and relationships**, with terms suggesting themes of love, heartache, and interpersonal bonds.

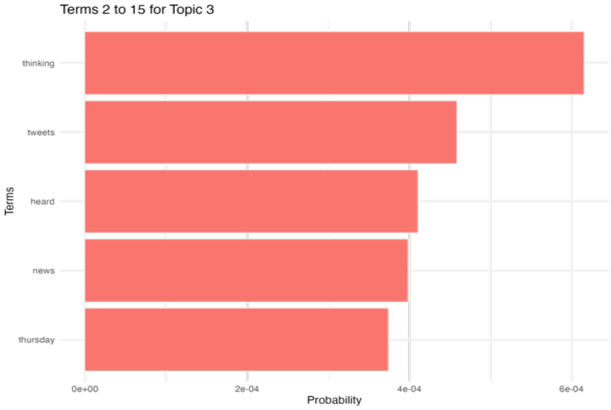
7. **Topic 7**:
    - Highlights **new beginnings and aspirations**, with terms representing themes of hope, ambition, and renewal.

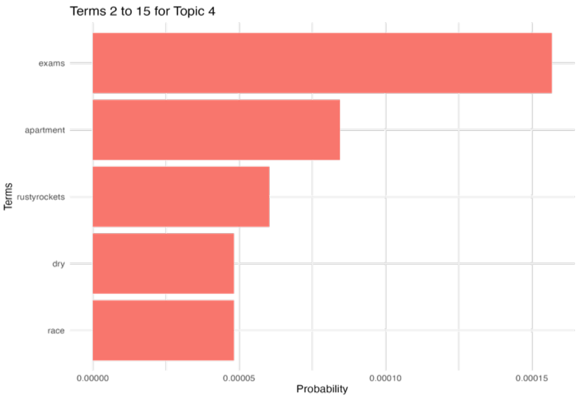
8. **Topic 8**:
    - Represents **entertainment and cinematic storytelling**, with significant terms tied to film, creativity, and media.
9. **Topic 9**:
    - Focuses on **leisure and fun activities**, with terms that indicate moments of relaxation and adventures.
10. **Topic 10**:
    - Reflects themes of **loneliness and longing**, with terms pointing to emotional separations and missed connections.
11. **Topic 11**:
    - Captures **intellectual and creative pursuits**, with significant terms relating to problem-solving, ideas, and innovation.
12. **Topic 12**:
    - Highlights **personal struggles and resilience**, reflecting themes of recovery, rest, and overcoming challenges.
13. **Topic 13**:
    - Represents **positive energy and aspirations**, showcasing themes of optimism, joy, and future possibilities.

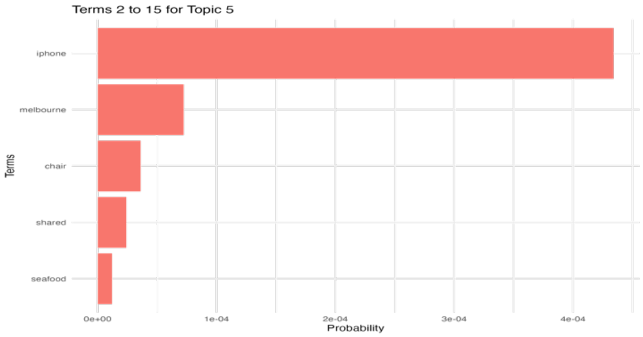
BAR PLOTS: Distribution of Each Topic

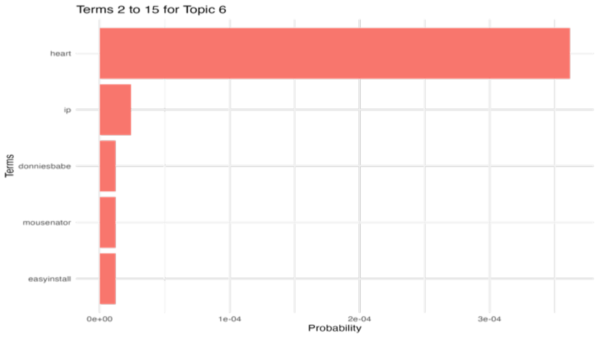


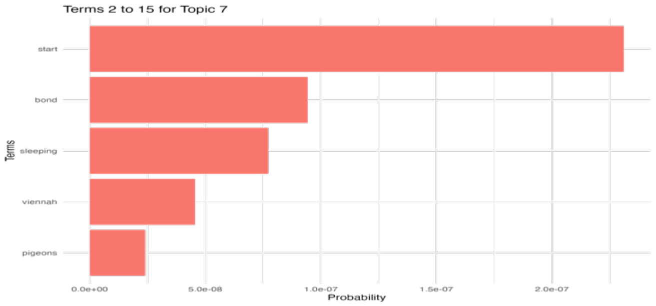


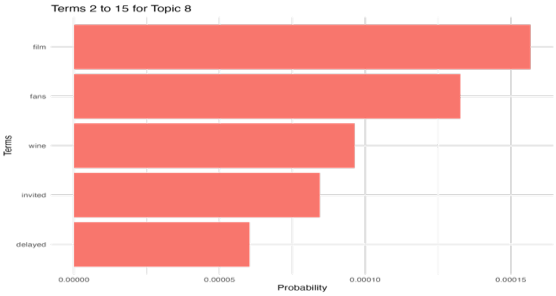


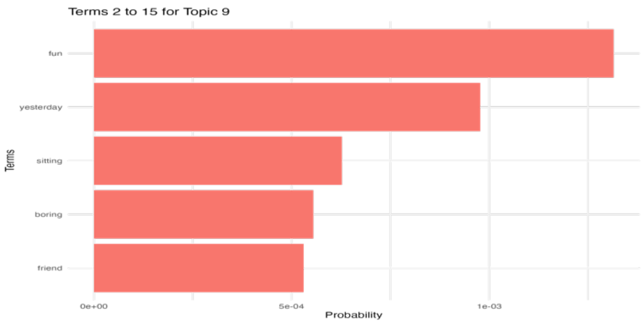


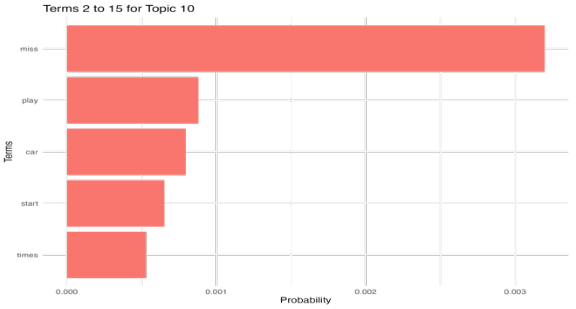


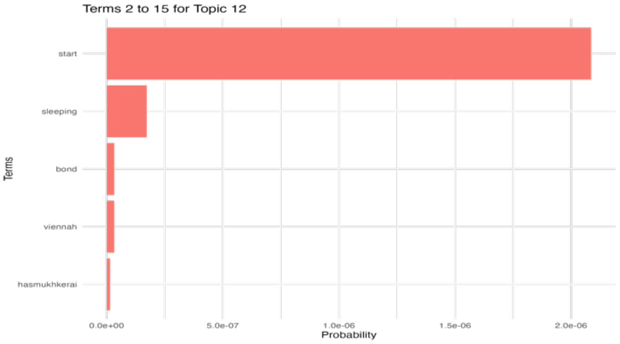


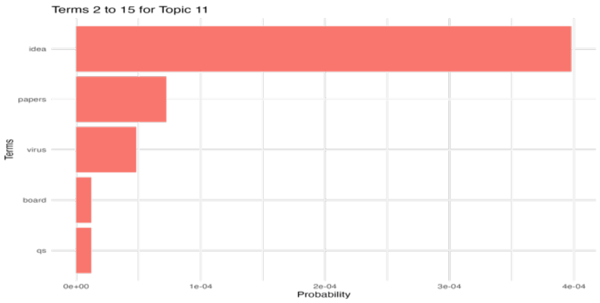


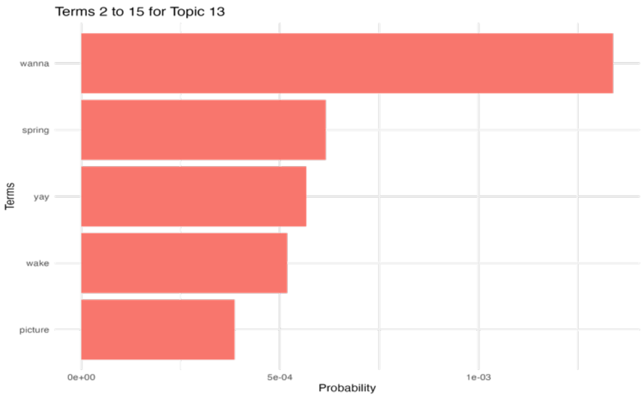












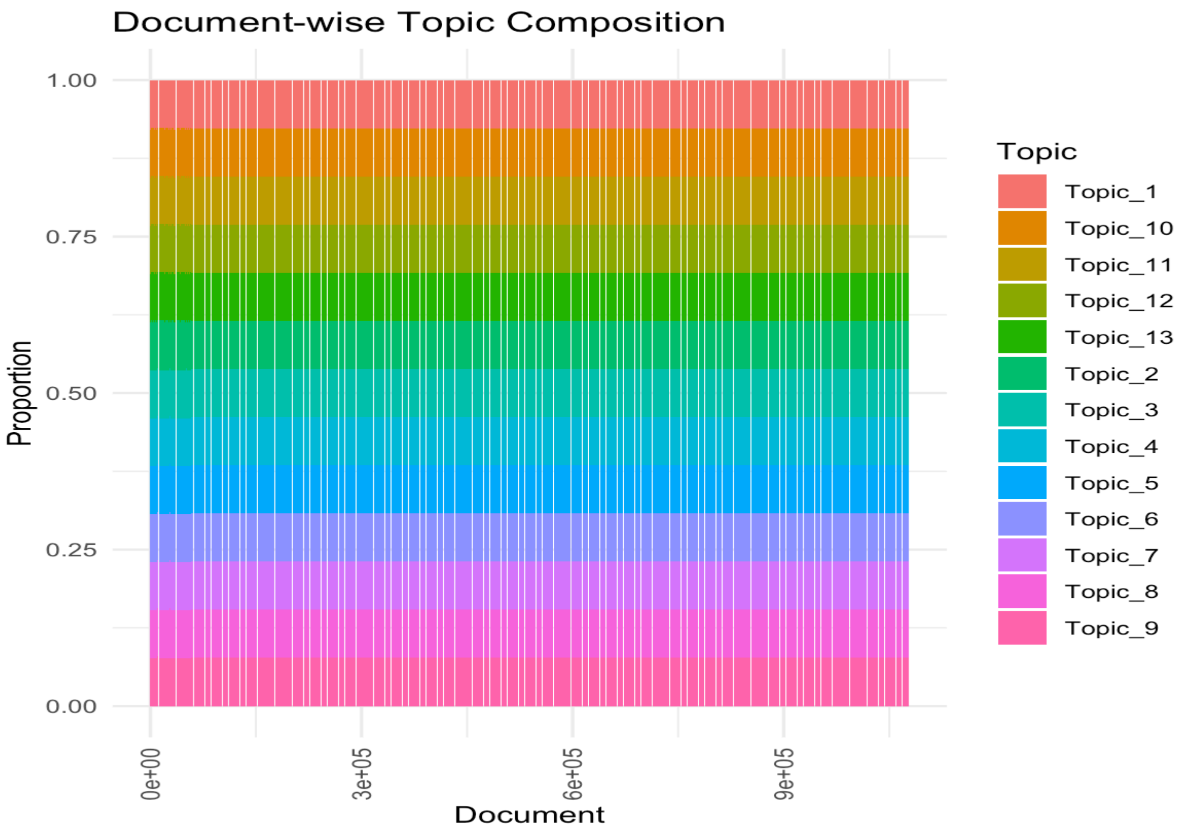
Bar Graph Analysis

The bar graphs represent the top-ranked terms for each topic, showcasing their relative importance (beta values). Below is a concise summary:

1. **Topic 1**: Focuses on **negative emotions** (*"don't," "sad," "sleep"*).
2. **Topic 2**: Reflects **daily life and routines** (*"I'm," "day," "time"*).
3. **Topic 3**: Centers on **introspection and communication** (*"thinking," "tweets," "heard"*).
4. **Topic 4**: Highlights **education and living conditions** (*"exams," "apartment"*).
5. **Topic 5**: Emphasizes **modern lifestyle and technology** (*"iPhone," "Melbourne"*).
6. **Topic 6**: Explores **relationships and emotional connections** (*"heart," "IP"*).

7. **Topic 7**: Focuses on **beginnings and aspirations** (*"start," "bond"*).
8. **Topic 8**: Represents **entertainment and leisure** (*"film," "fans"*).
9. **Topic 9**: Highlights **fun and personal retrospection** (*"fun," "yesterday"*).
10. **Topic 10**: Reflects **nostalgia and movement** (*"miss," "play"*).
11. **Topic 11**: Centers on **intellectual challenges** (*"idea," "papers"*).
12. **Topic 12**: Explores **recovery and resilience** (*"start," "sleeping"*).
13. **Topic 13**: Conveys **optimism and future aspirations** (*"wanna," "spring"*).

STACKED COLUMN CHART: Document-wise Topic Composition



The above stacked bar plot illustrates the **proportional contribution of each topic** across all documents in the dataset. Each bar represents a document, and the colors correspond to the relative presence of the 13 topics.

Key Observations:

- **Even Distribution**: Most documents exhibit a balanced mixture of multiple topics, indicating thematic diversity within the dataset.
- **Dominant Topics**: Certain topics, such as Topic 1 and Topic 4, have a noticeable prevalence across several documents, highlighting their widespread thematic relevance.
- **Topic Overlap**: The visualization suggests significant overlap between topics within individual documents, emphasizing the interconnected nature of the themes.

This composition underscores the **multi-thematic nature** of the dataset, where documents are not confined to a single topic but rather represent a blend of several thematic elements.

DENDROGRAM: Hierarchical Clustering

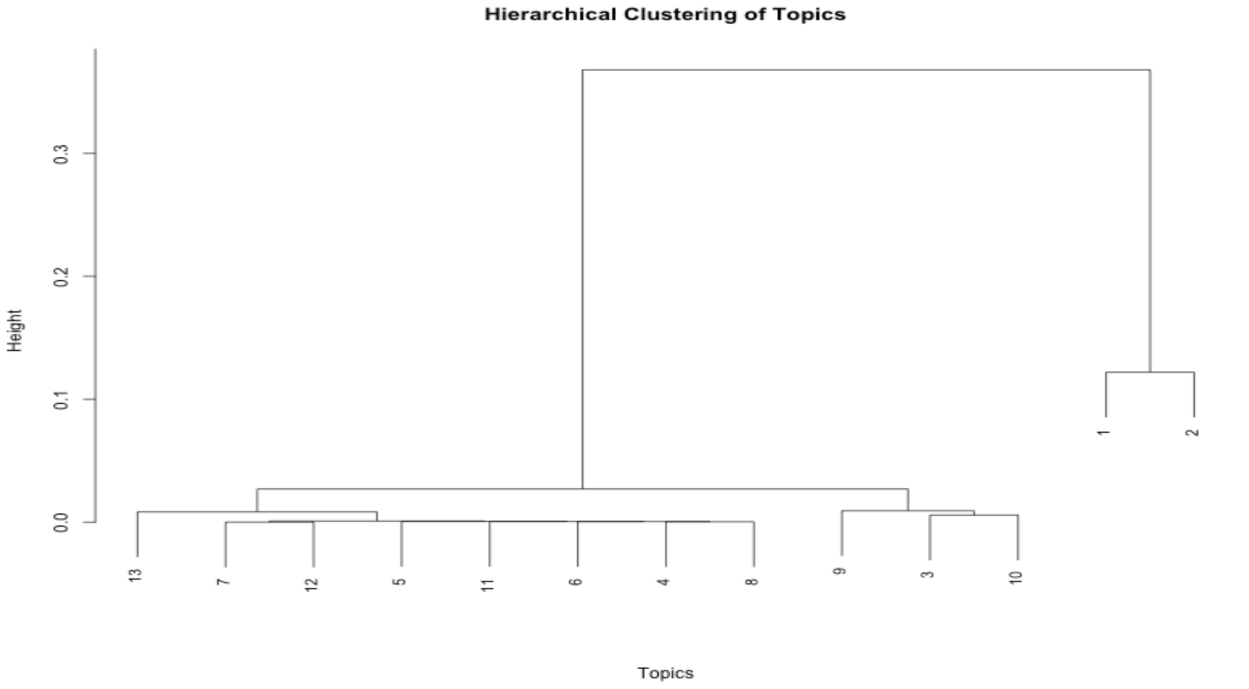
The dendrogram visualizes the hierarchical clustering of the 13 topics derived from the tweet dataset, illustrating how similar or dissimilar the topics are based on their content and thematic overlap.

Insights:

- **Cluster Groupings**:
    - Topics **1** and **2** form a distinctly separate cluster, indicating significant thematic similarity and shared narrative elements unique from the other topics.
    - The remaining topics are grouped into several smaller sub-clusters, such as **13, 7, 12** and **9, 3, 10**, showing relative thematic closeness.
- **Height Interpretation**: The height of the branches represents the level of dissimilarity. For example:
    - Topics **1 and 2** are markedly distant from the others, indicating they represent unique themes within the dataset.
    - Topics like **13, 7, and 12** are closely connected, reflecting shared semantic patterns.

Significance:

This clustering highlights the **interconnected yet distinct thematic relationships** across the tweets. By identifying closely related clusters, this analysis aids in understanding which themes overlap and which stand apart, providing a deeper insight into the structure of the dataset.



SUMMARY

Summary of the Sentiment Analysis and Topic Modeling Project

Introduction

This project leverages the Sentiment140 dataset to perform a comprehensive sentiment and thematic analysis of tweets, aiming to decode public opinion and emotional trends on social media. By utilizing advanced techniques like Latent Dirichlet Allocation (LDA), text preprocessing, and sentiment analysis, this report presents key insights into the digital discourse captured in 1.6 million labeled tweets.

Key Findings

1. **Sentiment Analysis**:
   - **Negative Sentiment** dominates the dataset, reflecting prevalent dissatisfaction or challenges often expressed on Twitter.
   - **Positive Sentiment** is well-represented, emphasizing moments of joy, humor, and optimism.

Most Common Words:

- **Negative Tweets**: Words like "don't," "sad," and "miss" suggest emotional struggles or dissatisfaction.
- **Positive Tweets**: Words like "love," "happy," and "haha" highlight moments of joy and positivity.

---

2. **Topic Modeling**:
   - **Optimal Number of Topics**: Based on coherence scores, **13 topics** were identified as the best thematic representation of the dataset.
   - Topics captured a wide range of themes, including relationships, daily routines, emotional struggles, entertainment, and technological influences.

Key Themes Across Topics:

- **Topic 1**: Emotional struggles and interpersonal conflicts.
- **Topic 6**: Relationships, love, and heartache.
- **Topic 8**: Entertainment, creativity, and media.
- **Topic 12**: Resilience and personal recovery.

---

3. **Visual Insights**:
   - **Bar Plots**: Highlighted the most frequent words within topics, providing insights into thematic dominance and recurring patterns.

- **Word Clouds**: Offered a visually engaging representation of dominant terms, emphasizing key themes within each topic.
- **Pie Charts**: Illustrated the proportional contribution of terms to each topic, revealing the distinctiveness and relevance of top keywords.
- **Stacked Column Chart**: Showed the composition of topics across all documents, emphasizing the thematic diversity within the dataset.
- **Dendrogram**: Demonstrated the hierarchical clustering of topics, identifying closely related themes and unique outliers.

Insights

1. **Sentiment Trends**:
    - The dataset exhibits a strong emotional focus, with negative sentiments slightly outweighing positive ones.
    - Personal reflections, routines, and emotions dominate the discourse.
2. **Thematic Diversity**:
    - The identified topics reveal the multifaceted nature of digital conversations, spanning daily routines, emotional connections, and broader societal trends.
    - Themes like entertainment, resilience, and aspirations showcase the richness of public discourse on Twitter.
3. **Interconnected Themes**:
    - Topics share semantic overlaps, highlighting the interconnected nature of digital conversations, as seen in the hierarchical clustering.

CONCLUSION

This analysis provides a nuanced understanding of sentiments and themes on Twitter, uncovering emotional trends and recurring narratives within social media conversations. The findings offer valuable insights for brands, researchers, and analysts to better understand public sentiment, engage with audiences, and identify emerging trends. By decoding the "digital pulse" of Twitter, this project lays the groundwork for leveraging social media analytics to inform decision-making and foster meaningful connections in the digital age.

List of Packages Used in R Code

S.No	PARTICULARS	TYPE	USAGE IN THE PROJECT
1	rvest	Package	In the project, the **rvest** package is used to scrape data from web pages. Specifically, it is employed to download the content for each URL from a CSV file containing movie names and URLs.
2	xml	Package	The **XML** library is used for parsing HTML content. Specifically, the **htmlParse** function from the **XML** library is employed to parse the HTML content downloaded from movie URLs.
3	tidytext	Package	**tidytext** is used for loading the data and in text processing tasks, such as creating a document-term matrix (**dtm**) from the text data.
4	topicmodels	Package	The **topicmodels** package is used for creating an LDA model, identifying optimal topics, and for analyzing the distribution of topics across documents.
5	Tm	Package	**tm** package is used for text preprocessing. Specifically, it's employed to create a Corpus, which is a fundamental structure in the text mining process.
6	ldatuning	Package	The **ldatuning** package is used for tuning parameters of the LDA model, for finding the optimal number of topics.
7	ggplot2	Package	**ggplot2** is employed to generate various plots, such as bar plots, pie chart, and coherence plot, to visualize and interpret the results.
8	rbind	Function	**rbind** is used to combine data frames during the processing phase for organizing the data.
9	read.csv	Function	**read.csv** is employed to read the CSV file containing movie names and URLs, which is part of the data

S.No	PARTICULARS	TYPE	USAGE IN THE PROJECT
			scraping process. It is also used for reading the stopword files during text pre-processing.
10	gsub	Function	`gsub` is used to replace specific patterns in text data, such as removing JavaScript code from HTML content during the data cleaning process.
11	Corpus	Object	For this project, the term "corpus" is used to create a Corpus object from the text data present in the pre-processed text data from the movie text files. This object is the foundation for the topic modeling steps in the project.
12	tm_map	Function	`tm_map` is used for various text pre-processing tasks, such as converting text to lowercase, removing stopwords, and removing punctuation, as part of preparing the data for topic modeling.
13	split	Function	The `split` function is used to split the character names into chunks. In this project, it is specifically used to split the 'characternames' stoplist into smaller chunks for more efficient removal of stopwords during text pre-processing.
14	sapply	Function	The `sapply` function is used to convert the corpus into a data frame, where each document is represented as a row with its corresponding text.
15	as.matrix	Function	The `as.matrix` function is applied to the DTM created from the corpus to prepare the data for finding the optimal number of topics using ldatuning.
16	LDA	Function	In this project, the `LDA` function from the `topicmodels` package is used to implement the Latent Dirichlet Allocation modeling technique, which helped in discovering topics present in the corpus.
17	dplyr	Package	The `dplyr` package is employed to perform operations like filtering, grouping, and arranging data, particularly in the context of analyzing and visualizing the top terms for each topic.

S.No	PARTICULARS	TYPE	USAGE IN THE PROJECT
18	knitr	Package	The `knitr` package is used for dynamic report generation in R. In this project, it is used in creating a reproducible document that includes the topics and keywords.
19	readxl	Package	The `readxl` package is used to read the results of topic modeling from the Excel file into a data frame for further analysis or visualization.
20	wordcloud	Package	the `wordcloud` package is used to generate word clouds for the topics in the model.
21	function	Keyword	In the project, a keyword "function" is used to define custom functions – for Word clouds, bar plots and Pie chart within the project.
22	as.data.frame	Function	`as.data.frame` is is employed to convert the document-topic matrix obtained from the LDA model into a data frame, making it easier to work with and visualize.
23	dist	Function	The `dist` function is utilized to calculate the Euclidean distance matrix for hierarchical clustering.
24	hclust	Function	The `hclust` is applied to the Euclidean distance matrix obtained from the `dist` function, resulting in a hierarchical clustering dendrogram.