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

[Sentiment Analysis of Cultural Heritage Texts: 1](#_Toc162304119)

[1 Introduction 5](#_Toc162304120)

[2 Literature review 5](#_Toc162304121)

[3 Background 7](#_Toc162304122)

[4 Specification 7](#_Toc162304123)

[5 Design 7](#_Toc162304124)

[5.1 Choice of Technologies 7](#_Toc162304125)

[5.1.1 Python 8](#_Toc162304126)

[5.1.2 NLTK (Natural Language Toolkit): 8](#_Toc162304127)

[5.1.3 NRC Emotion Lexicon 8](#_Toc162304128)

[5.1.4 Pandas 8](#_Toc162304129)

[5.1.5 Streamlit 9](#_Toc162304130)

[6 Design Method 10](#_Toc162304131)

[6.1 Data Cleaning and Preprocessing 10](#_Toc162304132)

[6.2 Evaluation of Technologies and Methodologies 10](#_Toc162304133)

[6.3 Selection of Natural Language Processing (NLP) Techniques 11](#_Toc162304134)

[6.3.1 Tokenization 11](#_Toc162304135)

[6.3.2 Regular Expressions (Regex) 11](#_Toc162304136)

[6.3.3 Stopword Removal 11](#_Toc162304137)

[6.4 Integration of Emotion Lexicon Method 11](#_Toc162304138)

[6.5 Sentiment Analysis Process 11](#_Toc162304139)

[6.5.1 Calculate Sentiment Score 11](#_Toc162304140)

[6.5.2 Label Sentiment 12](#_Toc162304141)

[6.6 Date Pair Analysis 12](#_Toc162304142)

[6.7 Visualization Using Streamlit 12](#_Toc162304143)

[6.8 Future Considerations 12](#_Toc162304144)

[7 Implementation 12](#_Toc162304145)

[7.1 Preprocessing and Data Cleaning 12](#_Toc162304146)

[7.2 Sentiment Analysis and Emotion Extraction 13](#_Toc162304147)

[7.2.1 NRC Emotion Lexicon Loading 13](#_Toc162304148)

[7.2.2 Emotion Extraction 14](#_Toc162304149)

[7.2.3 Sentiment Score Calculation and Labeling 14](#_Toc162304150)

[7.2.4 Date Pair Analysis 14](#_Toc162304151)

[7.2.5 CSV Data Export 14](#_Toc162304152)

[7.3 Visualization in Streamlit 14](#_Toc162304153)

[8 Visualization 15](#_Toc162304154)

[8.1 Understanding Streamlit 15](#_Toc162304155)

[8.2 Downloading Streamlit 15](#_Toc162304156)

[8.3 Sentimental Trends Visualization Requirements 15](#_Toc162304157)

[8.4 Visualization Method 15](#_Toc162304158)

[8.5 Graphs and Visualizations: 15](#_Toc162304159)

[8.5.1 Graph 1: Emotion Counts over Time 15](#_Toc162304160)

[8.5.2 Graph 2: Emotion Distribution Pie Chart 16](#_Toc162304161)

[8.5.3 Graph 3: Sentiment Score Distribution Histogram 17](#_Toc162304162)

[8.5.4 Graph 4: Correlation Heatmap 18](#_Toc162304163)

[8.5.5 Graph 5: Word Clouds for Emotions 18](#_Toc162304164)

[8.5.6 Graph 6: Distribution of Sentiment Labels with Extreme Scores 19](#_Toc162304165)

[8.5.7 Graph 7: Emotion Transition Diagram 20](#_Toc162304166)

[8.5.8 Graph 8: Animated Time Series 21](#_Toc162304167)

[8.5.9 Graph 9: Highest Count of Each Emotion with Date 22](#_Toc162304168)

[8.5.10 Graph 10: Emotion Timeline Analysis 23](#_Toc162304169)

[8.5.11 Graph 11: Customize Dashboard 24](#_Toc162304170)

[8.5.12 Graph 12: Sentiment Distribution Over Time 25](#_Toc162304171)

[8.5.13 Graph 13: Sentiment Labels Distribution 26](#_Toc162304172)

[8.5.14 Density of Emotion Words in Diary of Anne Frank 27](#_Toc162304173)

[9 Functionality 27](#_Toc162304174)

[10 Summary and Conclusions 28](#_Toc162304175)

[11 References 28](#_Toc162304176)

[12 Appendices 28](#_Toc162304177)

Table of Figures and tables

[Figure 1: Flow Chart of Sentimental Analysis of Historical Data 10](#_Toc162304236)

[Figure 2: Sample of Original data of Anne-Frank-The-Diary-Of-A-Young-Girl 13](#_Toc162304237)

[Figure 3: Sample of Filtered data of \_Anne-Frank-The-Diary-Of-A-Young-Girl 13](#_Toc162304238)

[Figure 4: Sample of emotions\_between\_dates\_with\_sentiment.csv. 14](#_Toc162304239)

[Figure 5: Emotion Counts over Time 16](#_Toc162304240)

[Figure 6: Emotion Distribution Pie Chart 17](#_Toc162304241)

[Figure 7: Sentiment Score Distribution Histogram 17](#_Toc162304242)

[Figure 8: Correlation Heatmap 18](#_Toc162304243)

[Figure 9: Word Cloud of selected emotion 19](#_Toc162304244)

[Figure 10: Distribution of Sentiment Labels with Extreme Scores 20](#_Toc162304245)

[Figure 11: Emotion Transition Diagram 21](#_Toc162304246)

[Figure 12: Animated Time Series 22](#_Toc162304247)

[Figure 13: Highest Count of Each Emotion with Date 23](#_Toc162304248)

[Figure 14: Animated Time Series: Evolution of Emotion Counts over Time 24](#_Toc162304249)

[Figure 15: Bar Chart of Sentiment Score against Start Date 25](#_Toc162304250)

[Figure 16: Sentiment distribution over time. 26](#_Toc162304251)

[Figure 17: Sentiment Labels Distribution 26](#_Toc162304252)

[Table 1: Density of Emotion Words in Diary of Anne Frank: Number of Emotion Words in Every 10,000 Words 26](#_Toc162262513)

[Table 2: Mean and Standard Deviation of polarity words density in Diary of Anne Frank 26](#_Toc162262514)

***Abstract* -** *Sentiment analysis applied to historical texts provides a unique lens through which we can delve into cultural heritage. The research focuses on Anne Frank's emotional diary, a crucial historical document that provides a vivid depiction of life during a challenging era.*

*By harnessing natural language processing (NLP) tools, we aim to dissect the sentiments embedded within its pages.*

*Anne Frank's diary provides a unique and profound glimpse into the struggles, worries, and dreams of a young Jewish girl amidst the backdrop of World War II. By parsing Anne’s sentences, we uncover the emotional nuances woven into her narrative. NLP algorithms allow us to identify joy, sorrow, fear, and resilience—emotions that resonate across time.*

*Through sentiment analysis, we gain insights into Anne’s inner world. Her moments of despair, fleeting joys, and unwavering determination come alive. We explore temporal trends—how Anne’s emotions evolve over time.*

*Together, this interdisciplinary study sheds light on the complexity of human emotions resilience, and the enduring impact of historical narratives. The diary of Anne Frank emanates a remarkable sense of bravery that reverberates through time.*

*Its pages, marked with ink, serve as a powerful symbol of resilience and hope, showcasing the unwavering strength of the human spirit.*

# Introduction

In the digital age, where algorithms parse language and pixels hold narratives, the intersection of natural language processing (NLP) techniques with historical texts beckons us toward uncharted territories. Within the faded ink of letters, the parchment of manuscripts, and the brittle pages of diaries, emotions linger—echoes of lives lived, dreams cherished, and struggles endured. Our study delves into a deep exploration of the emotional aspects of history, focusing on Anne Frank's diary, a highly regarded document from the 20th century.  
Anne Frank's diary emerges as an incredibly personal narrative—a fragile glimmer of hope and resilience amidst unimaginable adversity. Anne’s reflections—her joys, fears, and desires—become our guide. We tread cautiously, as Anne struggled with youth, love, and the threat of Nazi occupation in these pages.

Employing NLP algorithms, or computational tools, we deciphered phrases, untangled metaphors, and sought the subtle sentiments woven into Anne’s prose. Armed with advanced methods, the aim extends beyond mere joy or sorrow. Leveraging the NRC Emotion Lexicon, emotions were extracted, and sentiment analysis was conducted by calculating sentiment scores and labelling sentiment categories. Additionally, Streamlit was used for data visualization to improve the findings' interpretability. It explores the nuances shades; quiet bravery, fleeting moments of beauty.

The exploration goes beyond individual sentiments to examine the complex temporal currents of emotions. It invites reflection: In what ways did hope fade during airstrikes? Did fear cast a shadow over love's bloom? Analyses lead to vivid representations, a mosaic of emotional experiences that are shared. Anne's heartbeat resonates with myriad others—the soldiers, the refugees, the dreamers. This project, which brings together language, technology, and historical events, has the potential to enhance our understanding of cultural heritage. Examining Anne Frank's diary showcases the remarkable resilience of the human spirit and its lasting influence on our comprehension of history and bravery. In these penned contemplations, the role of guardians of memory is taken for granted. Anne's bravery resonates across generations, and her strength is a guiding light.

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. Engrossed in her narrative, one is transported to the past and inspired by her unwavering commitment to bravery, hope, and perseverance in challenging circumstances.

# Literature review

The fundamental step in preparing textual data for sentiment analysis and other natural language processing (NLP) applications is data cleansing. It includes a number of approaches and strategies used to guarantee the precision, coherence, and dependability of the data that is processed. The main elements of data cleaning, such as text normalization, tokenization, stop word removal, and date pattern recognition, will be thoroughly examined in this literature review along with their importance in the context of sentiment analysis.

Textual data must be standardized into a consistent format that can be analysed, and text normalization is the first stage in data cleaning. This procedure includes handling whitespace, converting text to lowercase, eliminating punctuation, and taking care of any other anomalies that might be present in the data. Researchers can reduce variances and inconsistencies in the text by standardizing it; this will make further processing stages easier and improve the quality of sentiment analysis results.

**Bird, Klein, and Loper (2009)** examined text normalization methods in detail in their seminal work "Natural Language Processing with Python," exploring the complexities of preparing textual data for sentiment analysis and other natural language processing (NLP) applications. Their thorough analysis covers a broad range of text normalization techniques with the goal of standardizing textual material into a format that is appropriate for further analysis. Lowercasing, punctuation removal, managing whitespace, and correcting other inconsistencies inherent in textual data are some of the approaches they explored.

By extensively discussing these text normalization techniques, Bird, Klein, and Loper shed light on the significance of this preprocessing step in the NLP pipeline. They highlight the crucial role of text normalization in mitigating variations and inconsistencies within textual data, thereby laying the groundwork for accurate and reliable sentiment analysis outcomes. Through their detailed examination, they provide valuable insights into the best practices for preprocessing textual data, emphasizing the importance of robust text normalization strategies in ensuring the quality and efficacy of sentiment analysis and other NLP tasks.Their work provides a solid platform for practitioners looking to leverage textual data for purposes other than sentiment analysis.

In "Introduction to Information Retrieval," **Manning, Raghavan, and Schütze (2020)** give a comprehensive review of tokenization strategies, outlining approaches and providing insightful information on methods crucial for efficiently preparing textual data. They offer a comprehensive and authoritative introduction to tokenization approaches, clarifying segmentation strategies at the word and sentence levels. Sentence-level segmentation divides text into distinct sentences, whereas word-level segmentation breaks text down into individual words or tokens. Manning, Raghavan, and Schütze provide researchers with the expertise necessary to effectively extract relevant information from textual data by outlining these strategies.

It becomes essential for researchers to comprehend these tokenization techniques before beginning sentiment analysis or other natural language processing (NLP) projects. Word-level segmentation makes it possible to extract significant features from text so that sentiment and other linguistic aspects can be analyzed later on. Conversely, sentence-level segmentation makes it easier to arrange text into cohesive parts, which helps with textual content interpretation and analysis. Manning, Raghavan, and Schütze enable researchers to efficiently preprocess textual data and create the groundwork for precise and perceptive sentiment analysis results by offering a thorough overview of these tokenization strategies.

Furthermore, Manning, Raghavan, and Schütze emphasize the importance of stop word removal as a preprocessing step. Researchers may minimize noise and improve the emphasis on important content in textual data by eliminating common, low-semantic words known as stop words. The significance of eliminating stop words in enhancing the precision and applicability of sentiment analysis results is highlighted by Manning et al. Researchers can achieve more precise and insightful sentiment analysis outputs by removing stop words from their texts. This helps researchers reduce distractions and focus on the text's fundamental meaning and sentiment.

Sentiment analysis and other natural language processing tasks rely extensively on date pattern recognition, particularly when working with text-based data that contains temporal information. In their thorough manual "Introduction to Information Retrieval," Manning, Raghavan, and Schütze (2020) examine its importance and provide methods for precisely identifying and retrieving date patterns from textual data. Texts frequently include temporal references, which are essential for comprehending context and event timelines, such as dates, timestamps, and durations. Researchers can better manage and evaluate textual material by identifying and extracting these date patterns, which offer deeper insights into sentiment and temporal dynamics.

Manning et al. go into detail on creating pattern matching algorithms and regular expressions specifically for date pattern recognition. Regular expressions make it possible to define text patterns, which makes it possible to recognize common date formats, such as differences in separators and month representations in text. Regular expressions are enhanced by pattern matching techniques, which use machine learning or probabilistic models to reliably identify and extract date-related data from a variety of textual sources. These algorithms effectively handle different date formats by combining linguistic elements and contextual cues, improving text analysis accuracy in real-world scenarios.

**Mohammad and Turney (2010)** have made a substantial contribution to emotion analysis approaches by developing the NRC Emotion Lexicon, in addition to data cleaning techniques. This lexicon is an essential tool for researchers who want to identify and measure the emotional content in textual data. The method used by Mohammad and Turney entails incorporating word-emotion relationships into an extensive lexicon in order to make it easier to analyze emotional cues that are embedded in text.

The NRC Emotion Lexicon, formed by Mohammad and Turney, is a large lexicon of words labeled with emotions, which makes it easier for researchers to understand the emotional content of textual data. By means of their rigorous annotation procedure, Mohammad and Turney create correlations between words and a range of emotions, such as joy, sadness, anger, and fear, among others. Researchers may do sophisticated sentiment analysis and capture the wide spectrum of emotional expressions seen in textual data because to this thorough coverage of emotional relationships.

In addition, the NRC Emotion Lexicon is created and validated through the integration of sophisticated linguistic and computational tools in the work of Mohammad and Turney. Their methodology involves advanced techniques for locating and classifying emotional expressions in text, utilizing computational algorithms and linguistic insights to guarantee the lexicon's dependability and correctness. Using a combined strategy that incorporates knowledge from computer science, psychology, and linguistics, Mohammad and Turney create a strong basis for sentiment analysis research.

The significance of lexicon-based techniques in sentiment analysis is demonstrated by the NRC Emotion Lexicon, which was created by Mohammad and Turney. Their approach supports the systematic examination of emotional content within textual data by giving academics access to a complete repository of word-emotion connections. This allows researchers to get significant insights into the underlying sentiments and attitudes stated by authors. Furthermore, the NRC Emotion Lexicon's accessibility makes it more useful in a variety of contexts, enabling researchers to investigate the emotional dynamics of textual data in disciplines including psychology, marketing, and social sciences.

**Hutto and Gilbert (2014)** also made a big impact on sentiment analysis by developing the VADER lexicon, a specialized tool designed to analyze 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. Fine-grained sentiment analysis is made easier by VADER, which takes into account both emotive symbols and linguistic clues to capture the nuances of casual, conversational language that is frequently used in social media conversations. This program is a great resource for researchers and practitioners looking to evaluate sentiment in online conversations and interactions because of its wide-ranging praise for its ability to capture sentiment nuances in brief, informal messages.

Moving on, A Python framework specifically designed for creating interactive data visualization apps, called Streamlit, was presented by Allaire and associates. The effortless integration of their framework with well-known data science libraries and its intuitive user interface make it stand out. **Allaire et al.** highlight the importance of intuitive visualization tools in increasing the effect and accessibility of data analysis outputs by stressing their ease of use and efficient integration. With the introduction of Streamlit, data visualization has made significant progress. Scholars and practitioners now have a more efficient way to produce visually appealing visualizations and share insights obtained from intricate datasets.

**VanderPlas (2016)** VanderPlas provides a thorough examination of Python-based data science tools and approaches in his well-known book "Python Data Science Handbook." Among the many subjects he covers, VanderPlas gives several visualization libraries—such as Matplotlib, Seaborn, and Plotly—a lot of attention. The book explains how these libraries can create stories that are both visually appealing and educational through in-depth talks and examples. Researchers who are unfamiliar with the wide range of data visualization technologies often find important insights and direction from VanderPlas's thorough guide. His work is an invaluable resource for practitioners who want to use the best visualization tools possible for their data analysis projects, especially when used in conjunction with frameworks like as Streamlit.

To put it briefly, textual data must be cleaned using methods like text normalization, tokenization, stop word removal, and date pattern recognition before it can be subjected to sentiment analysis. By utilizing sophisticated sentiment analysis techniques, such as lexicon-based methods and cutting-edge instruments like Streamlit, researchers can acquire a deeper understanding of the emotional dynamics inherent in textual data. Using these methods and resources will be essential for developing new understandings of human emotions and behavior through textual data analysis as sentiment analysis develops.

# Background

# Specification

# Design

## Choice of Technologies

In crafting the sentiment analysis framework for Anne Frank's Diary, adequate caution was taken to choose technologies that would support the pillars of interpretability, accuracy, and robustness. The recommended technologies were entrusted with creating a robust framework for the analytical process, given the intricacy and historical relevance of the diary's contents. From initial data preprocessing to final visualization, this pipeline necessitated a harmonious integration of tools that could handle a variety of language peculiarities, extract subtle emotional insights, and highlight findings in an understandable manner. In order to guarantee that the sentiment analysis framework would produce significant and useful results, each technology was thoroughly assessed for its suitability in achieving these goals.

### Python

Python serves as the foundational language for implementing various Natural Language Processing (NLP) techniques, data preprocessing steps, sentiment analysis algorithms, and visualization tools in this project. Renowned for its simplicity, readability, and versatility, Python offers an extensive array of libraries and frameworks tailored for diverse data science tasks.

Specifically tailored to the sentiment analysis of Anne Frank's Diary, Python's flexibility and rich library ecosystem allow for seamless integration of various components of the analysis pipeline. From data preprocessing to sentiment analysis and visualization of results, Python provides robust tools for each stage of the process. Its intuitive syntax and comprehensive documentation make it accessible to both novice and experienced practitioners alike.

Moreover, Python's compatibility with other languages and platforms enhances its utility in integrating external resources and deploying analysis frameworks across different environments. Its interoperability with libraries such as NLTK, TextBlob, Pandas, Matplotlib, Seaborn, and Streamlit enables a cohesive workflow, empowering researchers to fully utilize these specialized tools in sentiment analysis tasks without the need for model training.

Overall, Python's widespread adoption in the data science community, coupled with its extensive library support and user-friendly syntax, makes it an indispensable tool for developing sophisticated sentiment analysis frameworks tailored to textual data like Anne Frank's Diary.

### NLTK (Natural Language Toolkit):

One essential element in the field of Natural Language Processing (NLP) is the Natural Language Toolkit, or NLTK. With its wide range of tools, NLTK is essential for sentiment analysis tasks at different phases. Its flexible features cover fundamental operations such as text preparation, including tokenization, stemming, lemmatization, and part-of-speech tagging. While stemming and lemmatization standardize language by reducing words to their base forms, tokenization involves division of text into smaller parts to facilitate analysis which occurs after. Furthermore, deeper analysis is made possible by NLTK's part-of-speech tagging, which gives words grammatical labels. Beyond preprocessing, NLTK offers advanced tools for sentiment analysis, allowing lexicons, machine learning, and rule-based techniques to be used to extract sentiment polarity, intensity, and subjectivity from a variety of textual sources. Through the integration of various features into a cohesive framework, NLTK simplifies the sentiment analysis technique, improving an understanding of the subtle emotional undertones found in Anne Frank's Diary. All things considered, NLTK is an invaluable tool for NLP practitioners and researchers, making it easier to extract insightful information from textual data.

### NRC Emotion Lexicon

An essential external resource for gaining emotions from textual data is the NRC Emotion Lexicon; this is especially true when sentiment analysis is employed to analyse Anne Frank's Diary. A wide spectrum of human emotions, including joy, sadness, fear, anger, trust, disgust, surprise, and anticipation, are covered by the long list of words in this dictionary that have been annotated with their associated feelings. With the aid of this vocabulary, analysts may more accurately pinpoint the emotional expressions that are included into the text of the diary, so improving the accuracy of sentiment analysis.

Furthermore, the NRC Emotion Lexicon's extensive coverage of emotional vocabulary enhances the accuracy and richness of sentiment analysis findings, allowing for a detailed examination of the complex emotional subtleties described in Anne Frank's story. Its well-structured design makes it easy to integrate into sentiment analysis processes and efficiently maps words to appropriate emotional categories. As a result, analysts are able to recognize recurrent themes or patterns that contribute to the emotional landscape shown in the diary, as well as subtle emotional indicators and sentiment changes over time. In the end, Anne Frank's Diary's emotional complexity can be untangled thanks in large part to the NRC Emotion Lexicon, which offers subtle insights necessary for thorough sentiment analysis.

Using these tools allowed for a precise execution of the sentiment analysis of Anne Frank's diary, which allowed for a deep examination of the historical relevance and emotional nuance contained within the text.

### Pandas

One of the many reliable Python modules that can handle structured datasets and is especially helpful for data processing is called Pandas. Two basic data structures, Series and Data Frame, each having a specific function, are at the centre of it all.

* Series: Series operates as a labelled array to handle one-dimensional data. It allows for the flexible storing of different data kinds, such as texts, floats, and integers, and it also offers labelled indexing for convenient element access.
* Data Frame: With rows and columns, Data Frame is a two-dimensional structure that may be customized and looks like a spreadsheet or table. Users may effectively edit and analyse tabular data since each column, like a Series, saves data.

Integration with Data Visualization: Pandas may be used to create intelligent plots and graphs from Data Frame objects, even though it is not primarily a visualization tool. It easily connects with data visualization libraries such as Matplotlib and Seaborn.

Pandas is essential for organizing, and evaluating the structured text that has been taken out of Anne Frank's diary in order to analyse the sentiment of the diary. It simplifies the preprocessing and transformation processes, allowing users to efficiently explore the diary's data and extract insightful information.

### Streamlit

In this project, the Python framework Streamlit—which is intended for creating user-friendly online applications—was essential. Developers created an interactive platform for viewing sentiment analysis results from Anne Frank's diary by utilizing Streamlit's user-friendly interface. Streamlit's features allowed visitors to investigate the diary's emotional story through interactive graphs and charts, making for an engrossing and captivating user experience.  
The usefulness of the platform was further enhanced by the smooth integration of Streamlit with data visualization tools like Matplotlib, Seaborn, and Plotly. This connection made it easier for developers to add complex visualizations, which improved the user's comprehension of sentiment patterns and emotional swings in the journal. In general, Streamlit played a crucial role in creating an engaging and dynamic setting for interpretation and analysis Anne Frank's narrative.

* Matplotlib, Seaborn, and Plotly:

Three popular Python libraries that are well-known for their expertise in data visualization are Matplotlib, Seaborn, and Plotly. These libraries played a key role in improving the Streamlit application's aesthetic appeal and depth of analysis. They made it easier to create visually stunning and educational plots, charts, and graphs by integrating with Streamlit seamlessly.

#### Matplotlib:

Matplotlib, which is well-known for its adaptability and strong plotting features, was essential in creating the illustrations in the Streamlit app, which is used to analyse Anne Frank's diary. Developers created a variety of captivating plots and charts by utilizing Matplotlib's extensive plotting functions, which provided insightful information about the emotional journey that the diary's textual content described.

The creation of line plots that show sentiment trends over time is one important use for Matplotlib. With the use of these plots, readers were able to follow the development of sentiment over the course of many Anne Frank diary periods, highlighting changes in the author's emotional tone and highlighting significant occurrences or themes. Users were able to have a better understanding of the emotional trajectory of the diary and the contextual elements affecting sentiment shifts by using Matplotlib to visualize sentiment trends. Matplotlib also made it easier to create bar charts that illustrate the frequencies of the different emotions that were recorded in the diary. These charts gave readers a thorough overview of the frequency of particular emotions, such happiness, sadness, and rage, making it possible to identify patterns in emotional expression and investigate the underlying ideas that Anne Frank's story explores. Developers were able to customize visuals to the specific needs of sentiment analysis by utilizing Matplotlib's versatile plotting features, which provided users with insightful information about the emotional terrain of the diary's textual content.

#### Seaborn

The Seaborn package, which is based on Matplotlib, was instrumental in improving the graphics in the Streamlit application, which is devoted to Anne Frank's diary. Through the utilization of Seaborn's sophisticated statistical plotting features and high-level interface, programmers were able to produce aesthetically pleasing charts that were easier to read. The use of Seaborn's functionalities facilitated the refinement of narrative elements, including color schemes and story structures, guaranteeing that the illustrations proficiently communicated the sentimental undertones present in the diary's narratives.

Furthermore, the investigation of intricate relationships and patterns within the diary's textual content was made easier by Seaborn's integrated statistical plotting features. By using Seaborn's capabilities, developers were able to explore the emotional journey that was described in the diary in greater detail and were able to identify minute trends and changes in sentiment expression over time. Seaborn improved the user experience and gave readers a greater understanding of the emotional terrain of Anne Frank's Diary by presenting the analysis results in visually appealing and educational charts.

#### Plotly

Plotly is well-known for its adaptability and interactivity, added dynamic and interactive plot features that were specifically designed for the sentiment analysis of Anne Frank's diary, greatly increasing the utility of the Streamlit program. Plotly's wide range of chart types—which include line plots, scatter plots, and 3D visualizations—was utilized to improve the user experience and enable a more immersive dive into the sentiment analysis results obtained from the diary's textual content.

The incorporation of interactive plots from Plotly into the Streamlit program gave users the ability to dynamically change and engage with visualizations, allowing them to directly connect with the emotional journey that Anne Frank's diary portrays. The ability for users to zoom in on particular data points, hover over items to see more information, and switch between various views or perspectives allowed for a greater comprehension of the changing emotional landscape that the diary described. Additionally, Plotly's animation and dynamic update features allowed developers to clearly communicate intricate correlations and patterns, giving viewers the ability to learn more about the subtle emotional shifts and sentiment trends that are included in Anne Frank's story.

# Design Method

The goals and specifications specified for the sentiment analysis of Anne Frank's diary are carefully examined before the design process gets underway. Establishing a clear knowledge of the project objectives and user expectations is the purpose of this first stage.

A diagram of a process

Description automatically generated

Figure 1: Flow Chart of Sentimental Analysis of Historical Data

## Data Cleaning and Preprocessing

Prior to doing sentiment analysis, data preparation and acquisition are essential stages to guarantee the accuracy and consistency of the Anne Frank's Diary dataset. The following thorough procedure is carried out:

* An attempt is made to procure digital versions of Anne Frank's Diary from trustworthy sources. To obtain the text in a machine-readable format, this may entail gaining access to official digital archives, online repositories, or approved publishers.
* For analysis, existing digital copies of Anne Frank's diary that are accessible in electronic formats like PDF, DOCX, or TXT are accessed. Libraries, schools, and websites devoted to historical fiction are the places where one can find these versions.
* The diary dataset is carefully cleaned and pre-processed after it is obtained to make sure sentiment analysis can be performed on it. This entails eliminating any noise, superfluous characters, and formatting artifacts that can obstruct the study.
* Methods like tokenization, lowercasing, and stop word removal are used to normalize the text and get it ready for additional examination.
* The diary text's numerical values, dates, and other non-linguistic content are recognized and handled appropriately. Sentiment analysis processes are designed to make sure that only relevant language content is analyzed.
* In order to keep sentiment analysis algorithms from being interfered with, handling special characters and punctuation entails recognizing and properly handling non-alphanumeric symbols and punctuation marks included in the text input. This can mean deleting these characters or swapping them out with spaces or appropriate placeholders. The text is cleaned and standardized by removing unnecessary characters, which guarantees that sentiment analysis algorithms can concentrate on the important linguistic information without being distracted by noise. In the end, this procedure helps to ensure the consistency and quality of the data, which makes sentiment analysis results more accurate and trustworthy.
* Data integrity, to guarantee the accuracy of the processed text data from Anne Frank's diary, data integrity tests are carried out. These checks include validating data types, finding duplicates, recognizing missing values, and confirming the correctness of the data. The integrity of the dataset is maintained and any problems like redundant or incomplete data are reduced by carefully attending to these details. By using these steps, the sentiment analysis that follows will be more accurate and reliable, guaranteeing that valuable insights may be precisely and confidently drawn from the historical text data.

These procedures are carefully followed to ensure that the Anne Frank's Diary dataset is adequately ready for sentiment analysis, which in turn yields accurate and trustworthy findings.

## Evaluation of Technologies and Methodologies

A systematic approach takes into account several variables such as accuracy, scalability, computational efficiency, and adaptability to historical language fluctuations when evaluating sentiment analysis approaches for historical texts. For the purpose of accurately identifying feelings in historical texts, accuracy is essential, and scalability guarantees effective processing of vast amounts of data. Timely analysis depends on computational efficiency, while cultural context and outdated language present obstacles that must be addressed with flexibility. Robustness and dependability are evaluated via extensive testing, and possibilities for modification and adaptability meet particular requirements. Smooth integration into analysis workflows is ensured via compatibility and integration with current systems. This thorough assessment directs the choice of the best methodology, providing the groundwork for strong sentiment analysis solutions customized to the particular qualities of historical text data.

## Selection of Natural Language Processing (NLP) Techniques

Considering the intricacy and subtleties of historical language, the design places emphasis on using natural language processing (NLP) algorithms that can efficiently handle contextual ambiguity and historical vernacular.

### Tokenization

A crucial stage in natural language processing (NLP) is tokenization, which divides text into discrete pieces known as tokens. Word tokenization, in particular, breaks down text into discrete words to enable in-depth analysis at the word level. Many NLP tasks, including named entity recognition, sentiment analysis, and part-of-speech tagging, depend on this segmentation. Within the given context, word tokenization is carried out via the word\_tokenize() function found in the NLTK package, a popular Python NLP toolkit. Text may be efficiently broken up into words by using this feature, which paves the way for thorough linguistic processing and analysis.

### Regular Expressions (Regex)

The use of regular expressions, or "regex," is essential for locating and extracting particular patterns from text data. To be more precise, regex patterns are defined using the re module in order to identify dates in the text. In order to do this, create a regex pattern that matches the text's dates in the format "Monday, January 1, 2024." Next, using functions like re.compile() to compile the regex pattern and re.findall() to extract all instances of dates matching the pattern, the defined pattern is applied to the text data. The code can precisely find and work with date information in the text by using regex, which makes it easier to do other processing and analytical activities.

### Stopword Removal

Common words like articles (e.g., "a," "an," "the"), conjunctions (e.g., "and," "or "but"), and prepositions (e.g., "in," "at," "on") that have minimal bearing on the overall semantic meaning of a text are known as stopwords in natural language processing. The aim of eliminating stopwords is to focus on the important ideas in the text by eliminating these less useful terms. NLTK's stopwords corpus, which includes predefined lists of stopwords for English, is used in the above code sample. Using word\_tokenize(), the text is first tokenized into individual words. Subsequently, stopwords are extracted from the NLTK stopwords corpus, and every tokenized word is contrasted with this compilation. The text data is essentially refined for later analytical tasks like sentiment analysis or topic modeling by filtering out any stopwords that are found. By removing noise caused by frequently used yet meaningless terms, this procedure improves the quality of the text data.

Through the use of these NLP approaches, the code seeks to efficiently preprocess the text, extract pertinent data, and enable sentiment analysis on text data from the past.

## Integration of Emotion Lexicon Method

* **The NRC Emotion Lexicon loading**

The NRC Emotion Lexicon, which consists of words tagged with emotion categories and related sentiment values, should be loaded to start the sentiment analysis process. To extract words and the related sentiment values and emotions, parse the lexicon file. Put this data in an appropriate data structure so that it may be easily retrieved for use in later analytical operations.

* **Extract Emotions**

Iteratively go over each word in the pre-processed text, looking for matches in the NRC Emotion Lexicon. Determine the sentiment values and feelings that are connected to each word in the dictionary. Combine these feelings, such as "sadness," "anger," "fear," "joy," "trust," "disgust," "surprise," and "anticipation," for every word that appears in the text." In order to facilitate a deeper emotional analysis, keep track of counts for each emotion in order to assess its intensity and prevalence within the text.

## Sentiment Analysis Process

The sentiment analysis process involves preprocessing the text, extracting emotions, calculating sentiment scores, and labeling sentiment based on predefined criteria.

### Calculate Sentiment Score

There are several important procedures involved in determining sentiment scores. First, a predefined lexicon is used to create lists of both good and negative emotions. The frequency of each happy and negative emotion in the text is then counted by iterating over the emotions data dictionary. A positive score and a negative score are then obtained by adding these counts for the positive and negative emotions separately. Lastly, the total negative score is subtracted from the entire positive score to get the sentiment score. This technique offers useful insight into the overall sentiment expressed in the text by quantifying sentiment based on the frequency of both positive and negative emotions.

### Label Sentiment

Predetermined thresholds or criteria are developed to categorize sentiment into distinct labels in order to label sentiment based on sentiment scores. Sentiment labels are assigned based on the sentiment score by using conditional expressions. Positive feelings are classified as "Highly Positive," and negative feelings are classified as "Positive." When a score is zero, it is classified as "Neutral," and when it is between -10 and 0, it is classified as "Negative." Sentiment scores below -10 are categorized as "Highly Negative," offering a clear classification according to the sentiment score's polarity.

## Date Pair Analysis

In order to extract sentiment scores and emotions from the text between each pair of dates, Date Pair Analysis identifies date pairs inside Anne Frank's Diary. This approach provides insights into chronological patterns and emotional swings throughout Anne's story by breaking up the diary entries according to dates. The method offers a methodical way to comprehend how Anne's feelings change over time, from happy and trusting times to depressing, afraid, or furious ones. This approach, which quantifies emotional states using sentiment analysis, improves our comprehension of Anne's experiences and the emotional journey she chronicles in her journal.

The collected emotions and sentiment scores between each date pair can also be saved in a CSV (Comma-Separated Values) file to aid in additional analysis and visualization. This file would have lists and counts of words related with various emotions (e.g., Joy, Sadness, Anger, Fear, Trust, Disgust, Surprise, Anticipation), as well as columns such as "Start Date" and "End Date" to indicate the time span. In addition, the CSV would have columns labeled "Sentiment Label" and "Sentiment Score" to offer a numerical evaluation of the emotional content of each interval. Researchers can perform in-depth studies, investigate chronological patterns, and have a deeper understanding of Anne Frank's emotional journey through her diary thanks to this structured approach.

## Visualization Using Streamlit

In order to expedite the creation of the visualization tool using Streamlit, first arrange the cleaned dataset with sentiment scores, emotion categories, and metadata. Then, install Streamlit using pip install streamlit. Next, define the Streamlit application with interactive elements and visualizations by importing the required libraries, such as Matplotlib and Pandas. Create interactive visualizations, including line charts or bar plots, to show sentiment patterns and emotional dynamics after loading the dataset into the app. Create an intuitively arranged layout and user-friendly interface, tailoring the visualizations and interface elements to match unique needs. Lastly, give the program a full test for correctness and functioning, fix any problems that you find, and upload it to a hosting service that will allow users to access it. The entire development process is streamlined by this method, which combines dataset preparation, Streamlit installation, visualization creation, user interface design, testing, debugging, and deployment.

## Future Considerations

In order to make sentiment analysis better, assess existing approaches, investigate cutting-edge NLP strategies like deep learning models, and incorporate extra lexicons or resources for more comprehensive emotion coverage. Create a methodical enhancement plan that includes steps for data preprocessing, training models, and evaluation. For validation, work with experts and make iterations based on their input. For the sake of openness and upcoming advancements, record decisions and outcomes.

# Implementation

## Preprocessing and Data Cleaning

* Text Extraction: To extract text from the source document, Anne Frank's Diary, which is usually available in PDF format, use the pdfplumber library. Make use of pdfplumber's features to extract the text from each PDF page by iterating through the document.
* Text Preprocessing: To clean the retrieved text and get it ready for analysis, carry out a number of preprocessing procedures.
* Noise Removal: Get rid of any characters, symbols, or other artifacts that don't add anything meaningful to the text's linguistic substance. Eliminate any formatting artifacts or irregularities that may have resulted from the extraction of PDF files, including line breaks, hyphenations, and uneven spacing.
* Removal of Non-Linguistic Content: Determine and eliminate any non-linguistic components that are irrelevant to the textual analysis, such as dates, numbers, or other metadata.

A computer code with text

Description automatically generated

* Eliminate Stop Words: To concentrate on important text, remove often used stop words such articles, prepositions, and conjunctions. For this, use a custom collection of stop words or the stop words corpus from NLTK.
* Tokenization: To enable additional analysis, divide the text into discrete words or tokens. To separate the text into discrete pieces, use the word tokenize() function from NLTK or alternative tokenization methods.

A computer screen shot of a computer code

Description automatically generated

* Data integrity: Validating data types, spotting duplicates, and dealing with missing information are all part of maintaining data integrity. This entails ensuring that the extracted text and metadata have consistent and accurate data types, controlling duplicate entries to avoid redundancy, and correcting any incomplete or missing data by imputation or exclusion from additional analysis.
* Pre-processed Text Saving: Once the text data has been cleaned and pre-processed, save the content in a format that will allow for simple access and additional analysis, like a Word document (Filtered\_Anne-Frank-The-Diary-Of-A-Young-Girl.docx). Make sure the document is saved with the text intact and with any formatting or structure that may be required for further analytical work.
* By using this process, the text taken from Anne Frank's diary is guaranteed to be precise, dependable, and suitable for analysis and interpretation.

A text on a page

Description automatically generated

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

A screenshot of a text

Description automatically generated

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

## Sentiment Analysis and Emotion Extraction

The objective is to use the text taken from Anne Frank's Diary to do sentiment analysis and emotion extraction. The NRC Emotion Lexicon loading, sentiment score calculation, sentiment labeling based on predetermined thresholds, date pair identification within the diary text, sentiment score extraction from preprocessed text, and sentiment score export to a CSV file are just a few of the pivotal tasks that are covered by this scope. Through the inclusion of these chores, the design seeks to offer thorough understandings of the diary's emotional content while enabling additional examination and interpretation of the emotions conveyed in the text.

### NRC Emotion Lexicon Loading

Bringing up the NRC Emotion Lexicon 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.

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.

A screenshot of a computer code

Description automatically generated

### Emotion Extraction

Using matches from the NRC Emotion Lexicon, the design iteratively goes over the preprocessed text to extract emotions at this step. Iterating through the preprocessed text is part of the implementation strategy. The technology verifies if a word appears in the text in the NRC Emotion Lexicon. When a match is found, the term is linked to the lexicon's definition of that category of emotions. After that, emotions are combined for every category, enabling a thorough comprehension of the text's emotional content. By accurately identifying and extracting emotions, this iterative method guarantees that the foundation for later sentiment analysis and interpretation is laid.

A computer screen shot of text

Description automatically generated

### Sentiment Score Calculation and Labeling

In this stage, the sentiment score is determined by adding together the positive and negative feelings that were taken out of the text data and figuring out how different they are from each other. As part of the implementation method, counts of both positive emotions—like joy, trust, and anticipation—and negative ones—like sadness, wrath, fear, and disgust—are added up. The method calculates the positive and negative sentiment ratings by adding together these counts of emotions. The emotion score is then computed by deducting the negative value from the positive score. This method makes it possible to measure the sentiment present in the text, giving rise to insights about the general emotional tone and paving the way for additional analysis and interpretation.

Sentiment is intended to be labeled into discrete groups, such as highly positive, positive, neutral, negative, or highly negative, by using predetermined thresholds to enable classification. The implementation strategy is setting thresholds, or the ranges within which sentiment scores fall into a certain category, for each sentiment category. The system then uses these preset thresholds to determine a sentiment label for every sentiment score. The sentiment expressed within the text is classified by the algorithm by comparing the sentiment score with the predefined criteria and selecting the appropriate sentiment label. This method makes sentiment patterns easier to analyze and makes it easier to comprehend the emotional context that the text data is trying to express.

A computer screen shot of a program

Description automatically generated

### Date Pair Analysis

To enable sentiment and emotion analysis through segmentation, the goal of this step is to locate date pairings (each date entry) in the diary text. Use of regular expressions to find date patterns is part of the implementation process. The machine then goes over these pairs repeatedly to extract sentiment scores and emotions out of each one. Through the extraction of data and storage of start and end dates for subsequent analysis, this method offers insights into emotional tendencies throughout time.

### CSV Data Export

In order to facilitate further research or visualization, the extracted emotions and sentiment scores between date pairs have to be saved in a CSV file(emotions\_between\_dates\_with\_sentiment.csv). The data is written into the CSV format during implementation by use of file I/O operations. In order to make it easier to understand and handle the stored data further, it is imperative that the CSV file has the proper headers for every column.

By adhering to this thorough process, the design guarantees precise sentiment analysis and emotion extraction from Anne Frank's diary, facilitating additional research into emotional patterns across time.

A screenshot of a computer

Description automatically generated

Figure 4: Sample of emotions\_between\_dates\_with\_sentiment.csv.

## Visualization in Streamlit

To depict emotion distribution, sentiment score distribution, and word frequency, the approach entails creating visualization components such word clouds, pie charts, line charts, and histograms. The sentiment distribution, important emotion-word correlations within the text, and emotional changes over time are all revealed by these visualizations. Each visualization is created using Python code (app.py), which ensures interactivity and user-friendliness by utilizing Streamlit's functions and visualization libraries. The app can be released on platforms following extensive testing to confirm functionality. Clear communication between users and maintainers is ensured by documenting the development process, including design choices and implementation specifics.

In the subsequent section, a deeper exploration of each visualization component will be conducted, focusing on their design, implementation, and significance in detail.

# Visualization

**Analysing Sentimental Trends of Anne Frank's Diary**

The Streamlit framework, a potent tool for creating interactive web applications using Python, is used in the visualization process to analyse the emotive themes found in Anne Frank's Diary. Streamlit is the perfect tool for creating data-driven apps since it enables the smooth integration of user interaction, data processing, and visualization all within a single Python script.

## Understanding Streamlit

* Streamlit offers an easy-to-use interface that lets you create web apps straight from Python scripts.
* Developers can specify the application's structure, behavior, and layout using the well-known Python syntax.
* Standard code is not necessary because Streamlit performs operations like data loading, visualization rendering, and user interface element handling automatically.
* It is simple to add interactive widgets, including buttons, sliders, and dropdown menus, to improve user interaction and data exploration.

## Downloading Streamlit

Streamlit can be downloaded and installed using pip, the Python package manager. Using your terminal or command prompt, type the following command to install it:

pip install streamlit

* Streamlit's Use in Visual Studio:

Python must first be installed on your computer before you can utilize Streamlit in Visual Studio. After installing Python, launch Visual Studio, open a new Python project, and follow the instructions to install Streamlit. After that, you can start developing the code for your Streamlit application by creating a new Python script (.py file).

## Sentimental Trends Visualization Requirements

Make sure the following conditions are met before starting the visualization:

• Anne Frank's Diary dataset: Get the textual data in a format that works for you, such as a CSV file, from Anne Frank's Diary.

• Essential libraries for Python: Install the necessary Python libraries (Pandas, Matplotlib, Seaborn, WordCloud) for data processing, analysis, and visualization.

• Streamlit: Verify that Streamlit is set up in the manner previously mentioned.

## Visualization Method

Using Streamlit, an interactive online application is made so that users may examine emotive patterns in Anne Frank's diary. The application will depict emotion counts, sentiment score distribution, word frequency, and more using a variety of chart types, including word clouds, pie charts, histograms, and line charts. Interacting with these visualizations will allow users to learn about the distribution of sentiments within the text and emotional trends over time.

## Graphs and Visualizations:

### Graph 1: Emotion Counts over Time

The counts of different emotions throughout time in Anne Frank's diary are shown in Figure 4: Emotion Counts over Time. The graph is produced with Matplotlib and Streamlit. A multiselect widget lets users choose which particular emotions they want to examine.

Every emotion on the graph is represented by a different color. The graph displays the following emotions: surprise, joy, anger, sadness, fear, trust, disgust, and anticipation. The start date of the diary entries is shown by the x-axis, and the count of each emotion is represented by the y-axis.

The graph shows how the counts of each emotion change over time in a visual manner. Readers can gain a fuller grasp of the emotional experiences that Anne Frank chronicled by observing the patterns and variations in her feelings throughout the diary.

The graph will show the numbers of the user-selected emotions. As a result, the graph is modified to emphasize the selected emotions and give a clear visual depiction of their counts over time. A notice asking the user to choose emotions from the multiselect widget to visualize is shown if no emotions are selected. All things considered, the graph offers a thorough and dynamic representation of the emotion counts throughout Anne Frank's diary, allowing readers to investigate and evaluate the text's emotional elements.

* Selected Emotions for Counts over time: Joy and Anger

A close up of a text

Description automatically generated

A graph of different colored lines

Description automatically generated

**Insights:**

Joy and Anger are the two extreme emotions that the user has chosen for analysis. A timeline of these emotions' counts over time is shown on the graph.

It is evident from the graph that at first, there was a greater number of joy than anger. But as the timeline goes on, something changes, and by the conclusion, the count of Anger exceeds the count of Joy.

This discovery implies that Anne Frank's diary's emotional content changed over time. The diary notes might have first described happier events or feelings. But as time passed, there seemed to be a change in the diary toward more angry incidents or displays of rage.

This realization illuminates a dynamic facet of Anne Frank's emotional journey as chronicled in the diary, illustrating a shift from primarily joyful emotions to a growing degree of anger.

Figure 5: Emotion Counts over Time

### Graph 2: Emotion Distribution Pie Chart

As shown in the Figure 5: Emotion Distribution Pie Chart Pie charts are used to illustrate how different emotions are

distributed over a dataset. The python code that creates the pie chart tallies the totals for each emotion and uses Matplotlib and Streamlit to plot the data.  
The dataset's relative proportions of each emotion are

shown in a pie chart. The names of each emotion are listed together with the matching proportion of the overall count displayed next to each label. The chart's colours serve as a useful tool for distinguishing and comparing the various moods.  
In general, the pie chart offers a visual aid for interpreting and evaluating the emotional content.

**Insights:**

The emotional journey of Anne Frank shows the complex range of feelings she experienced while hiding. The frequency of fear (13.3%) and sadness (12.8%) reveals the significant obstacles and uncertainties she faced, yet moments of joy (14.1%) and anticipation (15.1%) reflect optimism and hopefulness. Furthermore, despite the danger and isolation, the high level of trust (17.7%) indicates a certain dependence on others for solace and support. The

fact that she expressed anger (10.7%), disgust (8.2%), and surprise (8%) highlights the complexity of her experiences, which included disgust, astonishment, and irritation in the

midst of the turbulent conditions of hiding during the Holocaust.

Ultimately, Anne's emotional journey highlights her tenacious spirit in the face of hardship by illuminating her complex reaction to adversity, which includes moments of resistance, despair, and resilience.

A pie chart with different colored circles

Description automatically generated

Figure 6: Emotion Distribution Pie Chart

### Graph 3: Sentiment Score Distribution Histogram

A graph of different colored bars

Description automatically generatedThe sentiment score distribution obtained from the examination of Anne Frank's diary is represented graphically in the "Figure 6: Sentiment Score Distribution Histogram" graph. The y-axis shows the frequency of occurrence for each sentiment score bin, and the x-axis shows the sentiment scores, which are separated into discrete bins. A custom colormap representing the range of sentiment scores from negative to positive is used to colour the histogram bars. The colours span from dark red to red to green to dark green. The height of each bar represents the

Figure 7: Sentiment Score Distribution Histogram

frequency with which sentiment scores fall into each bucket. A summary of the sentiment score distribution throughout the diary is given by this visualization, which also offers insights about the dominant sentiments and their frequency in Anne Frank's story.

**Insights:**

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

1. The bulk of the bars are grouped to the left of the zero mark, suggesting that the dataset contains a high proportion of negative sentiment scores.

2. Bars are most frequently seen in the negative score range, specifically in the range of -20 to -15, indicating that sentiment scores in this area are prevalent throughout the diary.

3. Scores are notably concentrated around the neutral zone, suggesting that a sizable portion of emotion scores fall within this range.

4. There are fewer occurrences of strongly positive sentiment in the dataset as the frequency of bars falls as the sentiment scores go toward the extreme positive end of the scale.

According to this histogram study, which has a peak in the negative region, Anne Frank's diary has a significant quantity of negative sentiment. Furthermore, neutral feeling is widely prevalent, whereas extreme positive sentiment is somewhat uncommon. Understanding the general sentiment trends depicted in Anne Frank's story can benefit greatly from these observations.

### Graph 4: Correlation Heatmap

Based on count columns, the correlation matrix is computed. The correlation matrix data is used to construct the heatmap(Figure7: Correlation Heatmap), where the colour of each cell indicates the correlation coefficient between two variables. The correlation values are displayed using the 'RdBu' colorscale, where variations in correlation strength are represented by shades in between, with red denoting positive correlation and blue denoting negative correlation. The layout is updated with a title after the heatmap is plotted using Plotly's Heatmap function. Streamlit's plotly\_chart method is then used to display the heatmap.

**A screenshot of a computer screen

Description automatically generated**

Figure 8: Correlation Heatmap

**Insights:**

A correlation heatmap, or graphical depiction of the correlation matrix between a set of variables, is displayed in the image. The variables in this instance appear to indicate the relative amounts of various emotions, including "Joy," "Sadness," "Anger," "Fear," "Trust," "Disgust," "Surprise," and "Anticipation."

This is a summary of how to interpret this heatmap:

1. Colours: The heatmap's colours span from red to blue, and the scale is shown by a colour bar on the right. Higher positive correlations are shown by red, higher negative correlations by blue, and values ranging from positive to negative are represented by the hues in between.

2. Correlation Values: The correlation between the emotions on the X- (horizontal) and Y- (vertical) axes is represented by each square (or cell) in the heatmap.

An ideal positive correlation is denoted by a correlation value of 1, an ideal negative correlation by a correlation value of -1, and no correlation by a correlation value of 0.

3. Interpretation: For instance, the cell with a correlation value of -0.611338 is shaded red and located at the intersection of the X- and Y-axes labelled "Disgust Count" and "Surprise Count." This shows that the counts of the emotions "Disgust" and "Surprise" have a somewhat negative association.

Since each variable's correlation with itself is always 1 (perfect positive correlation), the heatmap is usually symmetric along its diagonal. When analysing data, the

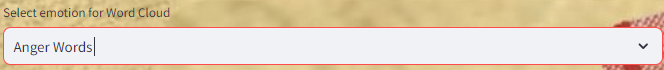
heatmap can be especially helpful in immediately showing relationships between variables so that patterns and correlations can be understood.

### Graph 5: Word Clouds for Emotions

A word cloud displayed in Figure 8: Word Cloud of selected emotion created using the chosen emotion from the dataset of Anne Frank's diaries. Users may choose the emotion for which they want to create the word cloud using a dropdown menu in the Streamlit app. When an emotion is selected from the dataset, the code concatenates all the words related to that emotion. It then creates the word cloud visualization by using the `WordCloud` class from the `wordcloud` package. Using a predetermined color scheme based on the

chosen emotion, each word in the word cloud is coloured. The app's interface shows the generated word cloud image.

.

- Selected emotion for Word Cloud: Anger Words  A close up of words

Description automatically generatedThis feature helps readers explore the emotional themes and sentiments portrayed throughout Anne Frank's diary by providing a graphic depiction of the key phrases associated with particular emotions.

Figure 9: Word Cloud of selected emotion

**Insights:**

 A word cloud, which is a graphic representation of text data where each word's size represents its relevance or frequency. This particular word cloud, which belongs to "Graph 5: Word Clouds for Emotions" and has the theme "Anger Words," is a portion of a bigger set or presentation.   
The word cloud highlights words related to rage in red, highlighting the emotional context. Words of all sizes are visible, including the following: "death," "fear," "lonely," "rage," "bad," "awful," "fight," "hurt," "crazy," "upset," "hate," and "scream," among others. In the context of the dataset used to create this word cloud, the larger the word, the more likely it is to be connected to the feeling of anger.

This visualization technique is frequently used to swiftly highlight the most important elements in a document, and it may also be used to organize words according to distinct themes or emotions by using color. Since red is frequently associated with strong emotions, especially aggression or anger, the usage of red for words that express wrath in this

instance is consistent with widespread cultural connotations.

### 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 9: Distribution of Sentiment Labels with Extreme Scores as a bar chart. In order to prepare data for graphing, the code defines sentiment categories, such as "Neutral," "Highly Negative," "Negative," and "Highly Positive." Subsets of the dataset are then extracted by iteratively going over each sentiment category. The number of sentiment labels and the extreme score—which might be either the maximum or minimum for positive or negative sentiments—for every subset are determined and recorded in a dictionary named `plot\_data'. After that, a bar chart is plotted using Matplotlib and this dictionary, with a bar for each sentiment label. The extreme score is indicated at the top of each bar, and the height of each bar reflects the number of sentiment labels. Each sentiment type has a unique color given to it to improve visual separation. Lastly, the `st.pyplot()` function is used to display the bar chart inside the Streamlit app interface. By displaying extreme scores linked to certain sentiments, this visualization A graph of a distribution of samples

Description automatically generatedprovides readers with insights into the distribution of sentiment labels inside Anne Frank's diary.

Figure 10: Distribution of Sentiment Labels with Extreme Scores

**Insights:**

A bar chart labeled "Graph 6: Distribution of Sentiment Labels with Extreme Scores" is displayed in the picture. The distribution of various sentiment labels, together with their corresponding counts and extreme scores, are shown in this graph.

Five categories are displayed on the graph's "Sentiment Label" x-axis:

Highly Positive, Positive, Negative, Highly Negative or Neutral

The "Count" of occurrences for each sentiment label is shown by the y-axis, which runs from 0 to 50.

The chart features color-coded bars with numerical values representing the number of occurrences above them and a "Extreme" score:

The green bar in the "Highly Positive" category has a count of 45 ,The green bar in the "Positive" category is lighter and has a count of 10.

A pink bar with a count of 10 represents the "Negative" category. The red bar in the "Highly Negative" category has a count of 40. A grey bar with a count of 0 and an extreme score of 0 represents the "Neutral" category.

Based on this data, we may deduce that most sentiments were categorized as "Highly Positive" or "Highly Negative" as a result of the sentiment analysis, with "Highly Positive" sentiments being somewhat more common. Few "Positive"

and "Negative" sentiments were found, and no instances of "Neutral" sentiment were noted. The extreme scores most likely reflect the sentiment's intensity or strength.

### Graph 7: Emotion Transition Diagram

A representation of the emotions transitioning between various states found in Anne Frank's diary dataset is called an Emotion Transition Diagram(Figure10: Emotion Transition Diagram). The DiGraph() function in NetworkX is then used to generate a directed graph object G. The method then extracts pairs of successive emotions—the present feeling and the subsequent emotion—by iterating through each row of the dataset. The transition from one emotion to the next is represented by an edge that is added to the graph G for each pair. Following the construction of the graph, the code uses the spring layout technique to determine the node positions, which are then stored in the pos variable. The nodes are positioned with the assistance of this layout technique to reduce edge overlap.

Next, using the subplots() function in Matplotlib, a new figure and axis are produced for the display, with the size of the figure specified. NetworkX's draw() method is used to draw the graph.

Using the st.pyplot() function, the graph is finally shown in the Streamlit app . Users can gain insights into the flow and order of emotions with this depiction.

A diagram of a negative and negative

Description automatically generated

Figure 11: Emotion Transition Diagram

**Insights:**

An "Emotion Transition Diagram," a kind of directed graph that shows how one emotional state might change into another, as depicted in the image. Four primary emotional states are shown as nodes in this diagram:

1. Highly Negative

2. Negative

3. Positive

4. Highly Positive

5. Neutral

Arrows connecting each node to the others show potential transitions between various emotional states. The fact that the arrows return to the same node indicates that an emotional state can endure for a long period.

This is a thorough description of the diagram:

- "Highly Negative" can stay that way or change to "Negative."

- The state "Negative" may change to "Neutral," "Positive," or return to "Negative."

- "Positive" can change to "Neutral," "Highly Positive," or stay "Positive."

It is possible for "Highly Positive" to change back to "Positive" or to stay "Highly Positive."

It is possible for "Neutral" to change into any of the four emotional states or to stay in "Neutral."

The figure only shows that these transitions are possible—not their likelihood or frequency. It's a visual depiction that could be utilized to comprehend and forecast emotional trajectories in user experience research, psychological investigations, or sentiment analysis algorithms.

### Graph 8: Animated Time Series

The Animated Time Series visualization displayed in Figure 11: Animated Time Series shows how Anne Frank's diary's emotion counts changed over time.

The relevant colors are established for the emotions of interest, which include Joy, Sadness, Anger, Fear, Trust, Disgust, Surprise, and Anticipation. Next, we use Plotly's `make\_subplots()` function to generate a subplot with a line chart showing the progression of emotion counts over time.

Furthermore, important events from Anne Frank's diary, such the "Warsaw Ghetto Uprising," "First entry after D-Day," and "First entry after Hiding," are specified together with the corresponding dates. Then, to represent these occurrences, vertical lines are added to the plot, and text that includes the event name, date, and sentiment for that day is displayed.

Lastly, Plotly's `st.plotly\_chart()` method is used to display the animated time series graph inside the Streamlit app interface, enabling users to interact with and see the shifting emotional trends throughout Anne Frank's diary.

**Insights:**

The diary of Anne Frank documents her personal experiences and her journey throughout the turbulent time she spent hiding from Nazi prosecution, from July 1942 to after August 1944. Anne's emotional reactions to different events are depicted in an animated time series graph that shows how her emotions have changed over this period.

**July 8 1942,**"**First entry after hiding"**: This is the first time Anne has thought about anything since she and her family went into hiding. As Anne struggles with the difficulties and uncertainties of concealment, the rise in A screen shot of a computer screen

Description automatically generatednegative emotions like despair, fear, and rage around this period represents the terrible reality of their circumstances.

Figure 12: Animated Time Series

**June 9, 1944, "First entry after D-Day**": Anne's diary entries after the Allied invasion of Normandy can show a range of feelings, such as relief and hope, but also possibly some concern or fear for what lay ahead. The comparatively numb emotional reaction on this particular occasion points to a wide range of complicated emotions as Anne negotiates the changing nature of the conflict.

**April 19, 1943 "Warsaw Ghetto Uprising":** Anne was probably deeply affected by this incident because she could relate to the suffering of individuals who were subjected to persecution and opposition in the Warsaw Ghetto. The increased negative feelings associated with this day are a reflection of Anne's compassion, grief, and rage over the atrocities committed during the Holocaust.

In summary, the animated time series graph provides an insight into Anne Frank's emotional state, showcasing her humanity, resiliency, and empathy throughout hardship. In the midst of conflict and persecution, it offers a moving depiction of her journey as she struggles with fear, despair, hope, and the complexity of human feeling.

### 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 12 : Highest Count of Each Emotion with Date" along with the dates that correlate to them. The code then pulls the dataset's columns that correspond to different emotions and creates a

dictionary that assigns a colour to each emotion to help with visual identification. It then determines the highest count and the date that corresponds to it for every emotion in the dataset. A bar chart with emotions on the x-axis, maximum counts on the y-axis, and colours chosen from a predefined colour dictionary is made using Plotly. The relevant dates of the maximum counts are displayed as text labels on each bar.

**Insights:**

The dates on the graph match significant events that Anne Frank experienced while she was hiding, offering insight into her emotional journey:

- **March 6, 1944**: The "Joy Count" peak may represent a joyful or consoling occurrence under the challenging circumstances of hiding. This contentment is probably a

result of their friendship, their shared laughing, and the enjoyment of small pleasures like music listening.

- **July 8, 1944**: Anne's feelings over her captivity and her desire for liberation are reflected in the top of the "Sadness Count". As Anne considers the risks and uncertainty of their circumstances, she struggles with extreme sadness. She feels a great deal of anxiety about being found out and about maybe losing her treasured journal.

- **April 11, 1944**: This date stands out since it was the apex of the "Anger," "Fear," "Surprise," and "Anticipation" tallies. On April 11, 1944, indignation is aroused by Anne's invasion of privacy, especially when it is suggested that her diary be destroyed in order to hide their hiding site. There is a strong sense of injustice and outrage at this invasion of her personal space and the danger to her treasured writings. During a break-in panic, Anne feels terrified since she knows that something will be discovered. When she waits for the police to arrive, she feels vulnerable and on edge due to the mystery surrounding the intruders' intentions.

**- June 13, 1944**: The "Trust Count" peak points to a period of hope or faith, perhaps resulting from a sense of trust or camaraderie among the hiding place's occupants.

**- March 14, 1944**: Anne's disgust or annoyance with the harsh circumstances of concealment or with other people's disagreements may be reflected in the "Disgust Count" peak.

A graph of different colored squares

Description automatically generated with medium confidence

Figure 13: Highest Count of Each Emotion with Date

The graph, when viewed against the backdrop of the Holocaust and the difficulties of confinement, offers a comprehensive perspective of Anne Frank's emotional landscape throughout her time in hiding. It captures moments of joy, sadness, fear, trust, disgust, and anticipation.

### Graph 10: Emotion Timeline Analysis

Using Streamlit and Plotly, this code section creates an interactive line chart named " Figure 13: Animated Time Series: Evolution of Emotion Counts over Time". Users can examine how various feelings' emotion counts have changed throughout time. Users can choose the emotion they wish to analyze from a dropdown menu in the sidebar. The chosen emotion's count over time is shown as a line graph, with various emotions represented by different colors. Users may see changes in emotion counts over the course of the timeline thanks to animation features, which improves the interactive experience. This graphic provides insights into Anne Frank's emotional journey during the era

of the diary entries and helps to comprehend the temporal patterns of emotions reflected in her writing.

**Insights:**

The "Emotion Timeline Analysis" with a particular emphasis on how fear counts have changed over time. Given the setting of Anne Frank's diary, this study is especially important since it sheds light on the emotional journey she went through while hiding. The "Fear Count," represented on the y-axis, represents the frequency of fear-related events or feelings that are recorded in Anne Frank's diary. On the other hand, the timeline is represented by the x-axis, which runs from July 1942 to July 1944 and represents the time that Anne was imprisoned in the Secret Annex.

Analysing the graph indicates variations in the fear score, which correspond to times when Anne Frank felt particularly anxious or uneasy. Fears of Nazi discovery or the difficulties of incarceration are likely associated with major events or times of increased stress, as evidenced by peaks in the fear count, such as the noteworthy rise around October1943.   
In summary, this examination illuminates the psychological distress that Anne Frank experienced throughout her incarceration, providing an insight into the worries and fears she faced on a daily basis. It emphasizes how deeply her experiences affected her mental health and how her journal will always be valuable as a reminder of the resilience of the human spirit in the face of hardship.

A screenshot of a computer screen

Description automatically generated

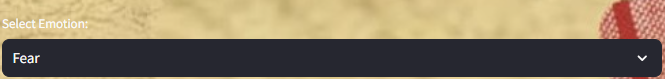
Figure 14: Animated Time Series: Evolution of Emotion Counts over Time

### Graph 11: Customize Dashboard

A customizable dashboard using Streamlit, named " Figure 14: Customize Dashboard - Bar Chart of Sentiment Score against Start Date" Users can personalize and see data on this dashboard in a way that suits their tastes. Users can choose the kind of chart (line, bar, or scatter plot) and the variables for the X and Y axes from dropdown menus using interactive widgets in the sidebar. The 'Start Date' column is converted to datetime format for appropriate processing of the data when the user makes selections. Then, using the graphing library Plotly, several visualizations are produced according to the user's selections.

These visuals, which display the chosen Y-axis variable against the 'Start Date' on the X-axis, can be line plots, bar plots, or scatter plots. Using a specified dictionary, the color of the plotted data points or lines is adjusted based on the associated emotion with the chosen Y-axis variable. Ultimately, the generated visualization appears inside the Streamlit application, enabling users to dynamically interact and examine the data according to their choices.

* Selected Emotion for Timeline Analysis:Fear



**Insights**:

The chosen bar graph visualization's X-axis denotes the "Start Date," which shows the chronology of observations or events that, based on Anne Frank's diary, most likely runs from July 1942 to July 1944. The "Sentiment Score" is a Y-axis that illustrates the degree of emotional intensity or polarity linked to every recorded entry or occurrence. With each bar representing the sentiment score recorded on a particular date, the bar graph offers a clear and succinct depiction of the sentiment scores over time.

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

A black and tan striped background

Description automatically generated

A screen shot of a graph

Description automatically generated

Figure 15: Bar Chart of Sentiment Score against Start Date

### Graph 12: Sentiment Distribution Over Time

The distribution of sentiment labels across the timeframe indicated by the "Start Date" column in the dataset is shown visually in a scatter plot named "Graph 12: Sentiment Distribution Over Time". Labels f or sentiment include "Highly positive," "Positive," "Highly negative," "Negative," and "Neutral." Every point on the scatter plot represents a note or observation made by Anne Frank; the date is indicated by the x-coordinate, and the sentiment label associated with that note is indicated by the y-coordinate.

The sentiment labels are color-coded to make them easy to distinguish: "Neutral" is light grey, "Highly Negative" is dark red, "Positive" is dark green, and "Highly Positive" is green.

This graph offers insights into the emotional patterns and oscillations within Anne Frank's diary entries by displaying the sentiment distribution over time.

**Insights:**

In the picture, there is a scatter plot with the title " Figure 15: Sentiment distribution over time." The graphic seems to be displaying sentiment analysis findings for the July 1942–July 1944 time period.

The date is shown on the x-axis, while the sentiment level is shown on the y-axis. Different colors correspond to five various feeling levels:

- Highly Positive (light green)

- Positive (dark green)

- Neutral (grey)

- Negative (red)

- Highly Negative (dark red)

Every dot on the plot indicates the sentiment score for that specific date. The dot's location on the y-axis represents the sentiment level, and its location on the x-axis represents the sentiment's recording date.

Figure 16: Sentiment distribution over time.

The frequency of Highly Negative thoughts increases noticeably toward the end of the timeframe, indicating a period of increased emotional suffering or despair. This increase in intensely negative feelings may be related to important occasions or difficulties that Anne and her family encountered, which would deepen our comprehension of the psychological effects of being confined during a war.

### Graph 13: Sentiment Labels Distribution

The distribution of sentiment labels in the dataset is shown graphically in the provided pie chart. The sentiment labels are assigned different colors to visually represent them. 'Highly Positive' is shown in green, 'Positive' in light green, 'Highly Negative' in dark red, 'Negative' in red, and 'Neutral' in light grey. The chart presents the percentage of each sentiment category in relation to the total, offering a concise summary of the sentiment distribution. In summary, this method successfully creates a concise and visually attractive display of sentiment label distribution, which assists in analyzing and understanding the sentiment patterns within the dataset.

**Insights:**

The pie chart provides an insight into the distribution of sentiments found in Anne Frank's diary entries. Here is a comprehensive overview of the sentiment categories:

Highly Negative: Indicated by the largest red section, which makes up 19% of the chart, a significant number of the entries express strongly negative sentiments.

Neutral: The smaller red slice next to it represents 23.8%, indicating a significant but somewhat less intense negative

sentiment compared to the 'Highly Negative' category.

**A pie chart with different colored circles

Description automatically generated**

Figure 17: Sentiment Labels Distribution

Academic: The analysis of Anne Frank's emotional expressions reveals that only a small grey segment, accounting for 2.4% of the entries, indicates minimal neutrality. The majority of her entries tend to lean towards either positive or negative sentiments.

Positive: The majority of the entries, about 29.2%, are filled with positive emotions, which is quite impressive considering the challenging circumstances described in the diary.

Positive: The lighter green slice, which makes up 25.6% of the entries, reflects moments of hope, resilience, and optimism in the face of adversity.

In general, the chart presents a sentiment analysis that shows a mostly positive sentiment (54.8% when combining Positive and Highly Positive), with negative sentiments (including Negative and Highly Negative) making up 42.8% of the data. The Neutral category remains relatively small compared to the others, highlighting the emotional intensity that permeates Anne Frank's diary entries.

### Density of Emotion Words in Diary of Anne Frank

Employing Streamlit's `st.markdown` function, "Table1: Density of Emotion Words in Diary of Anne Frank: Number of Emotion Words in Every 10,000 Words" was created. Statistics on the frequency of emotion words in Anne Frank's diary entries are shown in the table. For every emotion category—joy, sadness, anger, fear, trust, disgust, surprise, and anticipation—it computes and shows the mean and standard deviation.

A screenshot of a computer

Description automatically generated

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

A “Table 2: Mean and Standard Deviation of polarity words density in Diary of Anne Frank” showing the polarity word density in Anne Frank's diary together with its mean and standard deviation. the

columns that list the number of words with positive, neutral, and negative polarity and classify them appropriately. The word counts for each polarity are then represented by concatenating these columns into a single list. The code adds up the word counts in each column, computes statistics across the rows, and determines the mean and standard deviation for each polarity group. Using Streamlit's `st.write()` function, the results are arranged into a DataFrame for improved formatting and are shown as the mean and standard deviation of the polarity word density

for positive, neutral, and negative categories in Anne Frank's diary.

A screenshot of a computer screen

Description automatically generated

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

**Insights:**

This study provides an insight into the emotional terrain that the journal portrays, highlighting the frequency and diversity of feelings like happiness, despair, rage, fear, trust, disgust, surprise, and expectation.

For example, a prominent prevalence of trust-related themes or phrases in Anne Frank's work is suggested by the higher mean density of terms associated to trust when compared to other emotions. On the other hand, a relatively reduced emphasis on disgust expressions is indicated by the lower mean density of terms associated to disgust.

Additionally, the standard deviation values show the range of emotional expression in the diary and shed light on the variability or dispersion of emotion word density. Higher standard deviations are indicative of a wider range of emotional expression, whereas lower standard deviations are indicative of a more stable emotional tone.

The second table goes even farther into sentiment analysis, classifying words into sentiments that are favourable, neutral, and negative. This study provides a more sophisticated comprehension of the diary's overall emotional sentiment. Insights into Anne Frank's emotional experiences and views throughout the period recorded in the diary can be gained from the mean density and standard deviation values for each sentiment category, which highlight the emotional intensity and diversity in her writing.

# Functionality

Assessing the functionality of sentiment analyses for Anne Frank's diary entries requires a thorough evaluation of their ability to capture the subtle emotional nuances present in historical text. The main emphasis is on a custom sentiment analysis designed specifically for this purpose, using advanced methodologies, linguistic resources, and sentiment lexicons created for historical language. Conducting an exhaustive examination of sentiment analysis tools like NLTK and the NRC Emotional Lexicon provides valuable insights into their effectiveness in analysing historical data and maintaining the legitimacy of Anne Frank's narrative. The assessment takes into account various factors, including precision, significance in the historical context, comprehension, and the preservation of narrative integrity. By analysing empirical results, we are able to dig into the intricate performance of each tool, emphasizing their accuracy, contextualize sensitivity, and faithfulness to the original narrative tone. Suggestions for further improvement and future research recommendations focus on enhancing the effectiveness of the custom sentiment analysis and expanding its usefulness in historical research. Later enhancements could involve integrating additional lexicons or resources to achieve a more comprehensive coverage of emotions. This would enhance the analysis and offer a more profound understanding of Anne Frank's emotional journey. This assessment highlights the importance of sentiment analysis in shedding light on the emotional aspects of historical texts and enhancing our comprehension of the human experience throughout history.

# Summary and Conclusions

The sentiment analysis project sought to explore the emotional landscape portrayed in Anne Frank's diary entries, offering valuable insights into her experiences and reactions during the turbulent era of World War II. Through the application of natural language processing techniques, we were able to gain valuable insights into Anne's emotional experiences as documented in her diary entries. The analysis classified emotions into predetermined sentiment categories, such as highly positive, positive, neutral, negative, and highly negative, offering an in-depth comprehension of Anne's inner world amidst the chaotic time of World War II. By analyzing visualizations and reports, we discovered patterns, trends, and varying levels of emotions expressed over time. This gave insightful information about Anne Frank's struggles, fortitude, and hopeful moments amid hardship.  
Analysing Anne Frank's diary sentimentally has allowed us to see the emotional journey of a young girl living through one of the saddest periods in human history. Through close reading and visualization of her journal entries, we have been able to uncover the complex dynamics of her feelings, which included both profound sadness and worry as well as happy and hopeful moments.

This investigation has not only enriched our comprehension of Anne's personal challenges and determination but has also provided a glimpse into the wider experience of humanity during the genocide Looking ahead, the analysis provides valuable insights into the enduring effects of hope, courage, and the unwavering adaptability of humanity, even in the face of the most difficult situations.

# 11 References

* Allaire, J.J., Kronenberg, F., Esper, G., & Sequeira, D. (2020). Streamlit: Streamlit: the fastest way to build and share data apps. Zenodo. https://doi.org/
* Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc.
* Jurafsky, D., & Martin, J. H. (2019). Speech and Language Processing (3rd ed.). Pearson.
* Manning, C. D., Raghavan, P., & Schütze, H. (2020). Introduction to Information Retrieval. Cambridge University Press.
* 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).
* <https://www.docdroid.net/N8uGQ0o/anne-frank-the-diary-of-a-young-girl-pdf>

(More References needs to add)

# 12 Appendices