Sentiment Analysis of Cultural Heritage Texts:

Anne Frank Diary of Young Girl

Smita Sanjay Pable
ID: 2543320
AC52010 - MSc Project
(Data Science & Engineering 23/24)
University of Dundee, 2024
Supervisor: Christina Moir

Table of Contents

Table of Figures and tables

Abstract - *Sentiment analysis applied to historical documents helps us in peeling off the layers of cultural background in a novel way. The emphasis of the research is on Anne Frank's emotional diary, which is an important historical document. By harnessing natural language processing (NLP) tools allow us to dip into the sea of sentiments and bring them to a surface. Anne Frank's diary gives us access to the inner world of a young Jewish girl – her fears, thoughts and dreams that were set against the greatest tragedy the world has ever gone through the World War II. NLP algorithms enable us to discover joy, sadness, courage, and fear emotions that reflect throughout time. Sentiment analysis helps us unwrap such subtleties which give us a glimpse into Anne's inner universe. In her hiding inside that attic, her hopelessness, her glimpses of happiness and sheer defiance are all vividly depicted. Through the combined efforts of the disciplines this study sheds light on the complexity of human emotions including resilience and the lasting effects of historical narratives. The bravery of Anne Frank, depicted in her diary, sends out a powerful life-lesson that remains resonant throughout history.*

1 Introduction

In the era of digital technologies, where machines translate language convey stories, the place of NLP methods in historical discourse hints at the journey into the unexplored territory. In the fading colour of the ink from the letter, the parchment manuscript, and the brittle pages of the diary remain—the footprints of lives, of the dreams and the challenges that lie in the way. The emotional component of history is scrutinized by our research, beginning with Anne Frank's diary, an outstanding example from the 20th century. Anne, a teenager who was living in Netherlands, received a diary as her 13th birthday gift, she wrote this diary for two years from July 1942 to July 1944 while she was imprisoned in secret annex. Anne dies at the age of 15 during holocaust. Her father only survivor (from Frank family) of holocaust published her diary in 1947 two year after war ended. Anne Frank's diary appeared to be an incredibly personal narrative— a tiny sparrow making an attempt to sing under the storm. Anne's reflections—her joys, fears, and desires, helps us appreciate better what she goes through. We walked carefully, as Anne tried to deal with the consequences of youth, love, and the Nazi threat in the following pages. Using natural language processing (NLP) algorithms, or computational tools, we determined phrases, unveiled metaphors, and study the sentiments richly woven in Anne's writing. Armed with advanced methods, the aim goes farther from mere sadness and happiness with the use of the NRC Emotion Lexicon, emotions were extracted, and sentiment analysis was performed by computing sentiment scores and classifying them using sentiment categories. Besides, to illustrate the outcome, a data visualization technology was applied with Streamlit. It is revealed the nuances shades, quiet bravery, fleeting moments of beauty. The exploration does not stop at an individual sentiments tsunami but instead pushes further to surface the complex temporal patterns of emotions. It invites reflection: What a time was that when airstrikes eroded all hope! And if fear stole the lights of love's bloom? The analyses generate vivid representations that become a mosaic of emotional experiences that is shared. Anne's heartbeats echo the other hearts of soldiers, refugees, dreamers. This project, which incorporates the language with the help of technology and history, can extend our knowledge of cultural heritage. Examining Anne Frank's diary brings to light the uplifting power of the human spirit as well as long-term influence of the diary on our understanding of history and bravery. Anne's courage struck the hearts of many, and her strength became a lighthouse for all generations. In essence, the study of Anne Frank's emotional journey encompasses both space and time, encouraging contemplation on the ability for Understanding, kindness, and resilience. While emerging into her story, one would feel taking a trip back in time and be inspired by her unbeatable spirit of courage, hope and determination when precipitated by difficult moments.

2 Literature review

The primary step in preparing text data for sentiment analysis and other natural language processing (NLP) applications is data cleansing. It incorporates various techniques and methods employed to ensure data accuracy, overall structure and reliability of the data that is processed. The most significant part in data cleaning including the ones such as text normalization, tokenization, stop word removal, and date pattern recognition will be extensively discussed in this literature review with emphasis on their importance to sentiment analysis. First, textual information must be normalised to have a unified format that can be analysed, and text normalization is the initial stage in data cleaning. This step involves cleaning whitespace data, converting text to lowercase, eliminating punctuation in the text and any other anomalies that could be present in the dataset. Researcher can reduce variances and inconsistency in the text by standardizing it; that will make next processing stages easy and enhance the quality of sentiment analysis result.

2.1 Bird, Klein, and Loper (2009)

Highlighted the text normalization methods in their book "Natural Language Processing with Python" [1], by expanding this scope from sentimental analysis to natural language processing (NLP) applications. The Analysis shows the broad range of techniques to normalize the text that will standardize the textual material. These techniques basically aimed to lowercasing the text, removing punctuations, and dealing with whitespace or any other inconsistencies found in the text material. By carefully examine these techniques author highlights the importance of these initial steps in NLP framework. These techniques help to address such inconsistency in the text which can help to avoid inaccurate and unreliable sentiment analysis results. They have suggested useful tips how to preprocess the data, how to focus on important aspects to strategize these NLP techniques to ensure the quality and effectiveness of Sentiment analysis. researchers who are working on such textual data for sentimental analysis, this book work as a foundation. Furthermore, Recognizing the patterns in this textual data is also a crucial task, especially for sentiment analysis, the book also provides a guidance on how to observe such patterns in the text. For which it goes into detail about using special algorithm and expression to identify such patterns. It gives proper explanation on how these algorithm and expression works to find out these patterns like some common date style in text which helps to pull out date related information accurately. By combining language elements and context clues, these algorithms are able to handle various date formats, making text analysis more precise in real-life situations.

2.2 Jurafsky, D., & Martin, J. H. (2019)

In his book, "Speech and Language Processing" [2] (third ed., 2019) provide an insightful overview of basic natural language processing (NLP) methods. It basically covers the tokenization method. which is an initial stage in preparing text for analysis by diving text into meaningful token or words. In addition, author also discuss about the stop words, words with less semantic value end up being discarded Also, examine the normalization techniques. Stemming and lemmatization involves such word-stemming techniques where phrases are converted to their root or base forms to improve the consistency of the dataset and simplifying the analytical processes. Jurafsky and Martin significantly contributed to the field of NLP through their research.

2.3 Manning, Raghavan, and Schütze (2020)

In "Introduction to Information Retrieval," [3] Manning, Raghavan, and Schütze (2020) provides tokenization strategies, discuss different approaches gives valuable insights into methods which are essential for efficiently preparing textual data. These resource offers an in-depth overview of tokenization methods, explaining how to segment text at both word and sentence level. Manning, Raghavan, and Schütze offers outline of these methodologies to efficiently extract relevant information from textual data.

It is essential for researchers to understand these tokenization techniques before started working on sentiment analysis or other NLP projects. Word level segmentation makes it easy to extract key features from text which later will examine for sentiment analysis. On other hand, in sentence level segmentation text divides into sections which helps in textual analysis. Author provides overview of such tokenization methodologies to preprocess the textual data for precise sentiment analysis results.

Furthermore, Manning, Raghavan, and Schütze underline the importance of stop word removal as a preprocessing step. Stop words are the common words with low semantic value. By removing them researcher can reduce the noise and focus on relevant text. Manning et al. focus on the significance of removing these stop words in increasing the precision and quality of sentiment analysis result.

2.4 Mohammad and Turney (2010)

They have made a huge contribution to emotion analysis approaches by not only developing the NRC Emotion Lexicon, but also to data cleaning techniques. Researchers who want to determine emotional content from a text data, lexicon is the essential tool. The method used by Mohammad and Turney including word-emotion relationships into a large lexicon which make it easier to analyse emotional details that are embedded in text.

NRC Emotion Lexicon contains words labelled with emotions; it is easier for researchers to understand the emotional content of textual data. Mohammad and Turney created a systematic labelling method to create link between words and a range of emotions, including joy, sadness, anger, and fear etc. Researchers not only can capture a wide range of emotions but also conduct proper sentiment analysis of textual data using this NRC Emotion Lexicon.

In addition, Mohammad and Turney included the development and validation of this NRC Emotion Lexicon[4] using advanced linguistic and computing techniques. Which will help to ensure lexicon reliability and accuracy. Also, they have included various innovative ways for identifying and classifying emotional expressions in text. Mohammad and Turney created a strong foundation for sentiment analysis research by using these techniques.
They also proved that, in sentiment analysis how important lexicon-based methodologies are. Their approach helps to spot the emotional content in textual data by providing large collection of word-emotion correlations. This allows researchers to understand the useful insights into the underlying sentiments and attitudes expressed by authors. Furthermore, NRC Emotion Lexicon is essential to examine the emotional dynamics of textual data in fields such as psychology, marketing, and social sciences.

2.5 Hutto and Gilbert (2014)

Made a big impact on sentiment analysis by developing the VADER lexicon [13], it is a specialized tool designed to analyse sentiment in social media writing. VADER integrates sentiment scores for emoticons as well as words, in contrast to typical sentiment lexicons. This enables a more contextually relevant and sophisticated analysis of text data from social media networks like Facebook and Twitter. VADER supports fine-grained sentiment analysis using emotive symbols and linguistic clues to identify the nuances of casual, conversational language commonly employed in social media chats. This tool is useful for the researchers who are trying to evaluate sentiment in online conversations.

2.6 Streamlit

Further, A python framework called streamlit app[6] developed by Allaire et al. and companions for dynamic data visualization. Its user-friendly interface and effortless interaction with data science libraries make it more efficient for data analysis. Data visualizations became very easy after introduction of streamlit. Researchers can use various techniques to create attractive as well as interactive visualizations from complex dataset.

2.7 VanderPlas

"Python Data Science Handbook."[5], Book written by VanderPlas offers a thorough study of Python-based data science tools and methodologies in his book. VanderPlas covers various visualization libraries including Matplotlib, Seaborn, and Plotly. The book explains how to create stories that are both visually appealing and educational through in-depth talks and examples. Book can guide researchers who are not that familiar with various visualization techniques. It is a great resource for the people who wants to use best visualization tools for their data analytics project especially combined with frameworks like streamlit.

The textual data will have to be pre-processed, using various techniques, including normalization, tokenization, stop words removal and date pattern recognition before sentiment analysis into account. By Using Lexicon Based soft method and streamlit modern tool, the researcher can understand in detail of hidden emotions dynamics in textual information. This way, one can see human emotions and behaviour from the new direction through the textual analysis.

3 Specification

The Objective of the sentiment analysis of Anne Frank's diary is to explore the emotional journey mentioned in the historical text. Main goal is to extract emotions and analyse and interpret the sentiments communicated by Anne Frank in her diary. It provides the fascinating insights about her secret life in the annex, her experiences and responses throughout the chaotic period of World War II.

3.1 User Stories

3.1.1 User story 1:

As a researcher, my aim is to extract the sentimental aspects from Anne's diary, to understand the emotional journey she went through mentioned in the text.
Acceptance Criteria:
Apply Natural Language Processing techniques to extract emotional content from her diary. Analyse these emotions written by Anne which includes joy, sadness, anger and more.
It is necessary to extract precise emotions to know the depth of Anne Frank's writing.

3.1.2 User Story 2:

As a historian, my goal is to preserve the authenticity of Anne Frank's Diary while doing analysis.
Acceptance Criteria:
Implement precise sentiment analysis techniques which will help to maintain the authenticity of her text.
Avoid the incorrect interpretation of sentiments while doing analysis.
Review the result of sentiment analysis with respect to the historical events and personal experiences of Anne Frank.

3.1.3 User Story 3:

As a data Analyst, I final goal will be to do visualization of emotions or sentiments distribution over time to understand the change in emotions or sentiments of Anne over the period.
Acceptance Criteria:
Add general visualization like timeline graph to evaluate throughout emotions or showcase corelation between two emotions(heatmap). Review the strong emotions with respect to important events to check correctness.
Add interactive charts so that users can modify the graph according to their requirements.

3.2 Methodology and Requirements

In order to fully understand the sentiments folded in the historical data especially like darkest pages of Anne Frank's diary entries, it is crucially important to approach the analysis with deep sense of humanity and empathy.
The Methodology consists of a number of required steps, such as text preprocessing, emotions extraction, sentiment analysis and visualization. Following is the flowchart key Methodology we are adapting:

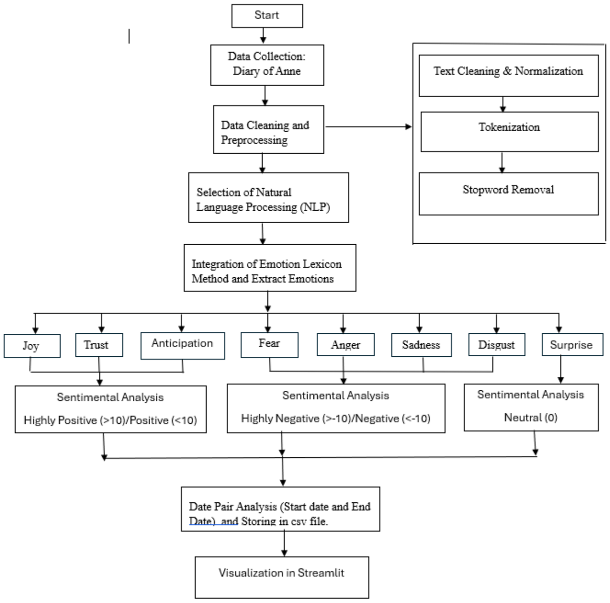

Start
Data Collection: Diary of Anne
Data Cleaning and Preprocessing
Text Cleaning & Normalization
Tokenization
Stopword Removal
Selection of Natural Language Processing (NLP)
Integration of Emotion Lexicon Method and Extract Emotions
Joy
Trust
Anticipation
Fear
Anger
Sadness
Disgust
Surprise
Sentimental Analysis Highly Positive (>10)/Positive (<10)
Sentimental Analysis Highly Negative (>-10)/Negative (<-10)
Sentimental Analysis Neutral (0)
Date Pair Analysis (Start date and End Date) and Storing in csv file.
Visualization in Streamlit

Figure 1: Flow Chart of Sentimental Analysis of Historical Data

- Data Collection: Getting Anne Frank's diary entries.
- Text Preprocessing: Text data processing and formatting for the analysis. This comes with methods that include noise removal, text normalization, tokenization, stop words removal and many more.
- NRC Emotion Lexicon Loading: The NRC Emotion Lexicon that had the assigned words for emotions and sentiments was loaded.
- Sentiment Score Calculation and Labeling: The difference between the positive and negative emotions can be measured. Thus, this helps in calculating the sentiment scores and you can place them in clearly predefined categories.
- Date Pair Analysis: Pick the date-entries in the diary text to know the sentiments and emotions trends over time.
- CSV Data Export: Save extracted emotions and sentiment scores along with the date entries in a CSV file for further analysis.
- Visualizing the extracted data in Streamlit.

3.2.1 Technical Requirements:

- Programming Language: Python was chosen as a primary programming language because of its versatility, and the large number of libraries for NPL such as NLTK. It is also useful for visualization.
- NPL Libraries: With use of NLP libraries like NLTK to preprocess the data and make it ready for sentiment analysis also use of lexical analysis to extract emotional factor from the text.
- Emotion Lexicon: utilizing emotion lexicon like NRC Word Emotion Lexicon, we can enhance the sentiment analysis by extracting emotion along with the sentiment analysis.
- Data Visualization Tools: Using data visualization tools like matplotlib, plotly, seaborn and combine this in streamlit to generate aesthetic and interactive visual representation that will successfully be able to communicate with the users and help them to understand the depth in the sentiment and emotional analysis of Diary.

This approach is mainly based on the NLP techniques with resources from a particular domain like NRC, Word Emotion Lexicon and this developed system performs the precise sentiment analysis on Anne Frank’s diary in a specialized way. This approach is based on text preprocessing, sentiment analysis and display of colours of emotions and sentiments through graph. By utilizing such a systematic approach, the analysis aims to offer a detailed and complex understanding of Anna’s emotional voyage.

3.3 Project Plan

3.3.1 Week 1 and week 2: Acquiring Data and Extracting Initial Text

The main objective of week 1 is to obtain Anne Frank’s diary in PDF format and extract the diary entries from it. It is necessary to get a digital version of the diary from trustworthy source. Verify the extracted data to make sure that no data has been lost or corrupted. Just in case any problem or inconsistencies occurs while extraction will fix to ensure the data integrity. For smooth sentiment analysis, this text will go through some preprocessing steps like normalization.

3.3.2 Week 3 and week 4: Data cleansing and text preprocessing

To make the extracted data ready for analysis, the data needs to be pre-processed and cleaned. So, the main aim is to Normalize the data, tokenize and then remove stopwards from the text. Normalization will reduce the noise from the text, tokenization will divide data in small units called tokens and eliminate the words which carry less importance and where analysis will only focus on the data which has some sentiment weight, which will increase the quality of sentiment analysis.

3.3.3 Week 5 and week 6: Emotion Extraction and Sentiment Analysis Implementation

Later, the focus lies on extracting emotions and applying sentiment analysis. Emotions will extract from the diary with the use of NRC Word Emotion Lexicon. The lexicon contains words labialized with emotions and sentiment rating. This helps to extract emotions. And sentiment score is calculated based on both positive and negative emotions. Next steps need to focus on is labelling sentiments like Highly Positive, Positive, Neutral, Negative, Highly Negative based on score.

Also, store the extracted emotions and sentiments along with date entries in csv file.

3.3.4 Week 7: Development and Interpretation of Visualisations

The project's goal in these weeks is to visualize the sentiment analysis and emotion extraction findings from Anne Frank's diary using techniques like matplotlib, seaborn and plotly in streamlit app. Streamlit is a python(.py) based application, which helps to develop interactive and aesthetic graphs like word cloud, histogram, line chart etc. Which allows to present animated emotions timeline or sentiment distribution and can calculate word density. Showcase sentiments related to the real events and diary entries is also possible which gives us emotional insights found in Anne Frank's diary.

3.3.5 Week 8 to week 11: Reports and Documentation Finalisation

During these weeks the goal is to conclude the project through writing report which will summarize the analysis findings and insights from Anne Frank's diary. Additionally, Project methodology, design, Implementation and Visualization methods are documented for future reference. To guarantee clarity of the project efforts are made. For effective presentation, Poster is prepared along with the power point presentation.

3.4 Time Scale

Start date	End Date	Description	Duration
15 Jan 2024	22 Jan 2024	Dataset exploration and background study, Finalized the Anne Frank's diary as my dataset	1 week
22 Jan 2024	29 Jan 2024	Download the Diary in PDF format, extracting data from PDF and verify the integrity of the extracted data.	1 week
29 Jan 2024	12 Feb 2024	Prepressed the cleaned the raw Anne Frank's diary, Normalized the text, after that tokenized the data and remove stop words to make ready for analyzing.	2 weeks
12 Feb 2024	26 Feb 2024	Extract emotion using NRC Word Emotion Lexicon and calculate sentiment score, based on score label sentiments accordingly. Stored in CSV file	2 weeks
26 Feb 2024	4 March 2024	Visualized the extracted emotions and sentiments in streamlit app and deploy it.	1 week
04 March 2024	11 March2024	Start writing a report to summarize the analysis finding.	1 week
11 March 2024	18 March 2024	Fine tune report and Test streamlit application	1 week
18 March 2024	25 March 2024	Take feedback from supervisor and finalize the report, Create Poster for presentation.	1 week
25 March 2024	01 April 2024	Create power point presentation for project presentation and ready for submission	1 week

3.5 Deliverables

- Sentiment Analysis: A software tool which helps to preprocess the raw diary of Anne Frank, extract emotion and analyze sentiments.
- Streamlit app: App where user can visualize Anne's emotional journey with the help of various interactive graphs.
- Comprehensive Report: Report with detailed summering analysis finding and insights of Anne's diary which will include everything with methodology, data source, implementation, visualization etc.
- Presentation Material: A Poster and Power point presentation for effectively communicating your project outcomes during project presentation.

3.6 Resources

Resources which help for analyzing the text of diary are firstly, Jupiter notebook to run python code to extract data, preprocess it. Furthermore, it is also employed to extract emotions and analyze sentiments After converting data of diary into clean CSV form which contains emotions and sentiments. We can use visual studio to write python code for visualization in streamlit app. Where we can import various techniques like matplotlib, Seaborn and plotly which will help to visualize data in interactive and aesthetic manner.

4 Design

4.1 Choice of Technologies

In constructing the sentiment analysis framework for Anne Frank's Diary, carefully selected the technologies that would support principles of interpretability, accuracy, and robustness. In order to build a robust platform for the analytical process, the chosen technologies were selected with the complexity and the historical nature of the diary's content keeping in mind. Through entire process of sentiment analysis of Anne frank’s diary from preprocessing the data till data visualization, this pipeline requires a seamless interaction of tools which will handle variety of language nuances and extract emotional insights. To achieve significant and relevant results for sentiment analysis purpose, each technology must be evaluated for its suitability in realizing this goal.

4.1.1 Python

Python serves as the fundamental language for the implementing of NPL techniques, such as: data preprocessing, sentiment analysis, and visualization, in this project. Python is famous for its simplicity, readability, and versatility. It gives a rich selection of libraries that are uniquely made to deal with a large set of data science tasks. Particularly for sentiment analysis of Anne Frank’s Diary, python’s flexibility and extensive library architecture allows smooth interaction of various components of analysis pipeline. Because of its simple syntax and detailed documentation, both new and experienced users can use it. Moreover, Python's compatibility with other languages and platforms makes it more useful for interacting with external resources and deploying analysis frameworks within different environments. Its interoperability with libraries such as NLTK, Pandas, Matplotlib, Seaborn, and Streamlit researchers can fully include these specialized tools in sentiment analysis without the need of training the module. Basically, Python is an integral part of creating a custom-built sentiment analysis for text data such as Anne Frank’s Diary since it is easy to learn (as its syntax is user-friendly) and has a large variety of libraries that the data science community commonly uses as a tool for analysis.

4.1.2 NLTK (Natural Language Toolkit):

One of the most integral parts of natural language processing (NLPro) is the Natural Language Toolkit [10], or NLTK. Due to its large set of tools, NTLK plays a major part in the sentiment analysis. Its flexibility implies the fundamental operations need to be executed including text preparation such as the tokenization, stemming, lemmatization, and part-of-speech tagging. While stemming and lemmatization take the words to their basic formation, tokenization is the one that partitions the text into smaller pieces to deal with later. NLTK part-of-speech tagging also allows for in-depth word analysis and providing parts of speech as well. Apart from preprocessing, the NLTK offers a variety of other advanced sentiment analysis tools. It is possible to implement lexicons, machine learning, and rule-based techniques for the purpose of capturing sentiment polarity, intensity, and subjectivity in various types of unstructured data. Including numerous tools within a single coherent structure, NLTK makes sentiment analysis easier to use by adding the ability to see into the nuances of the emotions that Anne Frank describes. Altogether, NLTK is a useful tool for NLP researchers that simplifies their task of getting useful information from textual data.

4.1.3 NRC Emotion Lexicon

NRC Emotion Lexicon is the most important external resource while extracting emotions from text data especially for sentiment analysis of Anne Frank's Diary. The words in this dictionary that have been annotated with their corresponding emotions cover the whole range of human feelings such as joy, sadness, fear, anger, trust, disgust, surprise, and anticipation. Such a vocabulary helps the analyst identify first those emotional expressions that are present in the diary and secondly improve the precision of sentiment analysis. On the other hand, the NRC Emotion Lexicon was created to be wide in order to support the accuracy and richness of the sentiment analysis findings. It, therefore, gives a detailed examination and covers the complex emotional subtleties described in Anne Frank's story. The well-defined architecture facilitates its smooth implementation into the sentiment analysis process where words are assigned to appropriate emotional categories. In the process of such alignment, researchers will identify recurrent topics depicted in the diary entries and their role in the emotional background of the diarist. Moreover, they can track the slight emotional signs and determine the nature of emotion. Her Diary's emotional depth is more understandable mostly through the NRC Emotion Lexicon, which provides insights that are necessary for doing an in-depth sentiment analysis.

Sentiment analysis of Anne Frank's carried out precisely due to use of these methods, which led to a deep exploration of the historical relevance and emotional depth that were in the text.

4.1.4 Pandas

A Python module which primarily works on structured data and especially is responsible for data processing is called Pandas which is known for its reliability. Data structures, Series and Data Frame, which have different purposes are the central element of the Python library.

- Series: It acts as pre-labelled an array intended to deal with one-dimensional data. It supports an array of data types such as strings, floats, and integers. Also, it allows indexing them by labelling and manipulating them in a fast and convenient way.
- Data Frame: The data frame is the two-dimensional structure which is easy to customize and looks like a spreadsheet or a table with rows and columns. When done, column (which is equal to the Series) can be saved and be easy to formulate and analyse the tabular data.

Integration with Data Visualization: Pandas is able to do a lot more than just generate charts and graphs, as these will end up being more intelligent data plots which can plot from Data Frame objects. It is in an effortless way linked to the data visualization libraries like Matplotlib and Seaborn. Panda is an important tool that can be used to organise, structured and to classify the text outtake from the diary of Anne Frank to understand the sentiments of the diary.

It makes these pre-processing and transformation steps straightforward; thus, users are enabled to efficiently dig into the data diary and get meaningful information analysis.

4.1.5 Streamlit

In this project, the Python framework of Streamlit—which is for building user-oriented web applications—acted as the base ingredient. Interactive platform for visualizing the sentiment analysis results of Anne Frank's diary database via Streamlit's user-friendly interface. Through the interactive graphs and charts, Streamlit, was able to tell the history of this diary's empirical expression, which created an interesting and engaging user interface. Apart from that, the Streamlit platform became more helpful with the complete integration of it with popular data visualization tools like Matplotlib, Seaborn, and Plotly. This connection allowed to include rich graphics as well, which eventually made it even more understandable for a reader to see a pattern of emotions and moods in a journal. In summary, Streamlit was very effective in allowing the environment for interpretation and analyses on Anne Frank's story to be interactive and to be maintained.

- Matplotlib, Seaborn, and Plotly:

Matplotlib, Seaborn and Plotly are the three popular Python libraries that are widely used for data visualization.

The Streamlit application's aesthetics and analytic depth were greatly increased by these libraries. They made it possible to come up with the most appealing and educational of plots, charts, and graphs by merging them with Streamlit natively.

4.1.5.1 Matplotlib:

The Streamlit application, that parses diary of Anne Frank heavily contributed by Matplotlib [8], which is one of the most well-known and highly appreciated open-source libraries with large set of plotting tools and flexibility, performs visualizations. By using Matplotlib's array of useful plotting functions, one can illustrate the rich variety of storylines with attractive plots and charts, and the information contained in the diary's text was depicted accordingly.

The ability to create line plot to track the sentiment over time is one of the best capabilities of Matplotlib. The reliance on these plots allowed the audience to trace the sentimental movement of the time during many Anne Frank diary entries experiments. It highlighted the changes in sentiments and themes of high importance. users can feel the emotions of the diary and see the contextual factors underlying the sentiment's variation with the matplotlib software. Matplotlib helped creating a bar chart which represented the frequency of the various emotions that were punctuated in the diary, these charts offered readers much details about the frequency of certain feelings she had, such as happiness, depression, and anger, which made them able to map the emotional and philosophical sides behind Anne Frank's story. Through Matplotlib, it is easy to design the visuals to suit sentiment analysis requirements and the nature of the diary's text using its wide range of plotting options, giving users important information about the emotional qualities of the text.

4.1.5.2 Seaborn

The Seaborn package, which is opened on Matplotlib, played an important role which made it possible for the Streamlit app to develop impressive graphics that are based on the diary of Anne Frank. Using a sophisticated plotting function of Seaborn and an intuitive interface, it is easy to create appealing charts that were simpler to understand. Through using Seaborn tools, it became the easiest to improve such elements as colour schemes and storylines, for which the diary’s emotional tone can be accurately visualized. Additionally, using Seaborn’s built in options for plotting statistical trade made it easy to identify the complex relationship found in the diary's textual content. Data visualization tools utilize Seaborn libraries to encourage developers to study the emotional route of journey recorded in the diary and reveal individual shifts and slight changes in emotional expression through differentiation. The Seaborn library allowed for a more satisfactory user experience and, in turn, reading maps, brought to life the emotional space of the book in a way easy to use and to appreciate.

4.1.5.3 Plotly

Plotly is particularly known for its adaptability and interactivity, consequently Plotly incorporated additional plot dynamic interactive features specialized for the the sentiment analysis of Anne Frank's diary and thus increased the usefulness of the Streamlit program. Using Plotly's complete variety of chart types, such as line plots, scatter plots, or 3D visualizations, was created for the better user experience and more realistic immersion in the final results from the sentiment analysis conducted on the diary text content. Through the implementation of interactive charts from Plotly into Streamlit technology, users has the opportunity to effectively modify and interact with the visualizations, which in turn enhanced connection with the tale of Anne Frank documented in her diary. The opportunity to zoom in on particular areas, hover the cursor over it to see more information, and switch between different viewpoints or angles gave a visual understanding of the evolving emotional terrain portrayed by the diary. Also, Plotly interactivity and animation features enabled to effectively illustrate complicated emotional change and sentiment fluctuations without overloading viewers with information. Finally, viewers could click on any plot data to see an expanding graph together with additional information and with the help of that able to gain more knowledge about Anne Frank's story.

4.2 Design Method

The goals and specifications for Anne Frank's diary sentiment analysis provided are thoroughly examined before designing, the aim of this first part is to be sure to set clear goals, know the project objectives as well as the user expectations.

4.2.1 Data Cleaning and Preprocessing

Before implementing sentiment analysis, the stage of data acquisition and preparation is crucial to make the data more accurate and coherent for the Anne Frank's Diary dataset. The following thorough procedure is carried out:

1. Attempts are made to find respective digital copies of Anne Frank's Diary from the sources considered as the most trustworthy. To get the text in a machine-thought-out format this could imply entering the official digital archives, online websites containing such information, or publishers approved to be doing that.
2. For analysis the existing digital copies of the Diary of Anne Frank, which can be easily downloaded in different electronic formats, i.e. PDF, DOCX or TXT are used for analysis. To know where to get the versions close to an authentic story, for example by looking at libraries, schools or sites for historical fiction are the places where these can be found.
3. These methods such as tokenization, lower casing and stop word removal are employed to transform the text into a suitable format for further checking.
4. The diary's numerical values, dates, and other non-linguistic content is processed and treated appropriately. Sentiment analysis processes have been created to ensure that only essential language content is analyzed.
5. To ensure that such algorithms are not compromised, special character and punctuation handling will involve spotting and dealing with a special symbol and punctuation sign which may be present in the text. In such a case, a user may remove these characters or change them by inserting a space or any other reasonable placeholder. The core of preprocessing tools includes cleaning and standardizing the text by removing unwanted characters, which ensures that sentiment analysis algorithms can focus on essential linguistic features instead of being distracted by noise. It turns out that this procedure is aimed at the assurance of data consistency and quality, which generates more accurate and credible sentiment analysis results.
6. Data integrity tests will be performed to ensure the accuracy of the processed text from the diary of Anne Frank. Besides these, verifications such as validating the data types, finding duplicates, recognizing punctuation marks; the wrong data types and confirmation of correct data are included. The accuracy of the data is protected, and errors such as duplicates, or incomplete data are minimized through the careful monitoring of these details. These approaches are critical because they not only make the sentiment analysis that will occur a refined one but also increase the reliability of the analysis and hence enable reliable conclusions to be drawn from the relevant historical text data.

These steps would be carefully followed to guarantee an adequate preparation for the “Anne Frank's Diary” database that would lead to rightful and reliable results.

4.2.2 Evaluation of Technologies and Methodologies

In a systematic approach, several dimensions, such as precision, scaling, computational efficiency, and ability to adapt to the language's evolution, are considered while selecting the sentiment approach for historical contexts. To successfully differentiate the sentiments in historical documents, the technology needs to have accuracy, capability and scalability which gives a way to work with big amounts of data effectively. Real-time analysis is related to computational efficiency and while the cultural context and cliche expressions appear to be a nuisance, they need to be handled with some flexibility. The systems that are integrated into your existing analysis can easily be employed through compatibility and integration with the current systems. The result of this comprehensive evaluation leads the way in determining the most suitable methodology. Based on such, the foundation for reliable sentiment analysis solution that is customized towards the historical text data is laid down.

4.2.3 Selection of Natural Language Processing (NLP) Techniques [10][5]

Considering the complexities and subtleties of historical language, design priorities the use of natural language processing techniques which is efficiently capable of handling contextual uncertainty and historical vernacular.

4.2.3.1 Tokenization

Tokenization is the primary step in Natural Language Processing (NLP) wherein text is divided into small units of tokens. In order to get word level analysis, word tokenization breaks text into unique words. This segmentation is a fundamental step in many of NLP tasks, like named entity recognition, sentiment analysis, and part-of-speech tagging. With NLP toolkit in Python, there is an NLTK package known as the word_tokenize() function, that helps in the breakdown of sample sentence shown into words or word tokens. Textual data breakdown to a word level provides for more deeper language processing and analysis.

4.2.3.2 Regular Expressions (Regex)

A regular expression, also known "regex", will be able to extract and identify its patterns from text data. As an instance, re module is used to locate dates in a text. This be achieved through the creation of a regex pattern that is able to locate the dates within the format "Monday, January 1, 2024". While for the purpose of this, compiling (re.compile()) functions will be used to specify the regex pattern explicitly for text related data, and findall() functions will be adopted to capture all the dates which were found , through regex code, will ensure precise location and processing of the date related text , which will assist later tasks of analysis.

4.2.3.3 Stopword Removal

Words like articles (“a," "an," "the"), conjunctions ("and," "or," "but"), and prepositions ("in," "at," "on"), which have just merely a little influence over the meaning of text in natural language processing, are known as "stop words". While stopwords may not be as significant, they are used for specific purposes and may be avoided to help the sentence to be more precise. The code utilizes STOPWORDS corpus offered by NLTK, which has predefined list of stopwords for English. word_tokenize() is then used to tokenize the text into individual words. Next all words occurring from the NLTK stopwords corpus are excluded and each tokenized word is compared with that compilation. In this particular step, the text data will be cleaned for tasks like sentiment analysis or topic modeling by removing all the stopwords from it. This feature refines the quality of text data by eliminating words which are frequently encountered yet lack any useful meaning. By virtue of the algorithms, these NLP approaches afford the code to easily pre-process the text, extract meaningful data, and make analysis of sentiment on text data.

4.2.4 Integration of Emotion Lexicon Method

4.2.4.1 The NRC Emotion Lexicon loading

Utilizing the NRC Word-Emotion lexicon consisting of words assigned into emotion categories and sentiments values will be the initial step for the sentimental analysis. Scan the lexicon file to extract the words associated with sentiment value and emotions. The information needs to be stored in a well-organized data structure to allow for easy retrieval at a time when other operations will be needed.

4.2.4.2 Extract Emotions

Iteratively look over each word in the pre-processed text, searching for matches in the NRC Emotion Lexicon. Figure out semantic values and emotions that are associated with each word in the dictionary. Add prominent emotions, e.g., "sadness," "anger," "fear," "joy," "trust," " disgust," "surprise," and "anticipation," related to every word that occurs throughout the text." In order to perform the depth of the emotional analysis, counts of every emotion need to be countered to analyse the intensity and frequency of words, within the text.

4.2.5 Sentiment Analysis Process

Sentiment analysis procedure involves preparing the text, retrieving the emotion as well as calculating the sentiment score, and then labelling sentiment based on the previously defined criteria.

4.2.5.1 Calculate Sentiment Score

Sentiment analysis involves several some important steps in determining sentiment scores. The predesigned lexicon is used to obtain both positive and negative emotions lists. The gather the number of each positive and negative emotion is done through the iteration of the emotions dictionary. The sum of counts of both positive and negative emotions is calculated separately to get the positive negative score. At the end, the negative score is deducted from the positive one and this give us a sentiment score. This method provides invaluable information on overall sentiment expression in the text, by computing sentiment analysis based on the frequency of both positive and negative emotions.

4.2.5.2 Label Sentiment

To label sentiment based on sentiment scores, Predefined threshold or criteria is set to classify this sentiment into different labels. Sentiment labels are assigned based on the scores after using conditional statements. As "Highly Positive"(score > 10) and "Positive"(score < 10) emotions are categorized as Positive sentiment. The zero score refers to "Neutral," the range between -10 and 0 is "Negative," while anything below -10 is considered "Highly Negative," giving a clear picture of polarity.

4.2.6 Date Pair Analysis

Date Pair Analysis works in two stages: first it detects date pairs of various types inside Anne Frank's diary (where date format is " Day, Month Date, Year") and second it extracts sentiment scores and emotions from the text between each pair of dates. This technique of classification allows the reader to follow not only temporal patterns but also see the diverse emotional moods portrayed whilst reading. The approach gets you a systematic way to understand how Anne feelings transform through time: happy, trustful times, to sad, unfair, and furious feelings. The application of sentiment analysis is used to determine emotional states which not only improves our understanding of Anne’s experiences but also takes us on a journey of her emotions as narrated in her diary.

Sentiment scores and sentiment collected between each date pair can also be saved in a CSV (Comma-Separated Values) file that may serve as a base for further the analysis and visualization. This file would have sections and count of words that relate to emotional states (e.g. Joy, Sadness, Anger, fear, Trust, disgust, surprise and anticipation), besides time scale that show their durations (e.g. starting point and end point). Moreover, the CSV would have two columns, "Sentiment Label" and "Sentiment Score" that would shed light on the emotional content of each time frame. Researchers can be conducted more in-depth investigation, see the chronological patterns, and have insight into Anne's emotional progress which can be made possible by this systematic style.

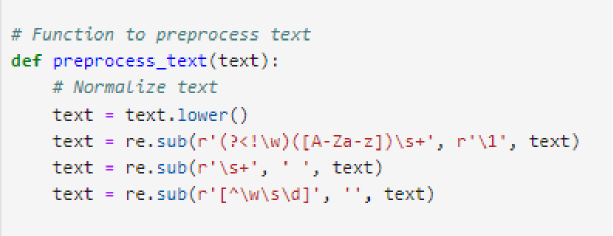
4.2.7 Visualization Using Streamlit

To speed up development of the visualization tool in Streamlit, ensure the data set is structured such that it contains sentiments scores, emotional categories, and metadata. Promptly after that, run pip install streamlit to have your Streamlit installed. Then, design a Streamlit application with options to use interactive features and plots by importing the needed libraries such as Matplotlib and Pandas. Include line and bar chart visualizations for displaying sentiment trend lines and emotional dynamics after the application has been loaded with a dataset. Create a unique and user-friendly design, with customized interface and visualizations that match to user's specific needs. Finally, perform testing of the program for precision and operation, debug and fix the bugs you find and finally add the program to a hosting system that will enable its usage. On a high level, the whole process of methodology is entirely composed of preparation of the data set, the Streamlit installation, the creation of the visualization, the User Interface design, the test, and the debugging, and the deployment.

5 Implementation

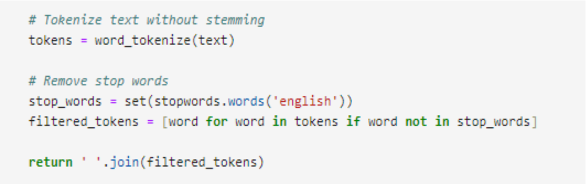
5.1 Preprocessing and Data Cleaning

- Text Extraction: for the purpose of obtaining the content of the PDF file (Anne Frank's Diary (PDF)), use the library of pdfplumber. It can be done by simply running the process on every page or over the whole PDF file.
- Text Preprocessing [5]: To acquire the scraped text and make it prepared for the analysis, one can use range of preprocessing methods as the following.
- Noise Removal: Eliminate the characters signs and other writings which do not have importance to the language character. Try to eliminate any irregularities that rise because of the digital objects extracted from the pdf documents such as line breaks, hyphens, and abnormal spaces.
- Removal of Non-Linguistic Content: Extract and remove unnecessary attributes that could be some information regarding numbers or meta data.



```
# Function to preprocess text
def preprocess_text(text):
    # Normalize text
    text = text.lower()
    text = re.sub(r'(?<!\w)([A-Za-z])\s+', r'\1', text)
    text = re.sub(r'\s+', ' ', text)
    text = re.sub(r'[^\w\s\d]', '', text)
```

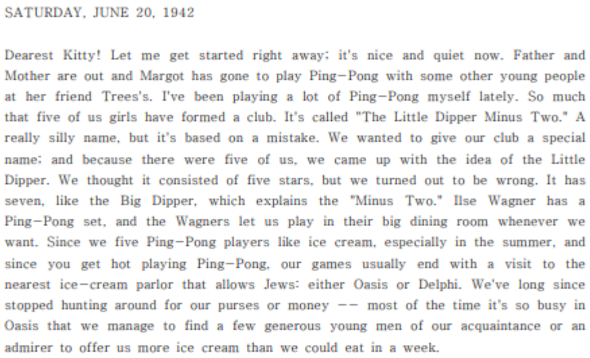
- Tokenization [5][10]: This step will help in follow up analysis by division of the text into smaller sections as either words or tokens. Divide the document into small text pieces using word tokenize() function NLTK or other tokenization methods.
- Eliminate Stop Words: Through preliminary Stage, attention should be paid to the main text by removing the most common stop words such as articles, prepositions, conjunctions. There you will use a custom stopwords set or NLTK's built-in standard set.



```
# Tokenize text without stemming
tokens = word_tokenize(text)

# Remove stop words
stop_words = set(stopwords.words('english'))
filtered_tokens = [word for word in tokens if word not in stop_words]

return ' '.join(filtered_tokens)
```

- • Data integrity: Checking viability of data type, revealing duplicates, and missing data are all included in the integrity of data. Which involves maintaining consistent and accurate types of data extraction and meta-data taken and filtering out duplicate entries to get rid of redundancy and correcting any miss or incomplete entries by substitution or elimination whatever is not perfect for additional analysis.
- Pre-processed Text Saving: Following the cleaning and pre-processing of textual data save it in a form which you can use to analyze, like .docx file (Anne-Frank-The-Diary-Of-A-Young-Girl.docx). Make sure that the file is not modified and in the desired format for analysis in the future.

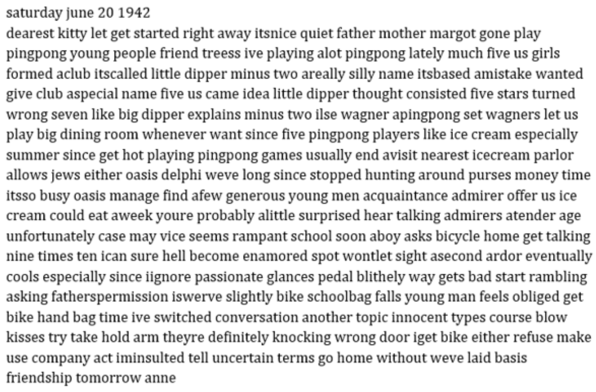
The textual data extracted from Anne Frank's diary by this process definitely holding all the qualities of being reliable, correct and best for analysis and interpretation.



SATURDAY, JUNE 20, 1942

Dearest Kitty! Let me get started right away; it's nice and quiet now. Father and Mother are out and Margot has gone to play Ping–Pong with some other young people at her friend Trees's. I've been playing a lot of Ping–Pong myself lately. So much that five of us girls have formed a club. It's called "The Little Dipper Minus Two." A really silly name, but it's based on a mistake. We wanted to give our club a special name; and because there were five of us, we came up with the idea of the Little Dipper. We thought it consisted of five stars, but we turned out to be wrong. It has seven, like the Big Dipper, which explains the "Minus Two." Ilse Wagner has a Ping–Pong set, and the Wagners let us play in their big dining room whenever we want. Since we five Ping–Pong players like ice cream, especially in the summer, and since you get hot playing Ping–Pong, our games usually end with a visit to the nearest ice–cream parlor that allows Jews: either Oasis or Delphi. We've long since stopped hunting around for our purses or money –– most of the time it's so busy in Oasis that we manage to find a few generous young men of our acquaintance or an admirer to offer us more ice cream than we could eat in a week.

Figure 2: Sample of Original data of Anne-Frank-The-Diary-Of-A-Young-Girl



saturday june 20 1942
dearest kitty let get started right away itsnice quiet father mother margot gone play pingpong young people friend treess ive playing alot pingpong lately much five us girls formed aclub itscalled little dipper minus two areally silly name itsbased amistake wanted give club aspecial name five us came idea little dipper thought consisted five stars turned wrong seven like big dipper explains minus two ilse wagner apingpong set wagners let us play big dining room whenever want since five pingpong players like ice cream especially summer since get hot playing pingpong games usually end avisit nearest icecream parlor allows jews either oasis delphi weve long since stopped hunting around purses money time itsso busy oasis manage find afew generous young men acquaintance admirer offer us ice cream could eat aweek youre probably alittle surprised hear talking admirers atender age unfortunately case may vice seems rampant school soon aboy asks bicycle home get talking nine times ten ican sure hell become enamored spot wontlet sight asecond ardor eventually cools especially since iignore passionate glances pedal blithely way gets bad start rambling asking fatherspermission iswerve slightly bike schoolbag falls young man feels obliged get bike hand bag time ive switched conversation another topic innocent types course blow kisses try take hold arm theyre definitely knocking wrong door iget bike either refuse make use company act iminsulted tell uncertain terms go home without weve laid basis friendship tomorrow anne

Figure 3: Sample of Filtered data of _Anne-Frank-The-Diary-Of-A-Young-Girl

5.2 Sentiment Analysis and Emotion Extraction

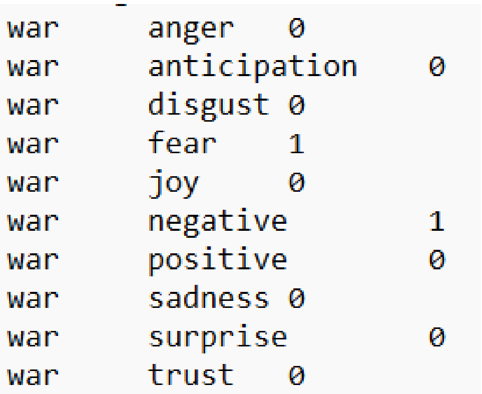
The aim is to analyze a text from Anne Frank's Diary for sentiment detection, and emotion extraction. The NRC Emotion Lexicon loading, sentiment score computation, sentiment labeling with predetermined thresholds, date pair identification within the diary text, emotion extraction from preprocessed text, and save these into a CSV file are vital tasks organization for processing. The addition of these chores to the design is intended to show the depth of the diary's emotional content making it available for further analysis and interpretation.

5.2.1 NRC Emotion Lexicon Loading

Bringing up the NRC Emotion Lexicon [9] is the aim of this step; it is a carefully selected lexicon that contains words that are labeled with emotions and have sentiment ratings associated with them.

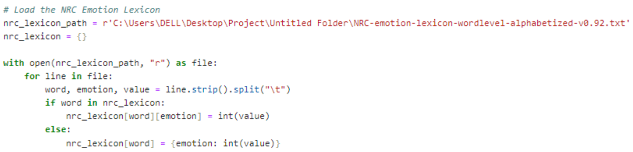
For example,

The word 'war' taken from the NRC Emotion Lexicon. While the term 'war' is typically considered a negative word with the emotion of fear therefore receiving a sentimental score of 1 for fear and 0 for the other emotions.



```
war	anger	0
war	anticipation	0
war	disgust	0
war	fear	1
war	joy	0
war	negative	1
war	positive	0
war	sadness	0
war	surprise	0
war	trust	0
```

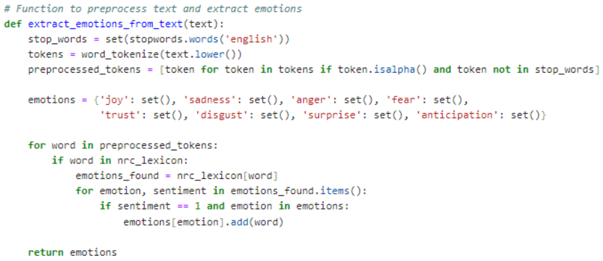
File input/output (I/O) procedures are used to access the NRC Emotion Lexicon, which is saved in a text file format, at the start of the implementation process. It is possible to extract individual words, the emotion categories they belong to, and the sentiment values associated with them by methodically parsing each line of the lexicon file. Next, the recovered information is arranged in a structured data format (a dictionary, for example) to make it easier to find and use later on in the emotion extraction procedure. Through a thorough integration of the NRC Emotion Lexicon into the sentiment analysis pipeline, this all-encompassing approach guarantees insightful information on the emotional content of the text data.



```python
# Load the NRC Emotion Lexicon
nrc_lexicon_path = r'C:\Users\DELL\Desktop\Project\Untitled Folder\NRC-emotion-lexicon-wordlevel-alphabetized-v0.92.txt'
nrc_lexicon = {}

with open(nrc_lexicon_path, "r") as file:
    for line in file:
        word, emotion, value = line.strip().split("\t")
        if word in nrc_lexicon:
            nrc_lexicon[word][emotion] = int(value)
        else:
            nrc_lexicon[word] = {emotion: int(value)}
```

5.2.2 Emotion Extraction

The code iteratively reads through the pre-processed text to identify emotions by the matches from NRC Emotion Lexicon. The code checks out if the word belongs to the NRC Emotion Lexicon. As soon as a match is found, the standard term is connected to the dictionary's description of that group of emotions. Then, these emotions are tied for each category, which provides a deep understanding of the emotional content of the text. The system uses an iterative approach to exactly understand and extract the emotions' and in this way it ensures that the base of further process of sentiment analysis and interpretation have a solid foundation.

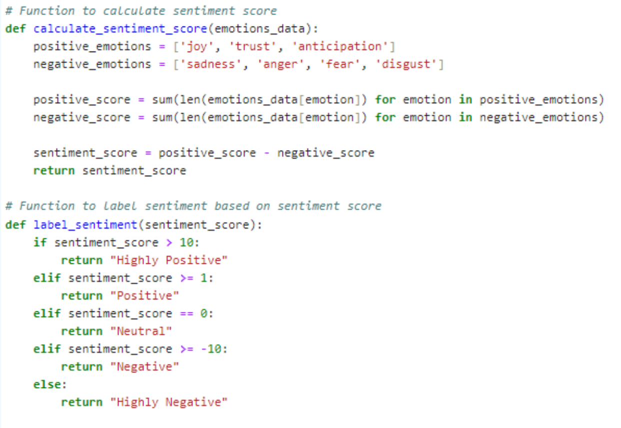


```python
# Function to preprocess text and extract emotions
def extract_emotions_from_text(text):
    stop_words = set(stopwords.words('english'))
    tokens = word_tokenize(text.lower())
    preprocessed_tokens = [token for token in tokens if token.isalpha() and token not in stop_words]

    emotions = {'joy': set(), 'sadness': set(), 'anger': set(), 'fear': set(),
                'trust': set(), 'disgust': set(), 'surprise': set(), 'anticipation': set()}

    for word in preprocessed_tokens:
        if word in nrc_lexicon:
            emotions_found = nrc_lexicon[word]
            for emotion, sentiment in emotions_found.items():
                if sentiment == 1 and emotion in emotions:
                    emotions[emotion].add(word)

    return emotions
```

5.2.3 Sentiment Score Calculation and Labeling

In this step, the sentiment score is calculated by adding the scores of positive emotions and negative emotions which were extracted from the text data and presenting the difference of the two in a comparative way. The approach to the implementation includes integration of both positive feelings including joy, trust, and excitement for instance, while also taking into account the negative emotions like sadness, wrath, fear, and disgust among others. This method sums up positive and negative scores contributed by the numbers of sentiments (either positive or negative). Then, total sentiment of the article is calculated via subtracting negative score from positive value. With the help of this technique, the text readers can detect sentiments that lay within the text; hence discovering the main emotional theme serves as a point of entry for other things such as analysis and interpretation.

Sentiments are detected and subsequently, they are grouped into distinct sets which include highly positive, positive, neutral, negative and highly negative based on the defined threshold values to allow for the classification of grouped sentiments. This method is accomplished by establishing the thresholds or bandwidths, within which the sentiment scores would fall in each of the categories of sentiments. The system next determines a sentiment label for each sentiment score using the predefined thresholds. The polarity of the text is annotated based on its sentiment score which is then matched with a pre-defined criterion to get the appropriate sentiment label. As a result, sentiment pattern analysis will become coherent and the emotional context behind the text data will be clear to reveal.



```python
# Function to calculate sentiment score
def calculate_sentiment_score(emotions_data):
    positive_emotions = ['joy', 'trust', 'anticipation']
    negative_emotions = ['sadness', 'anger', 'fear', 'disgust']

    positive_score = sum(len(emotions_data[emotion]) for emotion in positive_emotions)
    negative_score = sum(len(emotions_data[emotion]) for emotion in negative_emotions)

    sentiment_score = positive_score - negative_score
    return sentiment_score

# Function to label sentiment based on sentiment score
def label_sentiment(sentiment_score):
    if sentiment_score > 10:
        return "Highly Positive"
    elif sentiment_score >= 1:
        return "Positive"
    elif sentiment_score == 0:
        return "Neutral"
    elif sentiment_score >= -10:
        return "Negative"
    else:
        return "Highly Negative"
```

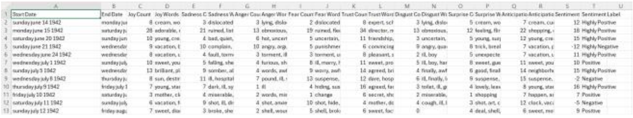
5.2.4 Date Pair Analysis

Sentiment and emotion analysis like this can be enabled through segmenting the diary text. Therefore, the purpose of this section is to find date pairings (the date entries). Finding the date pattern using regular expression is one of the implementation processes. Next, the code runs through these pairs several times to carry out sentiment scores and emotion extraction in each pair. Temporal analysis is made possible through data storage and identification of start and end dates. This method discovers trends in emotional trends across time.

5.2.5 CSV Data Export

To enable further investigation or visualization, the extracted emotions and sentiment scores between date pairs must be saved in a csv file (emotions_between_dates_with_sentiment.csv). The writing of data into CSV format is done by I/O Operations, including file writing. In order to help to understand and manage the stored data well, it is very important to ensure that all columns should have their proper headers.

The above comprehensive process makes sure that the design produces exact sentiment analysis and emotion extraction from Anne Frank's diary, which, in turn, assists in studying change in emotional patterns over time.



Start Date	End Date	Joy Count	Joy Words	Sadness C	Sadness W	Anger Cou	Anger Wor	Fear Count	Fear Word	Trust Coun	Trust Word	Disgust Co	Disgust Wo	Surprise C	Surprise W	Anticipatio	Anticipatio	Sentiment	Sentiment Label
sunday june 14 1942	monday jun	8	cream, wo	3	dislocated	3	lying, dislo	2	dislocated	8	expert, sch	3	lying, dislo	5	cream, wo	7	cream, cu	12	Highly Positive
monday june 15 1942	saturday ju	28	adorable, r	21	ruined, list	13	obnoxious,	15	ruined, fle	34	director, m	13	obnoxious,	12	feeling, flir	22	shopping,	18	Highly Positive
saturday june 20 1942	sunday jun	10	young, cre	4	bad, quiet,	6	hot, uncert	5	uncertain,	11	friendship,	3	uncertain,	5	young, sur	12	young, cre	15	Highly Positive
sunday june 21 1942	wednesday	9	vacation, t	10	complain,	10	angry, arg	5	punishmer	6	convincing	9	angry, qua	8	trick, breal	7	vacation, p	-12	Highly Negative
wednesday june 24 1942	wednesday	8	vacation, u	4	fault, torm	3	torment, ill	3	torment, u	8	pleasant, s	2	ill, boy	5	unexpecte	7	vacation, u	11	Highly Positive
wednesday july 1 1942	sunday july	10	sweet, you	5	falling, she	4	furious, sh	8	ill, marry, f	11	sweet, pro	5	ill, boy, har	8	sweet, gue	11	sweet, you	10	Positive
sunday july 5 1942	wednesday	13	brilliant, pl	9	somber, al	4	words, awf	9	worry, awf	14	agreed, bri	4	finally, awf	6	good, final	14	neighborh	15	Highly Positive
wednesday july 8 1942	thursday ju	8	sun, destr	11	ill, hospital	7	pound, ill,	13	suspense,	12	dare, hosp	6	ill, finally,	9	suspense,	15	suspense,	-2	Negative
thursday july 9 1942	friday july 1	7	young, star	7	dark, ill, sy	1	ill	4	hiding, sus	16	agreed, fa	3	toilet, ill, g	4	lovely, lea	8	young, sta	16	Highly Positive
friday july 10 1942	saturday ju	3	mother, ch	4	miserable,	2	words, mi	1	change	6	secret, sh	2	miserable,	1	shopping	7	happen, a	7	Positive
saturday july 11 1942	sunday july	6	vacation, f	9	shot, ill, di	4	shot, anxie	10	shot, hide,	4	mother, do	4	cough, ill, l	3	shot, art, c	12	clock, vaca	-5	Negative
sunday july 12 1942	friday augu	7	sweet, dis	3	broke, she	2	shell, wour	5	shell, brok	6	sweet, fac	0		4	deal, shell,	6	sweet, mo	9	Positive

Figure 4: Sample of emotions_between_dates_with_sentiment.csv.

5.3 Visualization in Streamlit

The sentiment visualization technique involves such visual components as a word cloud, histogram, pie chart and line chart that show sentiment score distribution and word usage. The sentiment distribution, crucial connections between the emotion-words found within the text, and emotional shift throughout the narrative are depicted in these charts. Each visualization is created via Python code (app.py), which is needed to make the application interactive and user-friendly by using Streamlit's functions and visualization libraries. The app can be released on platforms after the test version, which ensures the functionality has been completed.

In the next part, a more detailed assessment of the visualization elements, pivoting around their design, implementing, and contribution clauses will be performed.

6 Visualization

Analysing Sentimental Trends of Anne Frank's Diary

Streamlit a Python library capable of developing interactive web applications is used in the visualization stage to extract emotive themes in Anne Frank's diary. Streamlit is the suitable tool for development of data applications, and it makes it easy for you to combine interface, data processing, and visualization in a single Python script.

6.1 Understanding Streamlit

- Streamlit [6] provides a user-friendly interface whereby you can directly create web apps from Python scripts.
- A programmer can instruct the application about its structure, behaviour, and layout using the widely adopted syntax of Python.
- The standard command is irrelevant as Streamlit performs the operations of data loading, visualizing, and user interface element handling automatically.
- It is quite easy to use interactive widgets such as buttons, sliders and dropdown menus to increase user interaction and data exploration.

6.2 Downloading Streamlit

The installation of Streamlit can be done through the Python package manager i.e. "pip". With the help of your terminal or command prompt, type the following command to install it.

- pip install streamlit

- Streamlit's Use in Visual Studio:
  To use Streamlit in Visual Studio via Python, you must install Python on your computing device first. Once Python being installed, visual studio can start and a new Python project needs to be opened then follow the prompts to set up Streamlit. After that, get started with shaping the code of your Streamlit application by creating a new Python file (.py file).

6.3 Sentimental Trends Visualization Requirements

Before starting the visualization process make sure that the following requirements are satisfied.

- Anne Frank's Diary dataset: Extract the data from Anne Frank's Diary in the format like CSV files.
- Essential libraries: Install necessary external libraries like Pandas, Matplotlib, Seaborn, Plotly and word cloud for better and interactive visualization
- Streamlit: Also make sure that streamlit has been set up properly.

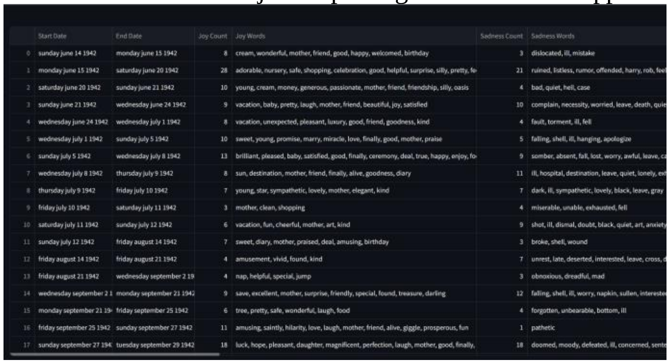
6.4 Visualization Method

Streamlet, this is a web–based application that aids the users to easy analysis of emotions within the diary of Anne Frank. The application will show the emotions number, the distribution of sentiment based on the score, frequency rate of words, along with others illustrated through different chart types of namely word clouds, pie charts, histograms, and line charts. Besides this, the users will be given the opportunity to engage with those visualizations so that they can get an insight into the distribution of the sentiments and patterns of emotions.

6.5 Graphs and Visualizations:

6.5.1 Extracted Dataset in CSV format:

The extracted CSV file just importing in the streamlit app:



	Start Date	End Date	Joy Count	Joy Words	Sadness Count	Sadness Words
0	sunday june 14 1942	monday june 15 1942	8	cream, wonderful, mother, friend, good, happy, welcomed, birthday	3	dislocated, ill, mistake
1	monday june 15 1942	saturday june 20 1942	28	adorable, nursery, safe, shopping, celebration, good, helpful, surprise, silly, pretty, fe	21	ruined, listless, rumor, offended, harry, rob, fee
2	saturday june 20 1942	sunday june 21 1942	10	young, cream, money, generous, passionate, mother, friend, friendship, silly, oasis	4	bad, quiet, hell, case
3	sunday june 21 1942	wednesday june 24 1942	9	vacation, baby, pretty, laugh, mother, friend, beautiful, joy, satisfied	10	complain, necessity, worried, leave, death, quie
4	wednesday june 24 1942	wednesday july 1 1942	8	vacation, unexpected, pleasant, luxury, good, friend, goodness, kind	4	fault, torment, ill, fell
5	wednesday july 1 1942	sunday july 5 1942	10	sweet, young, promise, marry, miracle, love, finally, good, mother, praise	5	falling, shell, ill, hanging, apologize
6	sunday july 5 1942	wednesday july 8 1942	13	brilliant, pleased, baby, satisfied, good, finally, ceremony, deal, true, happy, enjoy, fo	9	somber, absent, fall, lost, worry, awful, leave, ca
7	wednesday july 8 1942	thursday july 9 1942	8	sun, destination, mother, friend, finally, alive, goodness, diary	11	ill, hospital, destination, leave, quiet, lonely, ex
8	thursday july 9 1942	friday july 10 1942	7	young, star, sympathetic, lovely, mother, elegant, kind	7	dark, ill, sympathetic, lovely, black, leave, gray
9	friday july 10 1942	saturday july 11 1942	3	mother, clean, shopping	4	miserable, unable, exhausted, fell
10	saturday july 11 1942	sunday july 12 1942	6	vacation, fun, cheerful, mother, art, kind	9	shot, ill, dismal, doubt, black, quiet, art, anxiety
11	sunday july 12 1942	friday august 14 1942	7	sweet, diary, mother, praised, deal, amusing, birthday	3	broke, shell, wound
12	friday august 14 1942	friday august 21 1942	4	amusement, vivid, found, kind	7	unrest, late, deserted, interested, leave, cross, d
13	friday august 21 1942	wednesday september 2 19	4	nap, helpful, special, jump	3	obnoxious, dreadful, mad
14	wednesday september 2 1	monday september 21 1942	9	save, excellent, mother, surprise, friendly, special, found, treasure, darling	12	falling, shell, ill, worry, napkin, sullen, intereste
15	monday september 21 19	friday september 25 1942	6	tree, pretty, safe, wonderful, laugh, food	4	forgotten, unbearable, bottom, ill
16	friday september 25 1942	sunday september 27 1942	11	amusing, saintly, hilarity, love, laugh, mother, friend, alive, giggle, prosperous, fun	1	pathetic
17	sunday september 27 194	tuesday september 29 1942	18	luck, hope, pleasant, daughter, magnificent, perfection, laugh, mother, good, finally,	18	doomed, moody, defeated, ill, concerned, sente

Figure 5: Extracted CSV file in Streamlit App

6.5.2 Graph 1: Emotion Counts over Time

The counts of different emotions throughout time in Anne Frank's diary are shown in Figure 6: Emotionality is especially important over time. The plot is created using Python libraries Matplotlib and Streamlit. A multiselect widget enables people to select only certain emotions they want to analyse in detail. Each colour of the scale resonates each feeling represented in the graph. The graph displays the following emotions: surprise, joy, anger, sadness, fear, trust, disgust, anticipation. x-axis shows the start date, and the y-axis present the number of each emotion. The graph presents the frequency of each emotion through time in the most appropriate and directly visualizable way. Readers gains a better understanding of the emotional impressions as wrote down by Anne in her diary through observing the repeated patterns and variations in her emotions. The graph will include user-selected emotions. Consequently, the graph is adjusted to highlight the selected emotions and also present them in a clear pictorial manner through showing number of occurrences over time. In case, no emotion is chosen from the multi-widget option then a notice would be inferred which asks the user to select emotion from the multiselect widget to visualize. In general, the graph of the frequency of emotions in Anne Frank's diary presents a detailed and dynamic view of the emotional topics through the text, and it invites the reader to contemplate on the emotional significance of the text.
Selected Emotions for Counts over time: Joy and Anger

Insights:

Joy and Anger are the opposite extremes of the two selected emotions for the analysis. The timeline of these emotions' counts can be seen on the graph. Graph shows off that at first, number of joy emotions was more than anger in the beginning. However, as the calendar ticks by in the end the amount of Anger is greater than the amount of the Joy. The fact that this discovery suggests that Anne Frank's diary's emotional value altered from time to time.
The diary notes could have initially recorded some happier events or emotions. However, as days were passing there was noticeable shift towards those instances in the diary where the readers can find more display of wrath or anger. This perception highlights the dual nature aspect of Anne Frank's emotional life portrayed in the diary and demonstrates a shift from a dominant number of happy emotions to a stronger level of taste to anger.

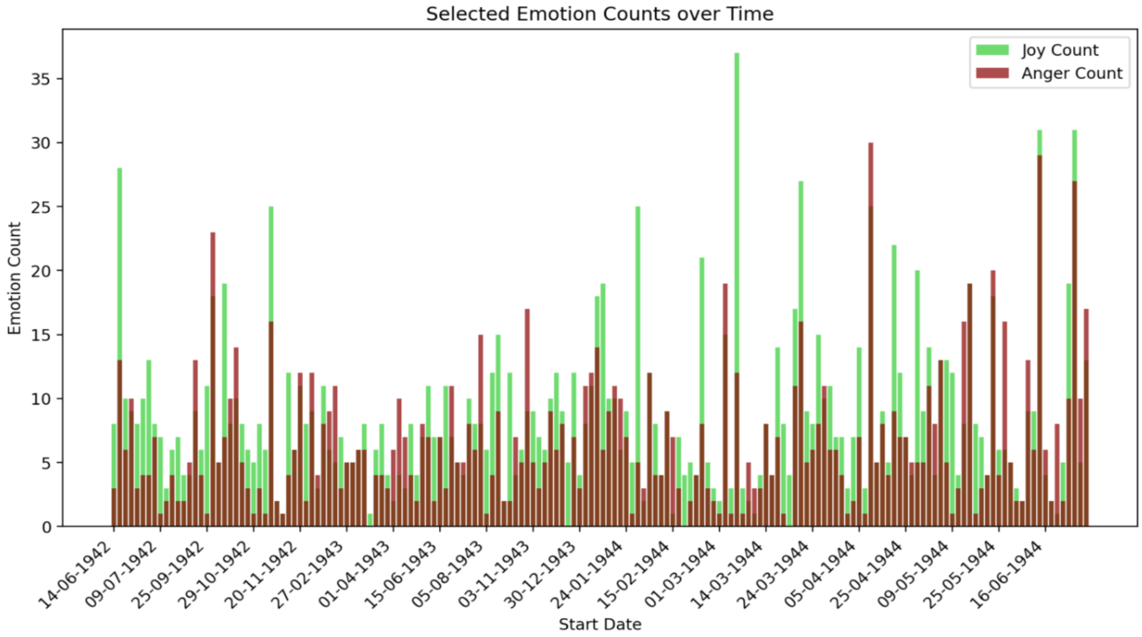


Figure 6: Emotion Counts over Time

6.5.3 Graph 2: Emotion Distribution Pie Chart

As shown in the Figure 7: The Emotion Distribution Pie Chart. The pie charts are used to show emotions distribution among the data set. The function that is written in Python for the pie chart adds the totals for each emotion and employs Matplotlib and Streamlit to make the diagram. The data chart demonstrates percentages of each emotion types. The pie chart above shows the percent distribution of each emotion. In this, the name of each emotion are in a list along with the proportion of the overall count which is adjacent with each label. The chart colours function as a useful tool by making it easier to differentiate the moods and contrast them. In general, the pie chart will provide the audience with a visual aid for the understanding and the evaluation of the emotional content.

Insights:

The journal, which Anne Frank penned while at the Nazi hideaway, went on to reveal the full range of emotions addressed by the young girl. The recurrence of a feeling of fear (13.3%) and sadness (12.8%) reflects the problems and relatively uncertain situation she has been experiencing. On the other hand, the narration might also speak a tone of joy (14.11%) and anticipation (15.12%) which are the words that indicate an optimist and some positive outlook. However, other than that, considered the type of trust (17.7%) which shows that people require each other for emotional support or happiness. Through the change of mood from 10.7% (which is angry), 8.2% (which is disgusted), and 8% (which is really surprised), you can see the depths of the emotions of a Holocaust survivor while trying to survive the harsh conditions of their struggle for survival. Moreover, Annes emotional journey shows a spirit of courage that is capable of boldly walking the hardest way under these emotions which cover the battle, grief and endurance.

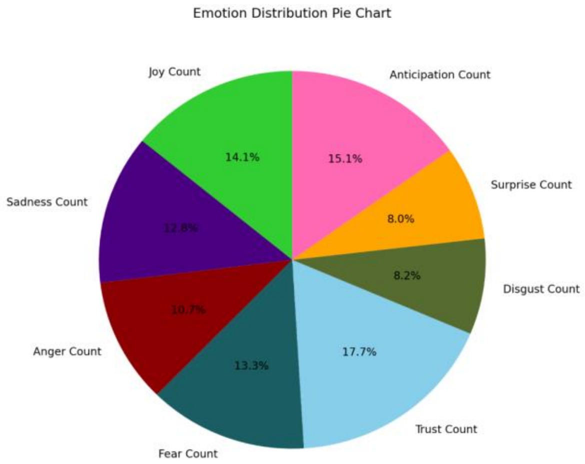


Figure 7: Emotion Distribution Pie Chart

6.5.4 Graph 3: Sentiment Score Distribution Histogram

The sentiment score distribution obtained from the examination of Anne Frank's diary is represented graphically in the "Figure 8: "Emotion Scoring Distribution - Histogram" graph. The y-axis represents the frequency of every sentiment score slice, and whereas x-axis is the one that shows the sentiment scores, which are separated into bins. each containing one of the assigned sentiments The colour-variety in the custom colormap is used for denoting

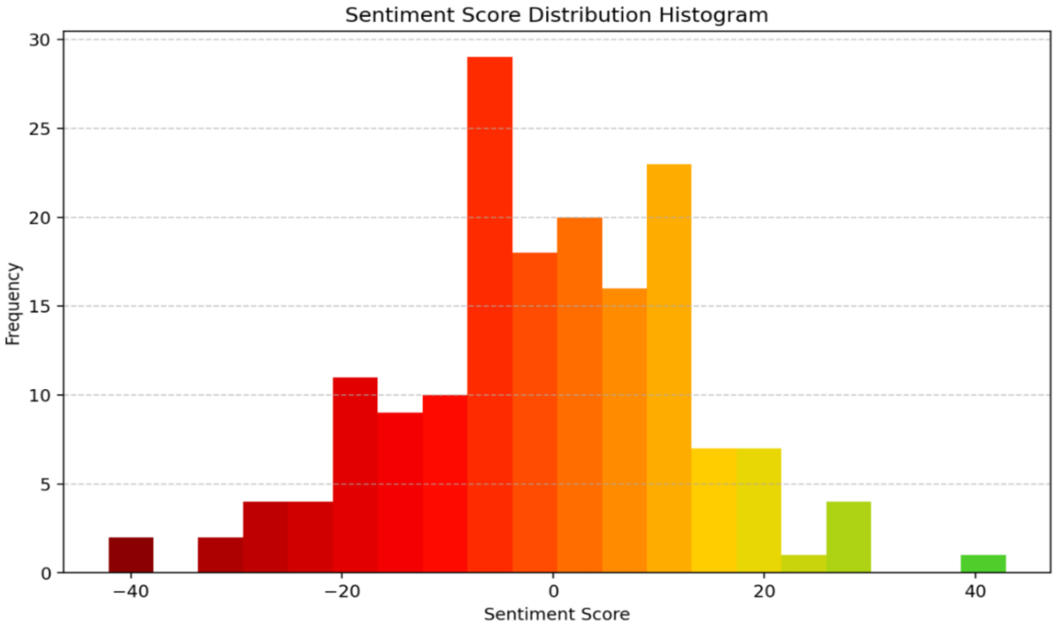


Figure 8: Sentiment Score Distribution Histogram

the range of sentiment score from negative to positive and is applied to bars of the histogram.

This colour range begins from dark red through red to green and ending at the dark green which shows the spectrum of emotions. This graph provides an overview of the distribution of the sentiment scores throughout the diary, and thus, offers a framework to capture the major sentiments and their frequencies recorded by Anne Frank in her story.

Insights:

There are various things to note about the histogram that represents Anne Frank's feelings:

1. Correspondingly, the majority of distribution for the data consists of the values to the left of zero, and so the data is likely dominated by a lot of these negative sentiment scores.
2. The most bars place between scores 0 and –20, which all are the negative range representing the fact that the sentiment score is likely to be frequently recorded in this area in the diary.
3. There is a high concentration of ratings around the neutral zone which suggests that numerous emotion scores fall into this category.
4. However, there is a noticeable decrease in the intensity of positivity (there exist fewer bars at the extreme positive end of the scale) and at the same time the frequency of highly positive sentiments is lower than the highly negative ones.

This histogram clearly shows that, a graph can be graphically shown where the position of maximum is negative, the diary of Ann Frank registers a large volume of negativity. Considering the broader picture, these observations may become beneficial to the readers after sharing them in regard to Anne Frank's story.

6.5.5 Graph 4: Correlation Heatmap

Based on the count column, the correlation matrix is formed. The correlation matrix data is used to construct the heatmap (Figure9: Correlation Heatmap) - A visualization method where each cell represents a degree of correlation coefficient between two variables and the colour saturation explains the correlation level. The correlation values are depicted using the 'RdBu' colorscale, which shows either the shades in between depending on the strength of the correlation; red indicates a positive correlation, while blue symbolizes a negative one. Updated layout is added and a title is included after the heatmap is plotted by Plotly's Heatmap function. For illustration purposes, heatmap is created using Streamlit's plotly_chart method.

Insights:

A correlation heatmap, which is a graphical representation of the correlation matrix between a set of variables, is shown in the image. In this case, it seems that the variables stand for the relative intensities of the different emotions ranging from "Joy" to "Sadness," "Anger," "Fear," "Trust," "Disgust," "Surprise" and "Anticipation."

This is a summary of how to interpret this heatmap:

1. Colours: Colour scale of the heat map covers from red to blue, and it is shown by a colour bar on the right side of the screen. For instance, the red shows the correlations become higher positive, the blue shows the negative, and the colours in between show different values ranging from positive to negative.
2. Correlation Values: The heatmap would highlight the emotions on the X- (horizontal) and Y- (vertical) axes where each square (or cell) represents the correlation

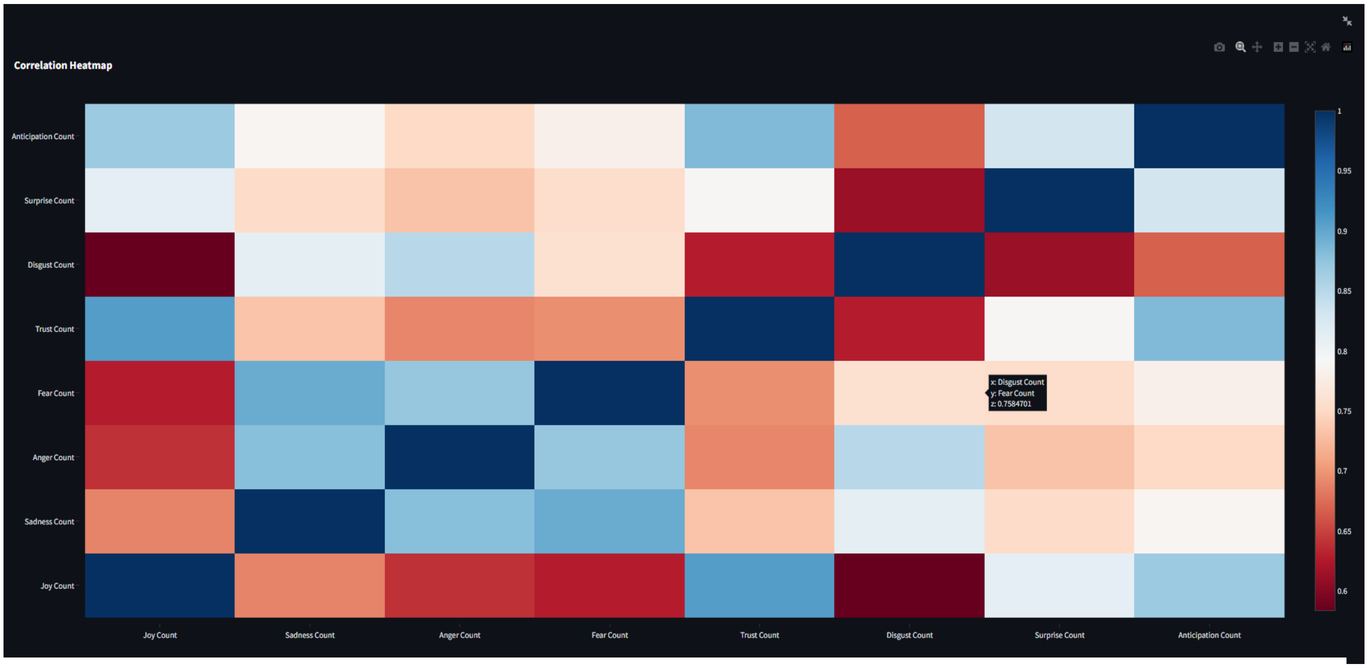


Figure 9: Correlation Heatmap

between them. A perfect positive correlation shown by a correlation value of 1 for an example and a perfect negative correlation shown by a correlation value of -1 in example, and no correlation stands for a correlation value of 0.

3. Interpretation: With regard to this, the cell marked -0.611338, which is red, is present at the intersection of the X-axis labelled "Disgust Count" and the Y-axis labelled "Surprise Count." As the correlation between itself and any particular variable is always equal to 1 (that is perfect positive correlation), the heat map is typically symmetric to its diagonal line.

When analysing data, the heatmap is more effective in directly displaying relationships between variables so that patterns and connections are easily grasped.

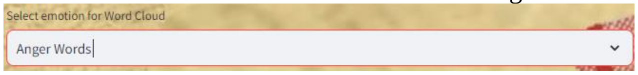
6.5.6 Graph 5: Word Clouds for Emotions

A word cloud displayed in Figure 10: Selected Emotion, Word Cloud created using the emotion from the dataset of Anne Frank's diary. A user can select the emotion for which they intend to create the word cloud by using a dropdown menu in Streamlit app. When an emotion is chosen from the database, the code connects all the words linked with this feeling. Finally, it generates the word cloud representation by employing the `Word Cloud` class of the `word cloud` library. Every word in word cloud is coloured with a predetermined colour scheme that is based on the given emotion. The app interface displays the created word cloud image. This allows the viewers not only to focus on the emotional themes and sentiments described by Miss Anne Frank in her diary, but they can also have a visual reflective which includes the main expressions associated with given emotional details.

Insights:

A word cloud which is a graphical way of a representation of text data and each size word is an indication of its correlation or its frequency. This particular word cloud, which belongs to "Graph 5: "Word Clouds for Emotions," and the theme "Anger Words" is one of the many parts of the collection or the presentation. With the red-coloured word cloud, the text illustrates the anger element. Words of all sizes are visible, including the following: "death," "fear," "lonely," "rage," "bad," "awful," "fight,", "hurt," "crazy," "upset", "hate", and "scream" and the others. In the light of the data used in this word cloud, the larger the word, the stronger it triggers the sentiment of anger. This visualization device is frequently used to quickly focus on the most important parts in the document, it may be also used to group words according to distinct themes or feelings categories by using colour. In this case, since red is often associated with intense emotions, especially anger or fury, this use of red on words that communicate wrath is in accordance with standard cultural nuances.

- Selected emotion for Word Cloud: Anger Words


Select emotion for Word Cloud

Anger Words

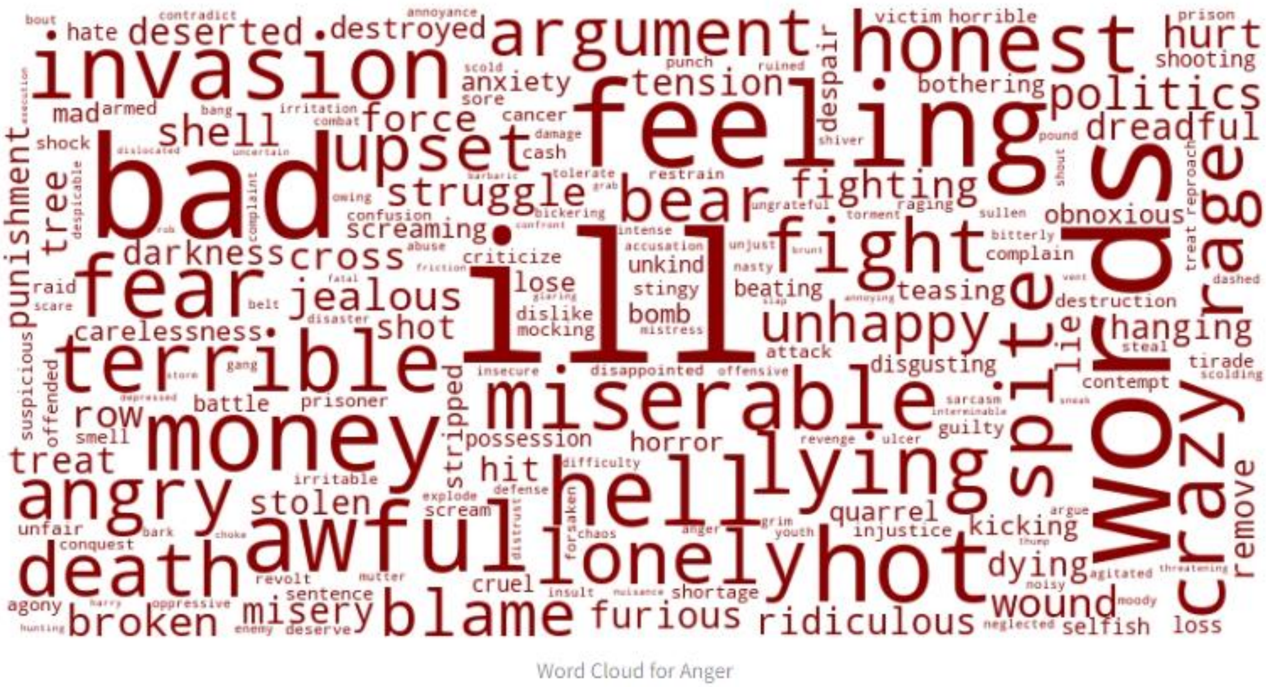


Figure 10: Word Cloud of selected emotion.

6.5.7 Graph 6: Distribution of Sentiment Labels with Extreme Scores

A visual representation of the sentiment label distribution with extreme scores in the diary dataset of Anne Frank, displayed in Figure 11: Sentiment Labels Distribution as a Bar Chart with the Extreme Scores. In the process of preprocessing data for graphing, code define the sentiment categories, i.e., "Neutral," "Highly Negative," "Negative," and "Highly Positive." The data is subdivided iteratively by going over each sentiment category. The sentiment number for each subset along with the extreme score (maximum or minimum number for positive or negative sentiments) is calculated and stored in a dictionary named `plot_data'. Subsequently, with the use of Matplotlib library and the dictionary, a bar chart is generated where every sentiment label is represented by a bar. The top is marked with an extreme score. The height of each bar indicates the number of sentiment labels. Each sentiment type comes with its own representative color enabling better understanding of the scheme. The last step is the `st.pyplot()` function that is used to show the bar chart inside the Streamlit app. Exhibiting the high scores that go along with certain sentiments in this visualization gives readers information on how the sentiment labels are distributed inside Anne Frank's diary.

Insights:

A bar chart labelled "Graph 6: Extreme score distribution in relations with the Sentiment Labels" is shown in the picture. The diagram below demonstrates the combined resultant figure of various sentiment labels used alongside their respective counts and extreme scores. Five categories are displayed on the graph's "Sentiment Label" x-axis.

Five categories are displayed on the graph's "Sentiment Label" x-axis: Highly Positive, Positive, Negative, Highly Negative or Neutral. This "Count" reveals chosen sentiment label values on the y-axis (0-50). The chart features color-coded bars with numerical values representing the number of occurrences above them and an "Extreme" score: The extreme score of the green bar behind the “Highly Positive” box counts 45 and the extreme score of the "Positive" category is lighter with 10. A pink bar shows extreme score being 10 is, therefore, showing the "Negative" category. The "Highly Negative" category has a bar that is red and in it you can see -40. A grey bar having 0 extreme score and maximum points of 0 denotes the "Neutral" category. This information may suggest that the sentiment polarity was either "Highly Positive" or "Positive" in the vast majority of cases, "Positive" being slightly more common. The results were mostly "Highly negative" and "Negative" sentiments, but very less "Neutral" sentiment was observed.

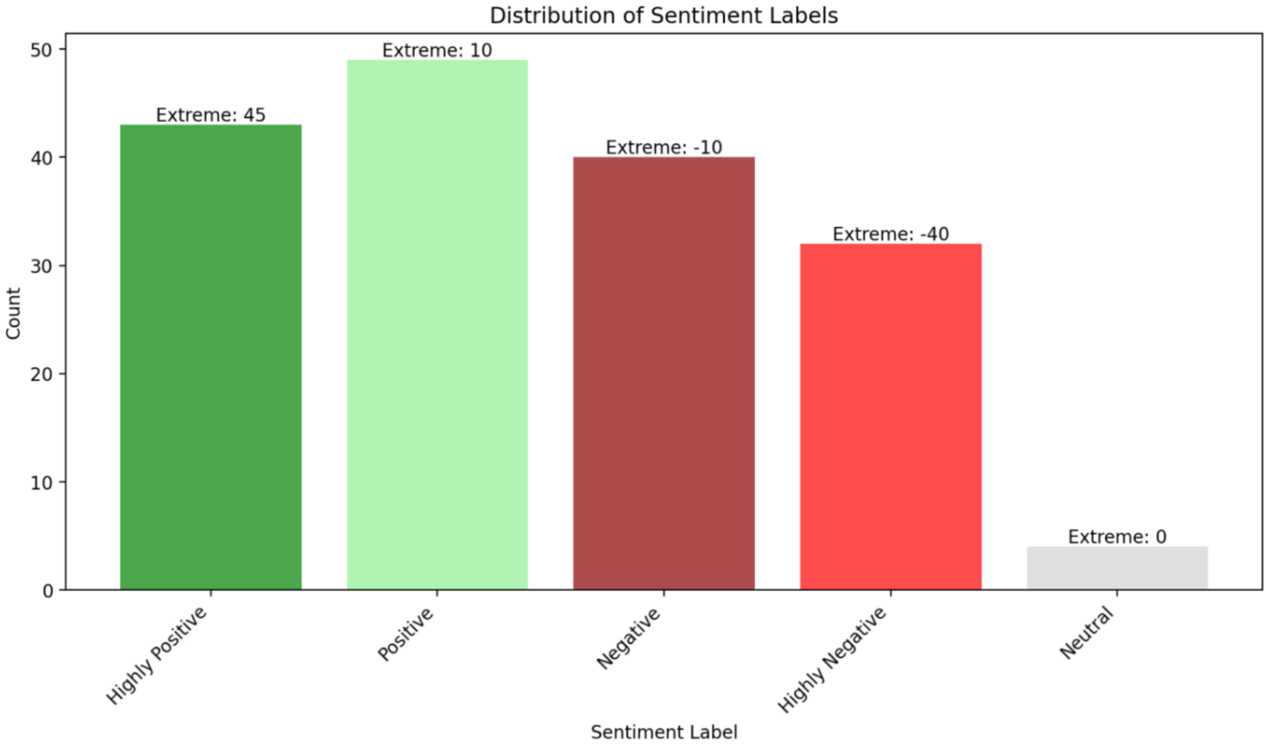


Figure 11: Distribution of Sentiment Labels with Extreme Scores

6.5.8 Graph 7: Emotion Transition Diagram

A representation of the sentiments transitioning between various states found in Anne Frank's diary dataset is called a Sentiment Transition Diagram (Figure12:"Sentiment Transition Diagram"). The NetworkX library's DiGraph() function is next employed to generate a directed graph object G. The method then goes through the pairs of successive sentiments (the present sentiment and the succeeding sentiment), by iterating through each row of the dataset. The transition from a certain emotion to the next one is linked by an edge that is then entered in graph G for each pair. After the graph is built, the code uses the spring layout technique to calculate the node positions and to store them in the pos variable. The placement of the nodes is obtained based on this layout technique to enhance the neatness and decrease the overlapping among the edges. Then, using the subplots() function in Matplotlib, a new figure is produced along with its own axis which are used to show the data. The size of the figure can be specified. The very draw() method of NetworkX is used in plotting the graph. Using st.pyplot(), the graph can be finally shown in the Streamlit app. By this display, the users could get an idea about the sequence and rhythm of the sentiments.

Insights:

A "Sentiment Transition diagram", kind of directed graph, which exhibits how one sentimental state might lead into another, which is shown in the figure below. Five primary sentimental states are shown as nodes in this diagram:

Five primary Sentimental states are shown as nodes in this diagram: 1. Highly Negative 2. Negative 3. Neutral 4. Positive 5. Highly Positive

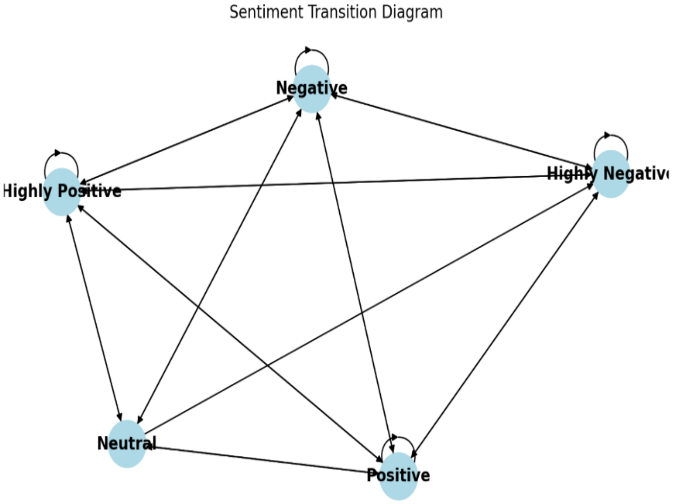


Figure 12: Sentiment Transition Diagram

Arrows are drawn between each node, representing the possibility of pattern shifts in mental states. The return of the arrow to the same node is a fact which demonstrates that emotional state could persist for a long time. This is a thorough description of the diagram: This is a thorough description of the diagram: -

“High Negative” can remain “High Negative” or can change to “Negative.” - The state of "Negative" may change to "Neutral" or "Positive" or go back to the state of "Negative." - "Positive" can turn to "Neutral, "Highly Positive" or continue "Positive." It can be "Highly Positive" that is changed over to "Positive" or will remain the same "Highly Positive." Neutral also can transform into any one of the four other emotional state. But neutral cannot go back to neutral state. Therefore, it is only demonstrated here that transition of these events is possible - not the probability or frequency. It is a visual illustration which can be utilized easily by people to understand and predict the emotional trajectory in user experience research, psychological studies or sentiment analysis algorithms.

6.5.9 Graph 8: Animated Time Series

The Animated Time Series visualization displayed in Figure 13: Animated Time Series is a way to show how the emotion in the diary of Anne Frank shifted overtime. Relevant colours are set for the emotions that include Joy, Sadness, Anger, Fear, Trust, Disgust, Surprise and Anticipation, of course. The next step consists of using the plotly function make_subplots(), to produce a subplot of the line chart illustrating emotion counts progression over time. More importantly, significant dates of the diary entries of Anne Frank, such the "Warsaw Ghetto Uprising," "First entry after D-Day," and "First entry after Hiding,”. Next, the incident lines are added to the plot and a text that describes the reason, date and sentiment for that day is displayed. Finally, the `st.plotly_chart()` method of Plotly is utilized to show the time series graph of changing emotional trends in Anne Frank's diary inside the Streamlit user interface, allowing users to interact with the graph and see these trends.

Insights:

Anne Frank’s diary provides a story of her personal life and of events she, her family and many unnamed people went through while hiding from the Nazi persecution between July 1942 and after August 1944. Anne reveals her feelings towards different occurrences that are portrayed by the animated time series graph which indicate the changes in her emotions during this period.

First Entry after Hiding (Negative): The moment that Anne went into the hiding place, it might have been terrifying and upsetting and made her much more aggressive and cynical. This was maybe due to the fact that the environment around her was so dangerous and that there were so many things that were terrible.

First Entry after D-Day (Neutral): The mixed feeling from Anne's tentative optimism followed by her individual reaction to the Allied attack might have also caused her neutral feeling. This was because she neither felt euphoria about freedom nor did she realized how uncertain the future would be.

Warsaw Ghetto Uprising (Highly Negative): Such extreme sensitivity and deep empathy that Anne could show for the actions that happened in the Ghetto for sure can be called as a crucial factor for Anne to get that uniquely pessimistic view of the world along the tragic events and the horrors that Jewish people could watch and for the worst, experience.

Basically, the animation sequence graph shows not only Anne Frank's fragility but also her strength and empathy, therefore, depicts her real human nature even during difficult times. In such a situation she is a courageous

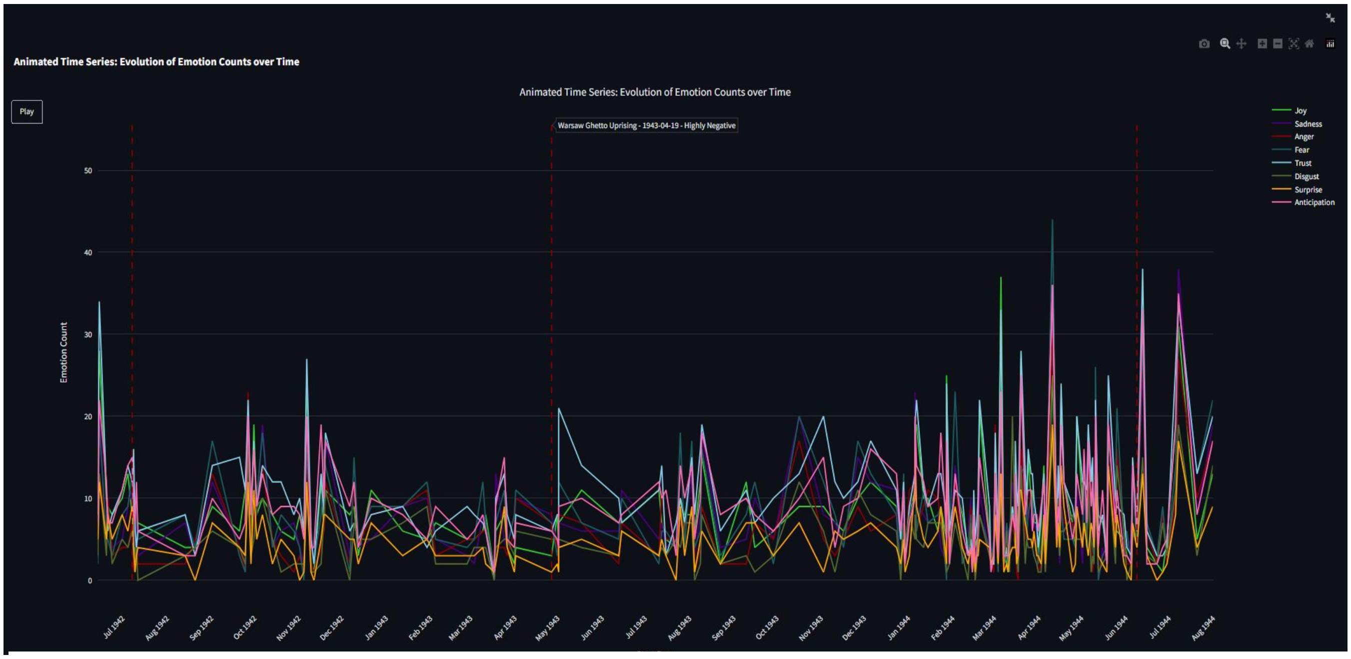


Figure 13: Animated Time Series

depiction of her fight against with fear and despair, hope and human emotions and their complexity.

6.5.10 Graph 9: Highest Count of Each Emotion with Date

The maximum counts of the various emotions found in the dataset are shown in a bar chart with the title " Figure 14: "Highest Count of Each Emotion with Date" followed by the corresponding days. Subsequently, the code selects the dataset's columns that match the emotions and generates an array of those. a dictionary where colours are allocated to each emotion so they can be easily differentiated visually. Hence, it calculates the highest count and the date that connects with each of the emotion in the dataset. A bar graph is made such that emotions lie on x-axis, maximum counts at the y-axis, and colours pulled from the defined colour dictionary by Plotly. The date of the highest of counts each emotion is indicated in each bar.

Insights:

Dates on the graph correspond with the noticeable events from the life of Anne Frank when she was hiding away providing us with the way into the emotional world of her.

Emotion Analysis:

Anne Frank goes through so all kinds of emotions and some of them are mentioned below.

11 April 1944(Fear, Anger, surprise and anticipation) –

A break-in incident happened in their hiding place that made Anne angry and fearful, catching them completely by surprise and filling them with anticipation for what would unfold next. The men in the house discovered that some burglars trying to break into the warehouse. That caused panic and tension situation among all. They had to hide in darkness fearing for their safety till police arrived. The fear escalated when they heard footsteps and rattling noises, imagining most horror of being exposed by the Gestapo(police), fuelling their anticipation of the worst-case scenario. Anne's anxiety maxed when she heard discussion about burning her diary to avoid incrimination.

6.5.11 Graph 10: Emotion Timeline Analysis

Using Streamlit and Plotly, this code section creates an interactive line chart named " Figure 15: The Animated Time Series: Emotion Counts through Time". Users will be able to see how the emotional tone frequency levels have changed over the course of the time. From the given dropdown list, the users can select what emotion they want to focus on. The given data can be noticed in a line graph with different emotions represented with different colors. Users can find the date and emotion count, simply hover the mouse over the line graph that day, which allow the user to be more interactivity. The depiction shows the emotional transition of Anne Frank during the diary writing time and so it is easy to understand the emotional pattern she describes as diary goes on.

Insights:

The "Emotion Timeline Analysis" talks about frequency of fear over time. While talking about the context specific of Anne Frank's diary, that uniquely reflects Anne's emotional roller coasters while in hiding. The "Fear Count" (y-axis) shows how many times the events and feelings of fears are examined in Anne Frank's diary. At the same time the x-axis is the timeline which spans from July 1942 to July 1944 showing period Anne spent in the Secret Annex. The line graph shows the fear score going up and down when Anne Frank was most anxious or irritated.

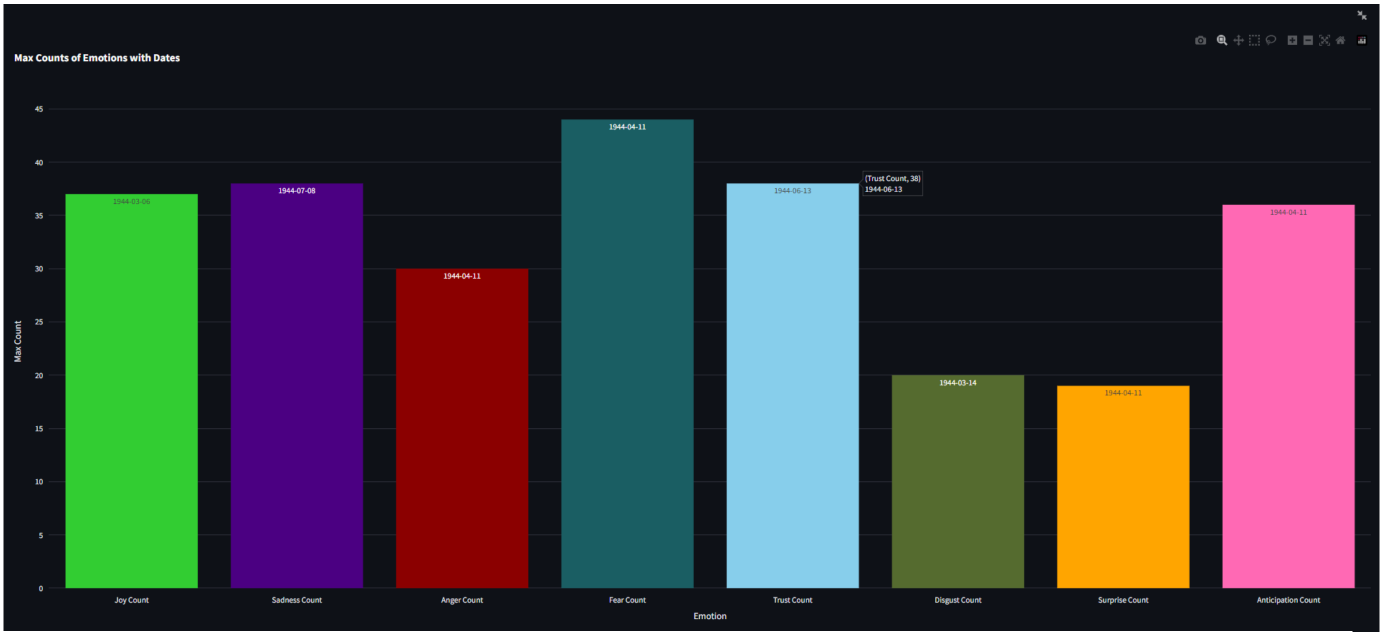


Figure 14: Highest Count of Each Emotion with Date

- Selected Emotion for Timeline Analysis: Fear

Select Emotion:

Fear

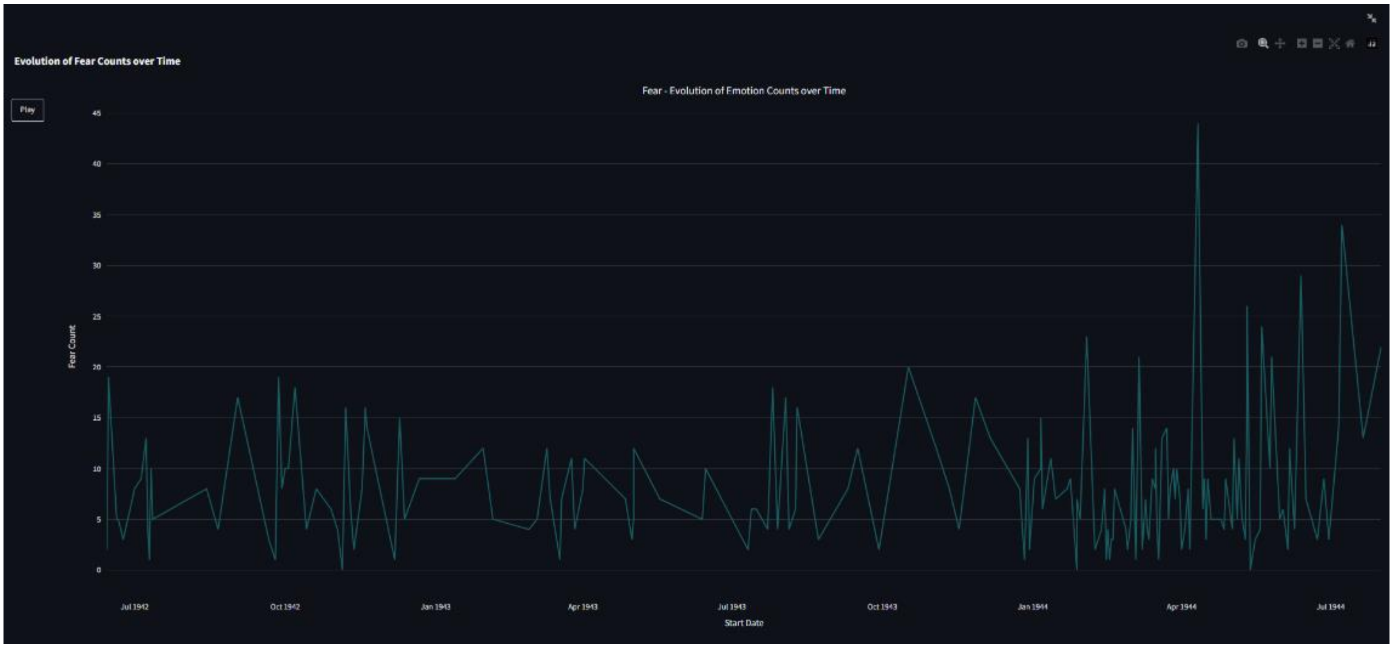


Figure 15: Animated Time Series: Evolution of Emotion Counts over Time

Fear index had gone up unexpectedly at the end, especially. reached its peak on 11 April 1944 when Break –in accident. happened during their hiding time in the secret annex. To conclude, the research sheds a light on the emotional struggles Anne went through during her hiding days and therefore enabling us to better understand her fear and worries in their day-to-day life. It shows that these events also hit her mental health quite seriously and that the diary is always a breakthrough for remembering the human spirit's strength.

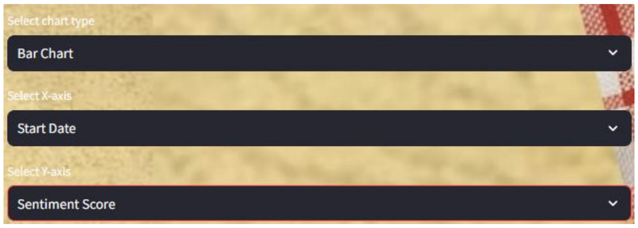
6.5.12 Graph 11: Customize Dashboard

A customizable dashboard using Streamlit, named " Figure 16: "Bar Chart of Sentiment Score against Start Date" shows that you can easily customize this dashboard. Employing dropdown menus, the users will be able to choose between the chart types (line, bar, or scatter plot) and click on the axis variables. Then apply the Plotly graphing library and show the data the user input as line plots, bar plots or scatter plots on the graph where Y-axis shows the selected variable against the X-axis. The different color codes linked to the emotions linked to the selected variable. In the end, the resulting visualization is added to the Streamlit application, and users can explore and analyse data based on their selected items.

Insights:

The graph's X-axis denoted as "Start Date" illustrates the starting date or the event happening from July 1942 to the end in July 1944 as written in Anne Frank's diary. "Sentiment Score" is an X-axis that displays the level of emotional intensity or polarity that is attached to every recorded entry. It each bar is the sentiment score of the specific date, the bar chart provides a direct and succinct illustration of the sentiment scores' over timeline graph.

- Selected Chart Type: Bar chart
- Selected X-axis: Start Date
- Selected Y-axis: Sentiment Score


Select chart type

Bar Chart

Select X-axis

Start Date

Select Y-axis

Sentiment Score

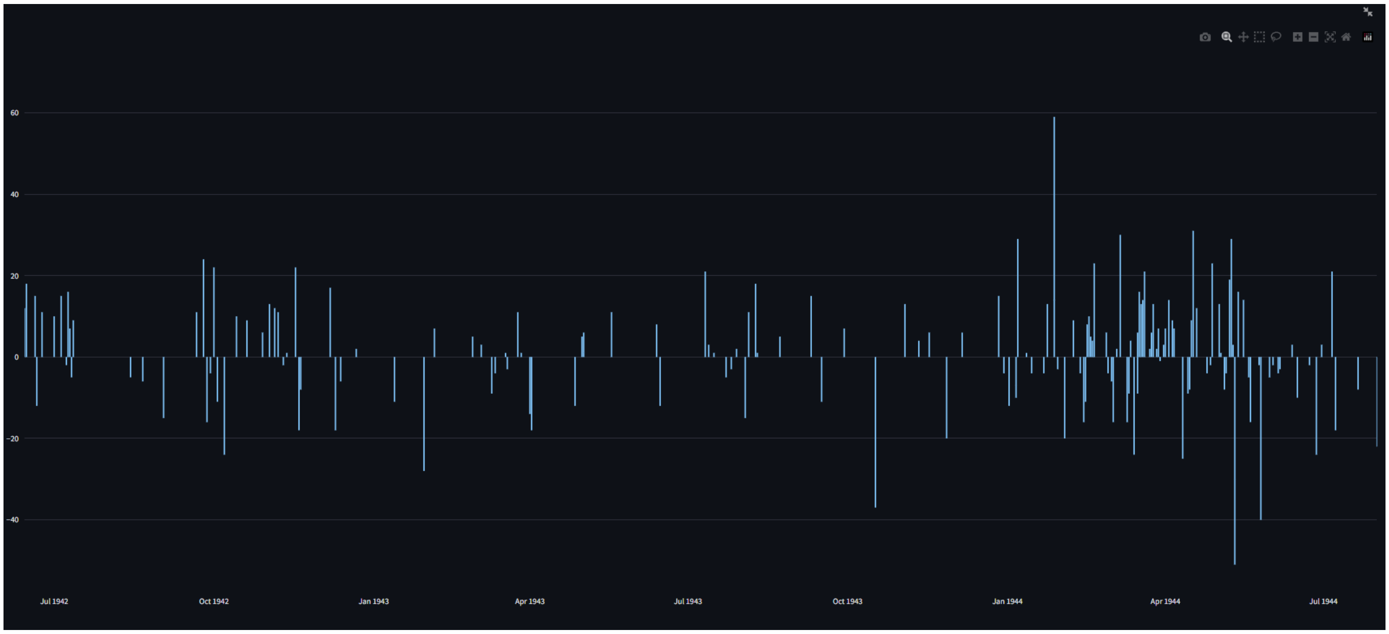


Figure 16: Bar Chart of Sentiment Score against Start Date

6.5.13 Graph 12: Sentiment Distribution Over Time

by the "Start Date" column in the dataset is shown visually in a scatter plot named "Graph 17: "Sentiment Distribution Over Time". These include: "Highly positive", "positive", "highly negative", "negative", and "neutral". the x-axis shows the dates, and the y-value represents the sentiment labels. The sentiment labels are color-coded to make them easy to distinguish: "Neutral" is light grey, "Highly Negative" is dark red, "Positive" is the dark green, and "Highly Positive" is green, "Negative" is light green. This graph will enable us to know the emotions that Anne Frank mentioned and the help of graph of sentiment labels we can also find out that sentiments differ. The distribution of sentiment labels over the time indicated.

Insights:

A scatter plot with the title " Figure 17: "Sentiment distribution over time" is showing in the image. The results of sentiment analysis are for the July 1942–July 1944 period. The x-axis represents the date, and the sentiment labels is plotted on the y-axis. Different colours correspond to five various sentiment levels:

Highly positive- green
Positive - dark green
Neutral - grey
Negative - red
Highly Negative - Dark red

On hovering, each dot of the plot shows the sentiment score on that particular date along with the sentiment label. The dot's location on the y-axis reflects the sentiment level, and its location on the x-axis is equal to the date of the recorded sentiment. Frequency of Highly Negative thoughts is considerably higher in the last portion of period, which suggests the most severe emotional abuse or despair. This development of super strong emotions is likely to be related to the emotional milestones the girl had to face or the struggles the family experienced in times of war, which will, in turn, help us appreciate the psychological dimension of the great number of difficulties during the war.

6.5.14 Graph 13: Sentiment Labels Distribution

Sentiment distribution in the data set is depicted in the following pie chart (Figure 18: Sentiment Labels Distribution). The sentiments are represented with different colours for easy perception. 'Highly Positive' here is green, 'Positive' is light green, 'Highly Negative' is dark red, 'Negative' is red, and 'Neutral' has light grey shade. Illustration of pie chart shows the percentages for the categories of sentiments over the entire dataset, this represents the summary of the distribution of sentiment. Broadly, this is a good way to come up with brief but visually appealing diagrams by showing how label distributions go. This is quite useful when interpreting the data and trying to understand the sentimental patterns it follows.

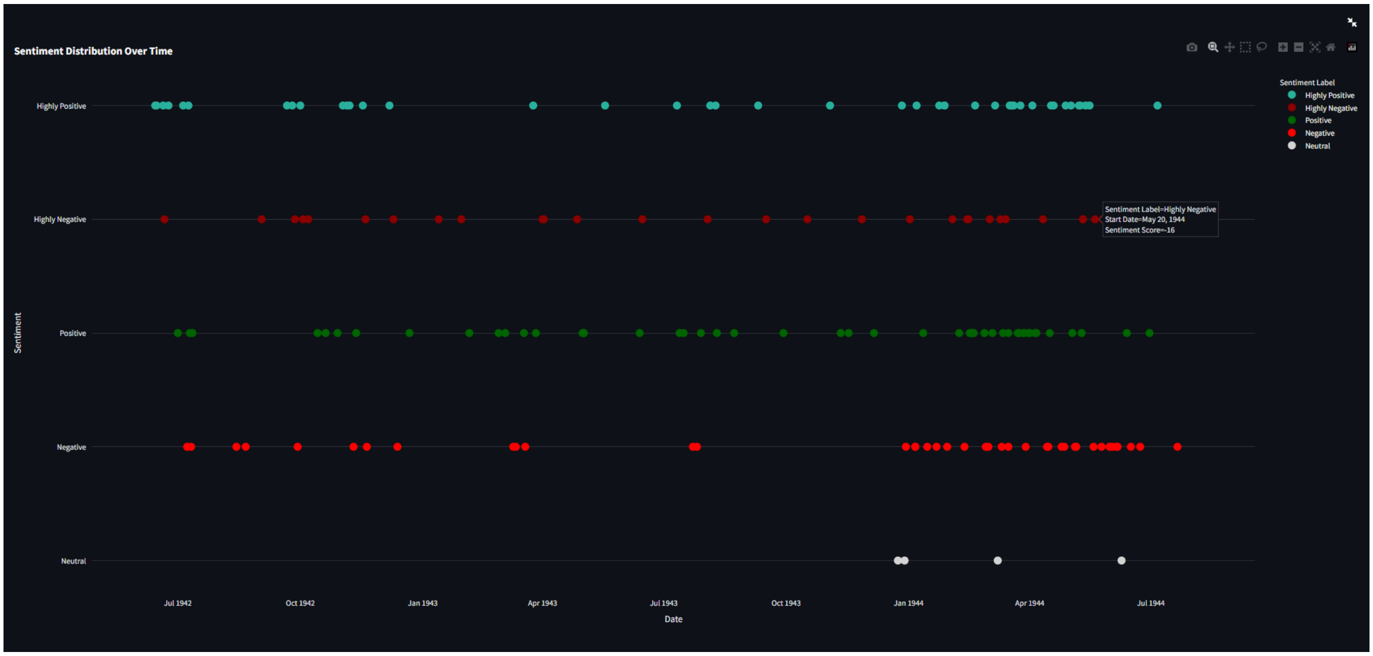


Figure 17: Sentiment distribution over time.

Insights:

The pie chart below indicates the distribution of sentiment labels across the dataset.

Below is the overview of sentiment distribution:

Highly Negative: The slice of dark red park of pie chart represents highly negative sentiment that is of 19%. Which says that most of the words are related to very negative sentiment. Neutral: The smallest piece of pie chart that is of 2.45% is illustrate Neutral sentiment, indicating that most of the sentiments either be positive or negative.

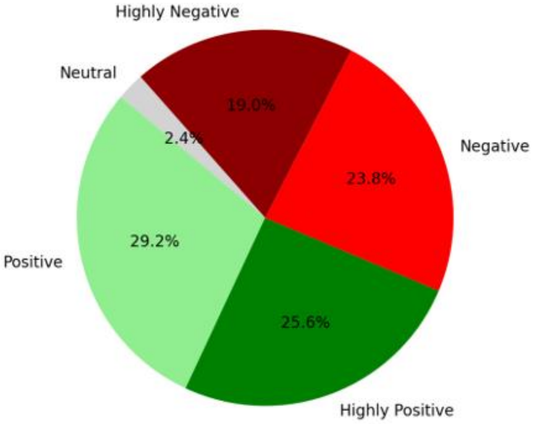


Figure 18: Sentiment Labels Distribution

Negative: The largest red chunk of pie chart represents Negative sentiment of 23% which is like Highly negative just bit more.

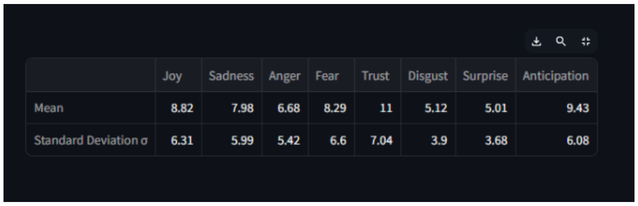
Positive: The biggest slice of pie chart is positive of 29.2%, it is more than a quarter which indicates Anne's strong spirit (Considering the time when the diary is written in the context of WWII)

Highly Positive: The dark green slice which represent 25.6% of pie chart indicates highly positive sentiment which is incredibly showcase hope, optimism, and resilience.

The general graph shows the results of a sentiment analysis that exhibits 54.8% positive sentiments set by positive and highly positive categories, while the percentage of negative sentiments which comprise those of negative and highly negative categories is 42.8%. The proportion of Neutral category is small as compared to the rest, illustrating the highly emotional nature of Anne's diary entries.

6.5.15 Density of Emotion Words in Diary of Anne Frank

With the use of Streamlit's mean() and .std() functions, "Table1: Density of Emotion Words in Novella of Anne Frank: "Number of Emotion Words in Every 10,000 Words" idea was developed. In the below table, the use of emotional words in Anne Frank's diary are shown. For each of the emotion categories—joy, sadness, anger, fear, trust, disgust, surprise, and anticipation, it computes and displays the mean and standard deviation.



	Joy	Sadness	Anger	Fear	Trust	Disgust	Surprise	Anticipation
Mean	8.82	7.98	6.68	8.29	11	5.12	5.01	9.43
Standard Deviation σ	6.31	5.99	5.42	6.6	7.04	3.9	3.68	6.08

Table 1: Density of Emotion Words in Diary of Anne Frank: Number of Emotion Words in Every 10,000 Words

Table 2: Mean and standard deviation of polarity words density in Diary of Anne Frank displaying the density of

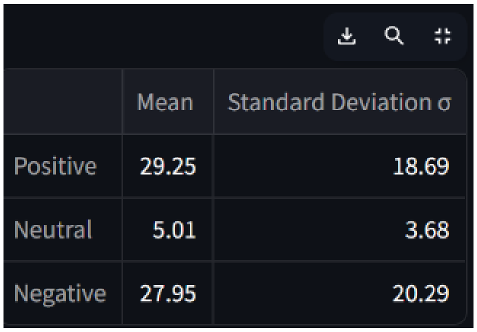
polarity words in Anne Frank’s diary with its mean and standard deviation.

The classification of the number of words with positive (joy, trust, anticipation), negative (sadness, anger, fear, disgust) and neutral (surprise) polarity. The code accumulates the words count of each column and compute the statistics. Then determines the mean and standard deviation for each polarity group. With the help of Streamlit’s write() function, results are arranged in to a data frame for better formatting and it shows Mean and Standard deviation for Positive Negative and Neutral polarities in the diary.

Insights:

This research gives an insight on the emotional world that the journal portrays, which showcase the frequency and verity of emotions like joy, sadness, trust, anger, fear, disgust, surprise, and anticipation.

For instance, words related to trust emotion has highest mean and standard deviation value compared to other emotions. Whereas disgust has lowest density of words.



	Mean	Standard Deviation σ
Positive	29.25	18.69
Neutral	5.01	3.68
Negative	27.95	20.29

Table 2: Mean and Standard Deviation of polarity words density in Diary of Anne Frank

Additionally, Standard deviation shows wide range of emotion diversity. Where Wide range of standard deviation states a wider dispersion of emotions and compare to lower standard deviation meaning a more stable emotional tone.

The second table calculated density of polarity words like Positive, Negative and Neutral. As we can see density of positive words are still more that negative one. User can gain a insights into Anne Frank's diary from mean density and standard deviation value of each sentiment category which helps to highlight emotional intensity and diversity in her diary.

7 Evaluation

Evaluation of the functionality of sentiment analysis on Anne Frank's diary entries requires a in-depth study on the capability of this tool to understand the invisible emotional undertones of historical writing. The focus will be on a custom sentiment analysis created exclusively for this objective, employing sophisticated approaches, linguistic assets and emotion lexicon meant for historical language in particular. An in-depth evaluation of NLTK and NRC Emotional Lexicon reveals their importance as tools for historical data analysis and preservation of the truthfulness of Anne Frank's account. This assessment includes the precision, the importance of the historical context, the understanding of the period, and the maintaining the narrative unity. By means of analysing empirical results, we can go into the intricate performance of each tool through measuring their accuracy, contextualizing sensitivities, and fidelity of the original narrative tone. Recommendations for further enhancement of the custom sentiment analysis and its wider applicability to historical research are suggested. A further improvement would be the incorporation of additional lexicons or other resources to get a wider grasp of the range of emotions. This would help in deep analysis and would therefore present a clearer picture of Anne Frank' emotional roller-coaster. Summing up, sentiment analysis underscores feelings behind the historical scheme and widens the horizons of our perception of the common experience of humanity in the past.

8 Summary and Conclusions

The purpose of the Sentiment analysis project is revealing the emotional content of Anne Frank’s Diary, which deals with some crucial aspects of her experience during the great tragedy World War two.

With help of Natural Language Processing algorithm, we were able to learn about various significant moments of her emotional journey as described in her diary. In fact, Anne’s diary entries are categorized into predetermined sentiment categories including highly positive, positive, negative, highly negative, neutral to measure the horizon of her emotions during the period of war.

Sentimental analysis of Anne Frank's diary gave us a glimpse of experience of young girl suffering in such hard times. Through the close reading of her diary entries as well as the visualization of these entries, we can notice that her emotion evolved from deep depression to little rays of optimism. The research moreover broadens our understanding of Anne's survival and strength of character providing insights into the human experience in such situations. Subsequently, it becomes evident that there are three main lessons that help to overcome the worst times that people experienced in the World War II: hope, courage and the flexibility to adapt to the new reality.

9 Appraisal

This project has been a great learning experience that offers an important insight into various aspects of data analysis, sentiment analysis and tool development. Considering the whole project, I realizes there are aspects that I would have done differently and there are some aspects that I would maintain:

What I Would Do Differently:

- In data Preprocessing, Lemmatization could have added along with the spell checker and stemming.
- Instead of using NRC Emotion Lexicon to extract data and customize sentiment analysis, I could have gone for existing techniques like Vader or other existing sentiment analysis modules that could increase the accuracy. Also, there are some amazing deep learning models like BERT. I could have even use textblob.

What I Would Do the Same:

- I would again select Anne Frank's diary as my dataset, Anne Frank's courage, optimism and resilience taught me a lot.
- Next, I would use Streamlit for visualization, it's a user-friendly software that creates interactive and aesthetic graphs. Even it is python based so user don't have to learn any other coding language.

Advice for others:

Invest sufficient time in in-depth project planning. Take feedback from others and implement it in your project. Also, Explore more on other sentiment analysis and text pre-processing approaches. In the end, do prioritize the documentation of your entire project, take feedback on it.

Lesson Learnt:

This project helps me to understand how to visualize the data properly using streamlit. Previously I worked on R and power BI for visualization. However, I would prefer streamlit over anything else. NLP techniques like tokenization, stop word removal I came across while doing literature review. Furthermore, I learnt about NRC Emotion Lexicon while I was doing research about how to extract emotions and do sentiment analysis. Overall, the project highlighted the importance of interdisciplinary collaboration as it also showed the data analysis is an iterative process which helps one to explore meaningful insights form historical texts.

10 Future Considerations

Improvement of sentiment analysis can be done by investigating and improving on the existing approaches and reading more about the most recent NLP strategies such as the deep learning concept and making the inclusion of different lexicons or resources that would lead to a more comprehensive tool for emotion or sentiment analysis. Develop a approach to enhancement which will consist of the following steps: data preprocessing, training models and evaluation. In a pursuit of validation, collaborate with experts and iterate according to their suggestion. For the purpose of transparency and big steps forward, make records of decisions and results.

11 Acknowledgements

I would like to appreciate the following people and institutions that played a significant role in the success of the project. I would like to first of all render my deep gratitude to my project supervisor 'Christina Moir' who was the source of my support, guidance as well as expertise all through the project. Their invaluable insights and approval comments made the project much richer than it otherwise would be. I would like to acknowledge the "University of Dundee" for providing resources and facilities which helped me carry out the research and the analysis. Furthermore, I also wish to express my gratitude to the developers of the open-source software tools and libraries used in this project which are NLTK, Matplotlib, and Streamlit. However, our research would have been impossible without the authors responsible for the creation of the NRC Emotion Lexicon (Mohammad and Turney). It was their detailed data set that helped to research emotions and do sentiment analysis for this project.

The people I mention above, and others indirectly or directly involved with this task I would like to sincerely acknowledge and give my thanks to. All your efforts and cooperation during this campaign really help in meeting the targets, and I truly appreciate the role you played.

12 References

[1] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc.

[2] Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.). Pearson.

[3] Manning, C. D., Raghavan, P., & Schütze, H. (2020). Introduction to Information Retrieval. Cambridge University Press.

[4] Mohammad, S. M., & Turney, P. D. (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (pp. 26–34).

[5] "Python Data Science Handbook" written by Jake VanderPlas

[6] Streamlit Documentation: https://docs.streamlit.io/

[7] Original Publisher of Anne Frank's Diary: Frank, A., & Pressler, M. (1952). The Diary of a Young Girl. Doubleday.

[8] Matplotlib Documentation: https://matplotlib.org/stable/contents.html

[9] NRC Word Emotion Lexicon link: https://ia801900.us.archive.org/10/items/nrc-emotion-lexicon-v0.92/NRC_emotion_lexicon_list.txt

[10] NLTK Project. (n.d.). NLTK 3.6.2 documentation. Retrieved from https://www.nltk.org/

[11] McKinney, W. (2017). Python for Data Analysis

[12] "From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales" by Saif Mohammad (2013)

[13] VADER: A Parsimonious Rule-based Model for Sentiment Analysis of social media Text by Hutto and Gilbert (2014)

[14] Img link: https://i0.wp.com/culturacolectiva.com/wp-content/uploads/2023/02/6F4HDIPBL5HFBFQ6PJZNROU6SI.jpeg?fit=1600%2C645&ssl=1

13 Appendices

1. Appendix 1: Streamlit App link https://sentimentalanalysisofannefrank.streamlit.app/

2. Appendix 2: Diary of Anne Frank in PDF format: Anne-Frank-The-Diary-Of-A-Young-Girl.pdf

3. Appendix 3: Pre-processed data of Diary for Anne Frank: Filtered_Anne-Frank-The-Diary-Of-A-Young-Girl.doc

4. Appendix 4: Extracted data in CSV format: emotions_between_dates_with_sentiment.

5. Appendix 5: app.py (Steamlit app code)

6. Appendix 6: FlowChart of Sentimental Analysis of Historical Data

7. Appendix 7: NRC emotion lexicon in text format: NRC-emotion-lexicon-wordlevel-alphabetized-v0.92

8. Appendix 8: Poster: Poster presentation 2543320

9. Appendix 9: Power point presentation: Sentiment analysis of historical text

10. Appendix 10: Background Images in app: Anne-Frank.jpg, Anne-Frank2.jpg

11. Appendix 11: Ethics Declaration Form

12. Appendix 12: Non-Clinical-Research-Ethics-CHECKLIST-1-Requirements-for-Ethical-approval-v3-29032019 (3)

13. Appendix 13: SSEN Risk Assessment Form for Student Projects -