6COSC023W – Final Project Report

Person's Behaviour Analysis with text messages via NLP - BAWT

Student: Mohammad Sabiq Sabry (19219495)

Supervisor: Anastasia Angelopoulou

This report is submitted in partial fulfillment of the requirements for the

BSc (Hons) Computer Science degree

BEng Software Engineering degree

at the University of Westminster.

School of Computer Science & Engineering

University of Westminster

Date: 02/05/2023

Declaration

This report has been prepared based on my own work. Where other published and unpublished source materials have been used, these have been acknowledged in references.

Word Count: 9200

Student Name: Sabiq Sabry (19219495)

Date of Submission: 02/05/2022

Abstract

This comprehensive report presents the findings of a case study exploring the development of an AI-based system aimed at improving the communication skills of individuals with Alexithymia or Autism. The study utilized a range of innovative techniques, including polarity analysis using sentiment analysis, sentiment analysis-based cloud words, topic modelling, emoji analysis using TF-IDF, unique words using TF-IDF, mood identifier, and exploratory data analysis (EDA), in order to gain valuable insights into the challenges and opportunities of text-based communication.

The primary objective of the study was to gain a deeper understanding of the challenges faced by individuals with Alexithymia or Autism when it comes to understanding and expressing emotions in text-based communication. To achieve this, the study used sentiment analysis, which provided a detailed picture of a person's behavioural pattern by analysing the positivity level of their text-based communication, including the use of emojis and gifs to relate to their feelings.

The study also followed a rigorous white box testing methodology, which involved testing each individual component of the NLP algorithm to ensure that they were working correctly and producing accurate results. The results of the study were highly encouraging, with the AI-based system proving to be highly effective at analysing and interpreting the emotional context of text-based communication for individuals with Alexithymia or Autism.

While the study was successful, there were some limitations that should be considered. For example, the cleaning for the text formats could not be fixed for iOS devices, and the data was not stored in an encrypted platform.

In conclusion, this study has made significant strides in the development of an AI-based system that can improve the communication skills of individuals with Alexithymia or Autism. The study used a range of innovative techniques and followed a rigorous testing methodology, providing valuable insights into the challenges and opportunities of text-based communication. The findings from this study have the potential to make a significant impact in improving communication skills for individuals with Alexithymia or Autism.

Acknowledgements

I would like to dedicate this report to my beloved parents, Muneera Sabry and Sabry Haniffa, for their unwavering support and encouragement throughout my academic journey. Their love, wisdom, and guidance have been instrumental in shaping my character and inspiring me to pursue my passions.

I am deeply grateful for their sacrifices and the countless hours they invested in providing me with the best possible education. Without their love and support, I would not have been able to achieve the goals I have set for myself. This report is a testament to their unwavering belief in me and their commitment to my success.

Thank you, Mom and Dad, for everything you have done for me.

I hope this report makes you proud and serves as a small token of my appreciation for all that you have done for me.

Table of contents

BAWT:

List of Tables

List of figures

1. Introduction

1.1 Problem statement

Text-based communication has become increasingly popular in recent years. Messaging applications such as WhatsApp have over 2 billion active users (Ruby, 2022) and are widely used by individuals of all ages and businesses for various purposes. However, individuals with conditions such as Alexithymia or Autism may find it challenging to understand the tone of conversations. This is because text-based communication offers individuals the freedom to express themselves without the pressure of face-to-face social interactions, which tends to lead to more honest and open communication. However, the lack of visual and auditory cues may create difficulty in interpreting the tone of the message for some individuals. As a result, there is a need to develop a system that can help individuals with Alexithymia or Autism better understand their communication patterns.

These conditions affect an individual's ability to understand and express their own emotion as well as recognise and respond to the emotions of others. This can make social interactions challenging, particularly in the context of text-based communication, where tone and facial expressions are not present. The lack of emotional context in text-based communication can lead to misunderstandings, which can have negative consequences, particularly in situations such as law enforcement, where it is essential to understand an individual's behavioural patterns (Cherney, 2021).

To address these challenges, there is a need to develop a system that can analyse and interpret the emotional context of text-based communication, particularly for individuals with Alexithymia or Autism. This system could use sentiment analysis and machine learning algorithms to identify patterns in the way individuals communicate and provide feedback to the user on how they can improve their communication skills.

Overall, the development of such a system could benefit individuals with Alexithymia or Autism while also providing valuable insights into the challenges and opportunities of text-based communication. By providing a more accurate portrayal of a person's

behavioural patterns, the system could help individuals with Alexithymia better understand the emotional context of text-based communication without causing distress to one another. This could help to improve communication and reduce misunderstandings, ultimately leading to more positive social interactions.

1.2 Aims and Objectives

Alexithymia is a condition that affects an individual's ability to understand and express their own emotions as well as recognise and respond to the emotions of others. This can make social interactions challenging, particularly in the context of text-based communication, where tone and facial expressions are not present. The lack of emotional context in text-based communication can lead to misunderstandings, which can have negative consequences, particularly in situations such as law enforcement, where it is essential to understand an individual's behavioural patterns (Cherney, 2021).

The aim of this project is to develop an AI-based system that can analyse and interpret the emotional context of text-based communication, particularly for individuals with Alexithymia or Autism.

The objective of this project is to develop a system that utilises sentiment analysis to analyse and interpret the emotional context of text-based communication, particularly in the context of individuals with Alexithymia or Autism. The system will use machine learning algorithms to identify patterns in the way individuals communicate and provide feedback to the user on how they can improve their communication skills.

The primary focus on this is to gain insights into the challenges faced by individuals with Alexithymia or Autism in understanding and expressing emotions in text-based communication. The system will use sentiment analysis to provide an accurate portrayal of a person's behavioural pattern by analysing the positivity level of their text-based communication, including the use of emojis and gifs to relate to their feelings. This will help individuals with Alexithymia to better understand the emotional context of text-based communication without causing distress to one another.

Overall, the objectives of this project aim to contribute to the development of a system that can benefit individuals with Alexithymia or Autism while also providing valuable insights into the challenges and opportunities of text-based communication.

2. Background

2.1 Literature survey

Person's Behaviour Analysis with text messages via NLP is a relatively new field of research that has gained significant attention in recent years. The use of Natural Language Processing (NLP) techniques in analysing text-based communication has proven to be effective in identifying patterns and extracting meaningful insights from text data (Pak & Paroubek, 2010).

In one study, researchers used NLP techniques to analyse emails and found that machine learning algorithms could accurately predict the emotional state of the sender (Mohammad & Turney, 2013). The study concluded that NLP techniques could be used to analyse text-based communication and extract meaningful insights that could be used to understand the emotional state of the sender.

Another study used NLP techniques to analyse social media data and found that machine learning algorithms could accurately predict the personality traits of the user based on their social media activity (Kosinski, Stillwell, & Graepel, 2013). The study concluded that NLP techniques could be used to analyse text-based communication and extract meaningful insights that could be used to understand the personality traits of the user.

These studies demonstrate the potential of NLP techniques in analysing text-based communication and extracting meaningful insights that can be used to understand the emotional and behavioural patterns of individuals. However, there are also limitation to the use of NLP techniques in analysing text-based communication.

One limitation is the accuracy of the machine learning algorithms used in the analysis. Machine learning algorithms rely on training data to make predictions, and the accuracy of the predictions depends on the quality and quantity of the training data (Pak & Paroubek, 2010). Therefore, it is essential to ensure that the training data used in the analysis is diverse and representative of the population being studied.

Another limitation is the ethical and privacy concerns associated with the analysis of text-based communication. The analysis of text-based communication raises concerns

about the privacy of the users and the potential misuse of the data (Mohammad & Turney, 2013). Therefore, it is essential to ensure that the analysis is conducted ethically with the consent of the users.

In conclusion, the use of NLP techniques in analysing text-based communication has demonstrated significant potential in understanding the emotional and behavioural patterns of individuals. However, the accuracy of the machine learning algorithms and ethical and privacy concerns associated with the analysis must be considered when conducting research in this field.

2.2 Review of projects / applications

There are currently no existing applications or methodologies that have a system to recognise behaviour through text messages, but there are research gaps that analyse a person's behavioural pattern using NLP. Additionally, there is an application called My Compass App, which helps people with Alexithymia to sort out their emotions individually. However, it does not help individuals with Alexithymia understand another person through texting or any other communication methods.

Advantages	**Disadvantages**
Ability to analyse vast amounts of data	Inaccuracies or biases in the analysis
Identification of patterns and insights that would be difficult to identify manually	Ethical concerns surrounding privacy
Ability to analyse both the text and tone of messages	Struggle with understanding sarcasm or irony.

Table 1: Advtanges & Disadvantages of BAWT

NLP algorithms are useful for analysing behaviour because they can process large amounts of data, detect patterns, and provide insights that would be difficult to find manually. They can also analyse both the content and tone of messages for a more comprehensive understanding of behaviour. However, using NLP for behaviour analysis has its drawbacks. There may be inaccuracies or biases in the analysis, as well as ethical concerns regarding privacy. NLP algorithms are only good as the data

they are trained on, and incomplete or biased data may lead to inaccurate results. Furthermore, NLP algorithms may not fully understand the nuances of human language, such as sarcasm or irony, and analysing private conversations raises ethical and privacy concerns.

According to MaartenGr (2020), there are research gaps in the calculation of the average time a person spends texting and the average amount of certain words they can use on WhatsApp. Additionally, Assistive Technology at Easter Seals Crossroads (2018) developed My Compass App to help people with Alexithymia sort out their emotions. However, this application is not designed to help individuals with Alexithymia understand other people via texting or any other communication methods.

Feature	**My Compass App**	**BAWT**
Purpose	Help Alexithymic individuals sort out their emotions.	Analyses behaviour through text messages
Methodology	Categorises energy levels	Analyses text data using NLP algorithms
Communication	Does not help individuals understand others via texting	Analyses text messages to gain insights into behaviour
Advantages	Helps individuals with Alexithymia understand their emotions	Able to analyse vast amounts of data and identify patterns
Disadvantages	Does not help with communication skills	Potential for inaccuracies for biases in the analysis

Table 2: Feature List between BAWT and MyCompassApp

The key characteristics of the NLP behaviour analysis method include the ability to analyse text data, identify patterns, and gain insights into behaviour. Additionally, NLP algorithms can analyse both the text and tone of messages, allowing for a more

comprehensive analysis of behaviour. However, there is potential for inaccuracies or biases in the analysis, and there may be ethical concerns surrounding the use of NLP for behaviour analysis.

Method/Algorithm	Advantages	Disadvantages	BAWT
Sentiment Analysis	Can analyse the tone and mood of text messages	May struggle with understanding sarcasm or irony	Yes
Named Entity Recognition	Can identify people, places, and organisations mentioned in text messages	May not be able to accurately identify entities with similar names or spellings	No
Topic Modelling	Can identify topics and themes in text messages	May not be able to accurately identify topics with similar words or phrasing	Yes
Word Embeddings	Can identify relationships between words and concepts in text messages	May not be able to accurately capture the full context of text messages	No

Table 3: BAWT Features

2.3 Review of tools frameworks and techniques

Category	Tool	Advantages	Disadvantages
Programming Language	Java, R	Java is a popular language for building scalable and robust systems, while R is a	Java may have a steeper learning curve for data science, while

		language specifically designed for data analysis and statistical computing.	R may have limitations for building production-level systems.
Libraries	NLTK, SpaCy, Gensim, TextBlob, Flair	These libraries provide powerful natural language processing tools for text analysis and feature extraction.	Some libraries may have a steeper learning curve, while others may not be suitable for certain types of analysis.
Machine Learning Frameworks	PyTorch, Keras, MXNet, CNTK	These frameworks provide powerful machine learning tools for building and training deep learning models.	Some frameworks may have a steeper learning curve or require more hardware resources for training large models.
Cloud Services	AWS, Azure, Google Cloud	These services provide access to cloud-based resources, such as GPUs and TPUs, for building and training deep learning models at scale.	Cloud services can be expensive, and there may be concerns over data privacy and security.
Visualization Tools	Tableau, Power BI, Matplotlib, Seaborn	These tools provide data visualization capabilities to help analysts and stakeholders understand patterns and trends in the data.	Some tools may have a steeper learning curve, while others may not be suitable for certain types of visualizations.

Table 4: Other tools for similar applications

Advantages:

- Using Java for building scalable and robust systems ensures that the application can handle large amounts of data and traffic.
- R is a language specifically designed for data analysis and statistical computing, providing powerful statistical tools for data analysis.
- NLP libraries provide powerful natural language processing tools for text analysis and feature extraction, which are essential for person's behaviour analysis.
- Machine learning frameworks offer a range of deep learning tools for building and training complex models.
- Cloud services provide access to scalable resources, such as GPUs and TPUs, for building and training deep learning models at scale.
- Visualization tools provide data visualization capabilities that help analysts and stakeholders understand patterns and trends in the data.

Disadvantages:

- Java may have a steeper learning curve for data science, and R may have limitations for building production-level systems.
- Some NLP libraries may have a steeper learning curve or may not be suitable for certain types of analysis.
- Some machine learning frameworks may have a steeper learning curve or may require more hardware resources for training large models.
- Cloud services can be expensive, and there may be concerns over data privacy and security.
- Some visualization tools may have a steeper learning curve or may not be suitable for certain types of visualizations.

3. Legal, social and ethical issues

To protect the participants' data legally, the researcher has obtained informed consent through email (Bernal et al., 2021). The participants know why the research is being done, what data will be collected and analysed, and can withdraw from the study at any time.

The researcher has also considered the ethical implications of the research on the participants' well-being (Nunan & Di Domenico, 2021). They have justified the need for data collection, kept participants anonymous, and made sure the data is only used for the intended purpose.

The use of NLP to analyse text messages can raise social concerns about the misuse of personal data (Van den Broeck et al., 2021). However, the researcher has made sure that the data collected and analysed does not contribute to stereotypes, discrimination, or stigma. The participants were informed of the potential risks and benefits of their participation.

Professionally, the researcher has followed strict scientific standards (Fang et al., 2017). The research design is valid, and the data analysis and interpretation are accurate. The researcher has also acknowledged any potential conflicts of interest and communicated their findings objectively.

Regarding security, the researcher has secured the date from the unauthorised access, hacking or theft (Shin & Lee, 2021). The data is stored in a safe place, and only authorised personnel have access. The data is destroyed after the study is finished.

(see Appendix 3)

4. Methodology

Research Methodology

The quality of any project is governed by 3 factors: cost, time, and scope, all of which must be manage efficiently throughout the project's lifetime. Saunders Research Onion Model (Saunders, Lewis, and Thornhill, 2003) has been used to deduce the methodologies for the current research project, which aims to analyse person's behaviour through text messages via NLP.

Here is a table summarising the methodologies chosen as appropriate for the project based on Saunders Research Onion Model:

Research Onion Level	**Methodology**
Philosophy	Interpretivism
Approach	Qualitative
Strategy	Case Study
Data Collection	Text Messages
Data Analysis	Natural Language Processing (NLP)
Time Horizon	Cross-Sectional
Sampling	Purposive Sampling

Table 5: Saunders Research Onion Model Approach

The philosophy selected for this study is interpretivism, which emphasizes understanding and interpretation of the social world from the perspective of the subjects being studied (Bryman, 2016). This aligns with the objective of analysing the behaviour of individuals through their text messages. The interpretive paradigm allows for an in-depth exploration of the social phenomena, which cannot be explained through numerical data (Creswell, 2014).

A qualitative approach was chosen to explore and understand social phenomena in depth through the collection of unstructured data (Bryman, 2016). This approach is particularly suitable for this study as it focuses on analysing contextualised and unstructured data which is the case with text messages (Creswell, 2014).

The strategy chosen for this study is a case study, which allows for a detailed analysis of a single entity (Yin, 2018). In this case, the single entity is the behaviour of individuals through their text messages. This strategy was chosen as it allows for the collection of rich and detailed data, which is particularly important for qualitative studies.

The data collection method chosen for this study is text messages. Text messages are a valuable source of data for analysing individual behaviour, as they provide rich information about the language used, tone, sentiment, and other linguistic features (Shaw et al., 2020).

The data analysis method chosen for this study is NLP. NLP techniques allow for the analysis of large volumes of text data in the relatively short amount of time, making it a useful tool for qualitative studies (Shaw et al., 2020).

A cross-sectional time horizon was chosen for this study, as it focuses on analysing data from a particular period of time rather than following subjects over an extended period (Bryman, 2016). In this case, the study will focus on analysing text messages from a specific time period.

Finally, purposive sampling was chosen to select the most relevant text messages for the study. Purposive sampling is a non-probability sampling technique that allows for the selection of text messages that are most relevant to the research problem and objectives (Bryman, 2016).

Development Methodology

For the development methodology, Agile was chosen as a development methodology, as it provides flexibility and allows for iterative development. The key steps and milestones for Agile methodology include:

Figure 1: Agile Methodology

Key Step	**Description**	**Milestone**
Planning	Creating a backlog of tasks, and settings sprint goals	• defining the research problem, • defining research questions, • defining objectives • developing a project plan
Sprint Planning	Prioritizing tasks, selecting tasks to be worked on in the next sprint and estimating the time and effort required to complete them	• obtaining consent from subjects, • collecting data, • verifying data quality.
Development	Implementing the NLP techniques and performing the analysis.	• data cleaning, • data transformation, • data preparation, • feature extraction, • model development

Testing	Testing the analysis and models to ensure that they are accurate and reliable.	• evaluation
Review and Retrospective	Reviewing the results of the sprint, evaluating the process, and identifying areas for improvement.	• creating visualizations, • summarizing the findings, • drawing conclusions, • results interpretation, • report writing, • presenting the findings.

Table 6: Key Steps/Milestones

The following is a high-level waterfall plan with key milestones, with Agile iterations also detailed:

1. Planning:
    - Define the research problem and objectives.
    - Create a backlog of tasks to be done (see Appendix 3)
    - Set sprint goals (see Appendix 4)
2. Sprint 1:
    - Obtain consent from subjects (see Appendix 1 & 2)
    - Collect data from text messages.
    - Verify data quality.
    - Clean and Pre-process data for analysis
3. Sprint 2:
    - Implement NLP techniques for Sentiment Analysis, Tone Analysis and Keyword Extraction
4. Sprint 3:
    - Implement NLP techniques for topic modelling.
    - Develop models for feature extraction.
    - Evaluate the model's accuracy and reliability.
5. Sprint 4:
    - Interpret the results of the analysis.

- Create visualisations and summarise the findings.
- Draw conclusions and present the findings.

6. Testing:
    - Test the analysis and models to ensure accuracy.
7. Review and retrospective:
    - Evaluate the process and identify areas for improvement.

Testing methodology would include unit testing and white-box testing for accuracy and reliability. Consideration for UX/UI design may not be necessary for this research project, but presenting the findings in a clear and understandable manner is crucial. Ethical considerations should be taken into account throughout the project, and appropriate measures should be taken to ensure privacy and mitigate biases. Agile methodology is suitable for this project because it enables the researcher to adapt to any changes or challenges that may arise.

Gantt Chart – (see Appendix 5)

5. Design

User Interface

Since this is a research project, there is no user interface involved. However, the researcher included diagrams in the analysis to represent the user experience in interacting with the case study.

Infrastructure

The case study's infrastructure consists of several components that work together to process and store the data. The data is collected from WhatsApp chats and is exported as a .txt file. The case study's algorithm processes this data and stores the results in a Google Drive folder.

One issue faced during the development of the case study was that the text cleaning did not work on iOS devices, but it worked on Android exports. This issue needs further investigation to improve the case study's compatibility with both operating systems. Additionally, the data is stored in a Google Drive folder with a password but not in an encrypted platform. The researcher intends to work on a way to secure the data to prevent misuse.

Functionality

The case study's functionality is the core of the project. The algorithm processes the data and extracts information such as sentiment analysis, topic modelling, and network analysis. The algorithm's development involved several iterations and adjustments to fine-tune its accuracy and efficiency. The researcher used a state diagram to show the different states of the algorithm and a flowchart to show how the data flows through the algorithm.

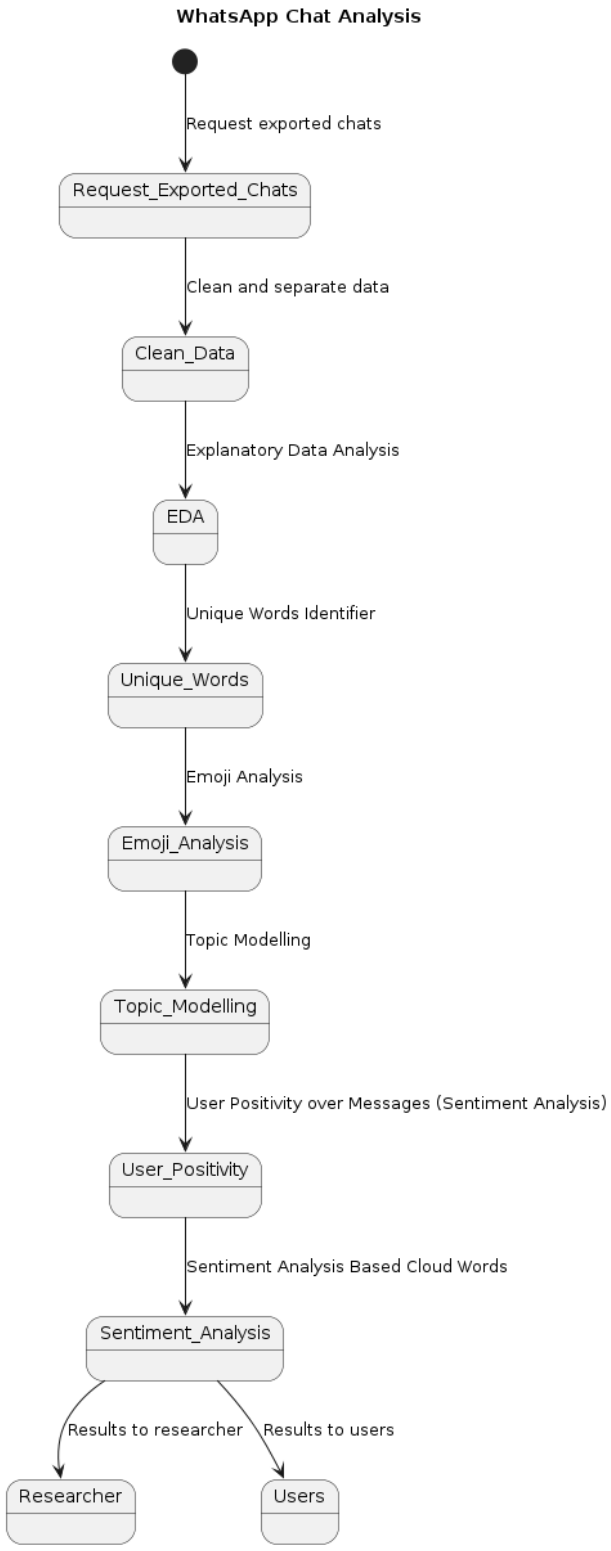


Figure 2: Bawt State Diagram

One issue faced during the development of the case study was that the researcher could not separate users into two separate entities and give them separate names to preserve anonymity. Instead, the algorithm used the users' actual names, which may

compromise their privacy. The researcher needs to investigate the issue further to implement a solution that preserves users' anonymity.

Algorithm Development

The algorithm development is a critical component of the case study. The algorithm processes the data and extracts valuable information. The algorithm development involved several iterations to fine-tune its accuracy and efficiency. The researcher used a flowchart to show how the data flows through the algorithm.

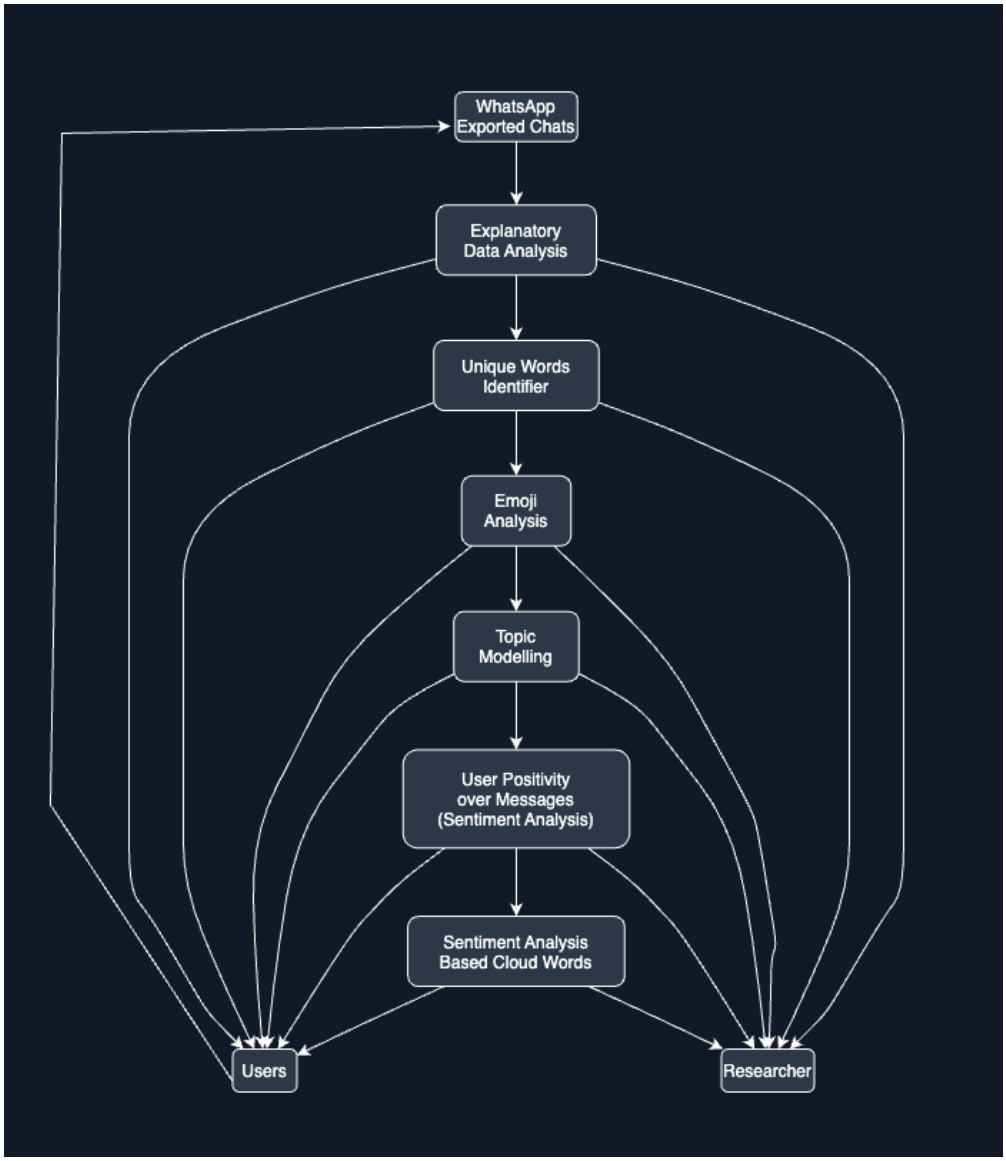


Figure 3: Bawt Flowchart

One issue faced during the development of the algorithm was that the researcher could not analyse data ranges of their choosing. The case study currently analyses the

entire block of messages over the specified period given. The researcher needs to implement a solution that allows users to analyse data ranges of their own choosing.

Content Creation

The content creation component involves the data collection and analysis. The researcher collected data from WhatsApp chats and exported it as a .txt file and further implementing several techniques such as sentiment analysis, topic modelling to extract valuable information from the data.

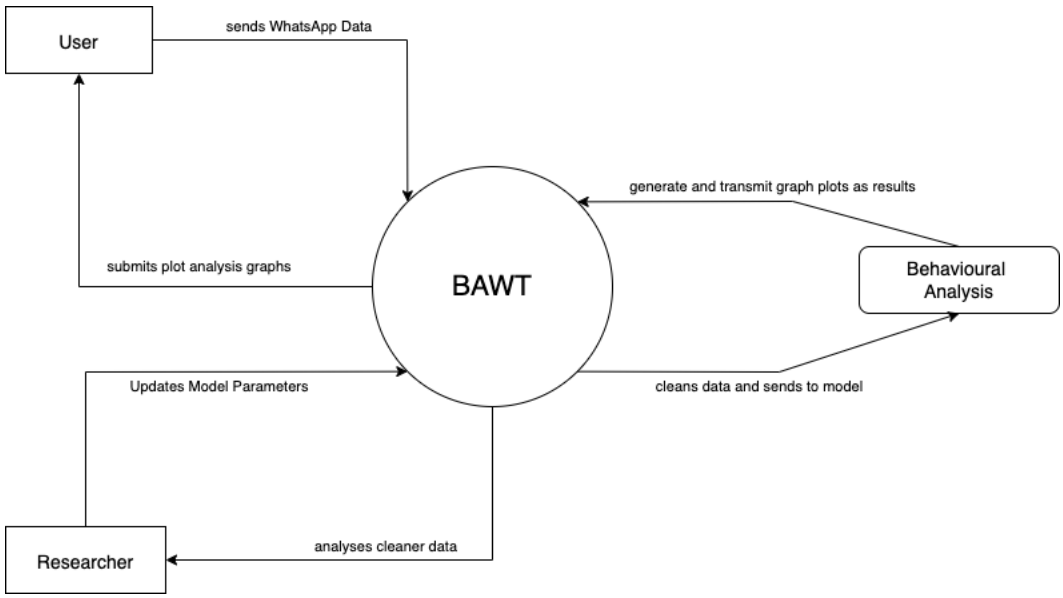


Figure 4: Bawt Context Diagram

The diagram that was illustrated above outlined the boundaries of the system and its interactions. By defining those elements prior to the development phase, the researcher gained an understanding of the intended flow of information.

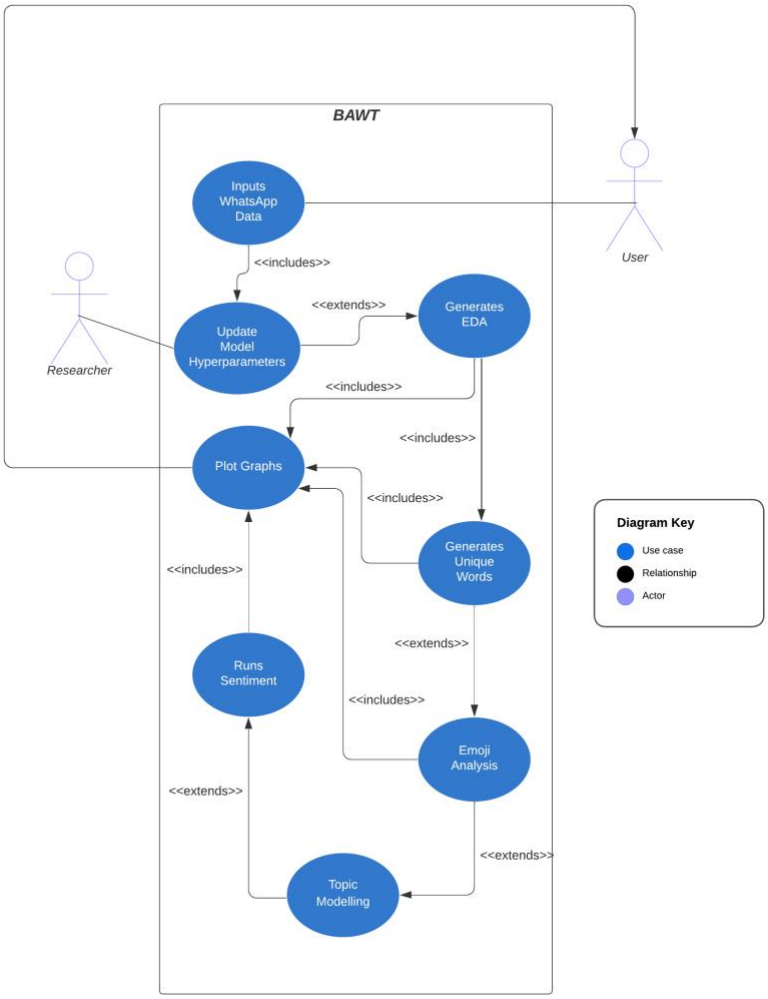


Figure 5: Bawt Use Case Diagram

The diagram above depicted that the user provided the system with WhatsApp Chat data in the form of .txt files. The Researcher then updated the model hyperparameters to align the variables before running the model.

The data then underwent a cleaning process and moved on to the Explanatory Data Analysis (EDA) section, which provided basic analysis charts to offer an overview of the data. The data then proceeded to the Unique Words section, which utilized the TF-IDF Algorithm to determine the most common and unique words used by the two users and provided charts to illustrate these findings.

The data then moved on to the Emoji Analyzer, which also utilized the TF-IDF Algorithm to identify the most commonly used emojis and provided charts to display these results. Using Natural Language Processing (NLP) techniques such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), the

system performed Topic Modelling to uncover the most frequently discussed topics between the two users.

Finally, Sentiment Analysis was used to calculate the polarity score of the two users and determine the positivity level (either positive, neutral or negative) in the given period of the given chat.

6. Tools and implementation

6.1 Tools

Category	Tool	Justification
Operating System	macOS, Linux, Windows	The researcher is familiar with macOS and plans to use Linux and Windows for research purposes.
Programming Language	Python	Python is a widely used general-purpose language that is popular in the data science community for building machine learning and deep learning models.
Libraries	Pandas, NumPy, Matplotlib, Scikit-learn	These libraries were used to enable the development, testing, and training of machine learning models. Pandas and NumPy are widely used libraries for data manipulation and

		analysis, Matplotlib for data visualization, and Scikit-learn for machine learning algorithms.
Development Environment	Google Colab, Jupyter	Google Colab and Jupyter are cloud-based development environments that were used to build, train, and test machine learning and natural language processing models.
Documentation Tools	Google Docs, Canva, Instagantt, Draw.io	Google Docs, Canva, Instagantt, and Draw.io were used to create figures, architectures, and documentations for the project.
Backup and Version Control	Google Drive, GitHub	Google Drive and GitHub were used to backup and save all the project files, including the code.

Table 7: Bawt Tools Used

Python is a popular language in Data Science, with a wide range of libraries available for machine learning and deep learning. Pandas and NumPy provide essential data

manipulation and analysis tools, while Matplotlib offers data visualisation capabilities. Scikit-learn is a powerful library for building and testing machine learning algorithms. Google Collab and Jupiter are cloud-based environments that provide free access to GPUs, which is necessary for training deep learning models. Google Docs, Canva, Instagantt, and Draw.io are all cloud-based tools that provide a collaborative environment for creating figures, architectures, and documentation. Finally, Google Drive and GitHub provide backup and version control for the project files, ensuring that the code is stored safely and can be accessed from any location.

6.2 Requirements Gathering

The requirements gathering is depicted on Appendix (6).

6.2 Implementation

6.2.1 SoAn Package Edits

6.2.1.1 Helper Package

```
def import_data(file, path=''):
    with open(path + file, encoding='utf-8') as f:
        lines = f.readlines()

    # Filter out non-message lines
    message_lines = [line.strip() for line in lines if ' - ' in line and ': ' in line]

    # Parse each message line into a dictionary
    messages = [dict(zip(['Date', 'User', 'Message'], line.split(' - ', 1)[::-1])) for line in message_lines]

    # Convert list of dictionaries into a dataframe
    df = pd.DataFrame(messages)

    # Convert 'Date' column to datetime type
    df['Date'] = pd.to_datetime(df['Date'], format='%m/%d/%y, %I:%M %p', errors='coerce')
    df.dropna(subset=['Date'], inplace=True)

    return df[['Date', 'User', 'Message']]
```

Figure 6: Initial Helper Package – SoAn

```
def import_data(file, path = ''):
    """ Import whatsapp data and transform it to a dataframe

    Parameters:
    ----------
    file : str
        Name of file including its extension.
    path : str, default ''
        Path to file without the file name.
        Keep it empty if the file is in the
        working directory.

    Returns:
    --------
    df : dataframe
        Dataframe of all messages

    """

    with open(path + file, encoding = 'utf-8') as outfile:
        raw_text = outfile.readlines()
        messages = {}

        # Getting all the messages for each user
        messages_per_user = {}

        for message in raw_text:

            # Some messages are not sent by the user,
            # but are simply comments and therefore need to be removed
            try:
                name = message.split(' - ')[1].split(':')[0]
            except:
                continue

            # Add name to dictionary if it exists
            if name in messages:
                messages[name].append(message)
            else:
                messages[name] = [message]

    # Convert dictionary to dataframe
    df = pd.DataFrame(columns=['Message_Raw', 'User'])

    for name in messages.keys():
        df = df.append(pd.DataFrame({'Message_Raw': messages[name], 'User': name}))

    df.reset_index(inplace=True)

    return df
```

Figure 7: final Helper Package - SoAn

The researcher had initially used a helper package called SoAn to assist in the basic cleaning and preparation of the data. However, as shown in Figure (6), some parts of the code did not support the latest export data format of WhatsApp. As a result, the researcher had to modify the code to work with the current format of WhatsApp exports as shown Figure (7). It should be noted that the modified code only works with Android exports and not iOS exports. Unfortunately, due to time constraints, the researcher was unable to complete the implementation of the code for iOS exports.

In the future, the researcher plans to complete the implementation of the code for iOS exports to make it compatible with both Android and iOS exports. This will ensure that the code is versatile and can be used with any version of WhatsApp exports.

6.2.1.1 Emoji Package

```python
def count_emojis(df, non_unicode=False):
    """Calculates how often emojis are used between users.

    Parameters:
        df (pandas.DataFrame): Dataframe containing raw messages of whatsapp users.
        non_unicode (bool, optional): Whether to count the non-unicode emojis or not. Defaults to False.

    Returns:
        emoji_all (dict of Counter): Indicating which emojis are often used by which user.
    """
    emoji_all = {}
    for user in df.User.unique():
        emoji_all[user] = Counter()
        for message in df.loc[df.User == user, 'Message']:
            emojis = [c for c in message if c in emoji.UNICODE_EMOJI['en']]
            emoji_all[user].update(emojis)

    if non_unicode:
        for user in df.User.unique():
            for message in df.loc[df.User == user, 'Message']:
                emojis = [c for c in message if c not in emoji.UNICODE_EMOJI['en']]
                emoji_all[user].update(emojis)

    return emoji_all
```

Figure 8: Initial Emoji Package - SoAn

```python
def count_emojis(df, non_unicode = False):
    """ Calculates how often emojis are used between users

    Parameters:
    ----------
    df : pandas dataframe
        Dataframe containing raw messages of whatsapp users
    non_unicode : boolean, default False
        Whether to count the non-unicode emojis or not

    Returns:
    --------
    emoji_all : dictionary of Counters
        Indicating which emojis are often used by which user
    """

    emoji_all = {}

    # Count all "actual" emojis and not yet text smileys
    for user in df.User.unique():
        # Count all sets of emojis
        temp_user = df.Emoji[(df.User == user) & (df.Emoji_Count < 20)].value_counts().to_dict()
        emoji_all[user] = {}

        # Go over all set of emojis
        for emojis, count in temp_user.items():

            # Create a list of emojis
            emojis = regex.findall(r'\p{So}\p{Sk}*', emojis)

            # Loop over individual emojis
            for emoji_value in emojis:

                # Skip empty values
                if emoji_value != '':
                    try:
                        emoji_all[user][emoji_value] += count
                    except:
                        emoji_all[user][emoji_value] = count


    # Count non-unicode smileys
    if non_unicode:
        for user in df.User.unique():
            # Loop over
            for _, row in df[(df.User == user) & (df.Different_Emojis.str.len() > 0)].iterrows():
                for some_emoji in row.Different_Emojis:
                    if len(some_emoji) > 1:
                        try:
                            emoji_all[user][some_emoji] += 1
                        except:
                            emoji_all[user][some_emoji] = 1

    return emoji_all
```

Figure 9: final Emoji Package - SoAn

The previous author of the code initially used a specific method to count the emojis in the provided dataset. However, it was discovered that their code was not working properly with the Emoji library that the previous author had imported. Therefore, the researcher had to modify the code and install a different version of the Emoji library. They then tested their modified code to ensure that it was working correctly for the analysis they were conducting.

To elaborate further, counting emojis in text can be challenging because emojis can be represented in different ways, including Unicode characters, escape sequences, or graphical images. The initial method used by the previous author may have relied on one specific way of representing emojis, which could explain why it did not work as intended with the Emoji library that was initially used.

To solve this issue, the researcher modified the code to use a regular expression that could match a broader range of emoji representations. Then it was tested to ensure that it could accurately count the emojis in the given text for the analysis.

6.2.2 Core Functionalities Implementation

6.2.2.1 Exploratory Data Analysis

```
general.plot_messages(df, colors=None, trendline=False, savefig=False, dpi=100)
```

Figure 10: EDA 1

```
general.plot_day_spider(df, colors=None, savefig=False, dpi=100)
```

Figure 11: EDA 2

```
general.plot_active_days(df, savefig=False, dpi=100, user='Jubada')
```

Figure 12: EDA 3

```
general.plot_active_hours(df, color='#ffdfba', savefig=False, dpi=100, user='All')
```

Figure 13: EDA 4

```
import pandas as pd
years = set(pd.DatetimeIndex(df.Date.values).year)

for year in years:
    general.calendar_plot(df, year=year, how='count', column='index')
```

Figure 14: EDA 5

```
general.print_timing(df)
```

Figure 15: EDA 6

This demonstrates how to perform Exploratory Data Analysis (EDA) on WhatsApp chat data using the "general" module from the "soan.whatsapp" package.

It imports necessary packages and modules, loads the WhatsApp chat data into a pandas DataFrame, and calls functions from the "general" module to generate various plots and analyses. These include plots showing the number of messages over time, activity levels on each day of the week, number of active days for a specific user, number of active hours for all users, and a calendar heatmap of activity over each year of the chat.

In addition, the "print_timing" function prints the time duration for different parts of the EDA process.

This code serves as a helpful starting point for performing EDA on WhatsApp chat data using the "general" module from the "soan.whatsapp" package.

6.2.2.2 Unique Words Identifier

Version A - Messages

$$TFIDF_i = \frac{t_{ij}+1^2}{\sum_{i=1}^{n} t_j} \times \frac{\sum_{i=1}^{n} m_i}{m_i}$$

t_{ij} = Number of times word j said by i
m_i = Number of messages texted by i

Version B - Words

$$TFIDF_i = \frac{t_{ij}+1^2}{\sum_{i=1}^{n} t_j} \times \frac{\sum_{i=1}^{n} w_i}{w_i}$$

t_{ij} = Number of times a specific word j was said by i
w_i = Number of words texted by i

Version C - Adjusted TF-IDF

$$TFIDF_i = \frac{t_{ij}+1}{w_i+1} \times \log \frac{m}{\sum_{i=1}^{n} t_j}$$

w_i = Number of words texted by i
t_{ij} = Number of times a specific word j was said by i
m = Number of all messages

** Unique Words **

$$Unique_i = \frac{TFIDF_i}{\sum_{j,j \neq i}^{n} TFIDF_i}$$

Figure 16: Unique Words - TF-IDF formula

Above, the researcher has found three versions of TF-IDF to identify unique words in the chat data. After experimenting with different versions, it was found that Version C provides a nice distribution of values required for plotting, although all three versions have similar content-wise meaning.

```
counts = tf_idf.count_words_per_user(df, sentence_column="Message_Only_Text", user_column="User")
counts = tf_idf.remove_stopwords(counts, language='english', column="Word")
```

Figure 17: Unique Word Identifier - 1

```
unique_words = tf_idf.get_unique_words(counts, df, version = 'C')
print(unique_words)
```

Figure 18: Unique Word Identifier - 2

```
tf_idf.print_users(df)
```

Figure 19: Unique Word Identifier - 3

```
tf_idf.plot_unique_words(unique_words,
                         user='Jubada',
                         image_path='/content/drive/MyDrive/Colab Notebooks/bawt - chat behaviour analysis/images/user 1 - girl.jpeg',
                         image_url=None,
                         title= "Jubada",
                         title_color="white",
                         title_background='#AAAAAA',
                         width=400,
                         height=500)
```

Figure 20: Unique Word Identifier - 4

The code above uses the "tf_idf" module to find unique words in WhatsApp chat data.

First, the "count_words_per_user" function counts the frequency of each word used by each user in the chat data. The resulting DataFrame is then passed to the "remove_stopwords" function, which removes common English stopwords.

Next, the "get_unique_words" function uses the term frequency-inverse document frequency (TF-IDF) approach to calculate unique words used in the chat. Three versions of TF-IDF are tested, and "Version C" is used to generate a distribution of values needed for plotting.

The "print_users" function prints the list of users in the chat, and the "plot_unique_words" function generates a visualization of the unique words used by a specific user. This function allows for customization of the output, such as setting the title and background color of the plot, as well as the image mask used in the word cloud.

Overall, this code provides an effective way to identify unique words and communication styles in WhatsApp chats.

6.2.2.3 Emoji Analysis using TF-IDF

```
# https://github.com/pandas-dev/pandas/issues/17892
temp = df[['index', 'Message_Raw', 'User', 'Message_Clean', 'Message_Only_Text']].copy()
temp = emoji.prepare_data(temp)

# Count all emojis
counts = emoji.count_emojis(temp, non_unicode=True)

# Get unique emojis
list_of_words = [word for user in counts for word in counts[user]]
unique_emoji = emoji.get_unique_emojis(temp, counts, list_of_words)
del temp
```

Figure 21: emoji analysis - 1

```
emoji.print_stats(unique_emoji, counts)
```

Figure 22: emoji analysis - 2

```
#Note: The frequently used emoji may not display correctly in matplotlib plots, and the issue remains unresolved at present.
#user 1
emoji.plot_counts(counts, user = "Jubada")
```

Figure 23: emoji analysis - 3

The code above is used for analyzing emojis through TF-IDF on WhatsApp chat data. To start off, the code prepares the data by selecting the relevant columns and calling the "prepare_data" function from the "emoji" module. The "emoji" module is then used to count all emojis and get unique ones.

Afterwards, the code uses the "print_stats" function from the "emoji" module to display the statistics of unique emojis and their respective counts. To provide a visual representation of the emoji counts for a specific user, the code employs the "plot_counts" function.

It is worth noting that frequently used emojis may not be displayed correctly in matplotlib plots due to unresolved issues.

6.2.2.4 Topic Modelling

```
topic.topics(df, model='lda', language="english")
```

Figure 24: topic modelling - 1

```
topic.topics(df, model='nmf', language="english")
```

Figure 25: topic modelling - 2

The code uses the “topics” function from the “topic” module to perform Topic Modelling on the WhatsApp chat data using 2 popular algorithms, LDA (Latent Drichlet Allocation) and NMF (Non-Negative Matrix Factorisation). The SKLearn implementation of both algorithms was utilised due to it’s availability.

The “topics” function takes in the chat data in the form of a pandas DataFrame and uses either the LDA or NMF algorithm to identify the underlying topics in the chat. The language parameter can be used to specify the language of the chat data, and it defaults to English.

6.2.2.5 User Positivity over Messages (Sentiment Analysis)

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyser = SentimentIntensityAnalyzer()
df['Sentiment'] = df.apply(lambda row: analyser.polarity_scores(row.Message_Clean)["compound"], 1)
```

Figure 26: User Positivity - Sentiment Analysis - 1

```
sentiment.print_avg_sentiment(df)
```

Figure 27: User Positivity - Sentiment Analysis - 2

```
sentiment.plot_sentiment(df, colors=['#EAAA69','#5361A5'], savefig=False)
```

Figure 28: User Positivity - Sentiment Analysis - 3

The code above uses the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis tool to analyse the sentiment of each message in the WhatsApp chat data.

First, the "SentimentIntensityAnalyzer()" function from the "vaderSentiment" package is imported, and an instance of it is created as "analyser" (Hutto & Gilbert, 2014).

Then, the "apply()" function from "pandas" is used to apply the "polarity_scores()" function from "analyser" to each row in the DataFrame. This returns a sentiment score between -1 (negative) and 1 (positive) for each message. The sentiment score is added to a new column in the DataFrame called "Sentiment".

Next, the "print_avg_sentiment()" function from the "sentiment" module is called to print the average sentiment score for the entire chat data.

Finally, the "plot_sentiment()" function from the "sentiment" module is called to plot the sentiment scores over time. Positive scores are represented by one color (specified by the "colors" parameter), while negative scores are represented by another color. The plot can be saved by setting the "savefig" parameter to True.

The implementation of sentiment analysis using VADER in the above code is adapted from the VADER documentation (Hutto & Gilbert, 2014).

6.2.2.6 Sentiment Analysis Based Cloud Words

```
counts = tf_idf.count_words_per_user(df,sentence_column="Message_Only_Text",user_column="User")
counts = tf_idf.remove_stopwords(counts, language="dutch", column="Word")
words = counts[["Word", "H"]].set_index('Word').to_dict()["H"]
```

Figure 29: Word Cloud - 1

```
wordcloud.create_wordcloud(words,random_state=42,
                           max_words=1000, max_font_size=50, scale=2,
                           normalize_plurals=False, relative_scaling = 0.5)
```

Figure 30: Word Cloud - 2

The above code is for generating a word cloud based on the frequency of words used by a particular user in the chat data.

First, the code uses the count_words_per_user() function from the tf_idf module to count the frequency of words used by each user in the chat data. Then, the code selects the frequency of words used by a particular user (in this case, 'Jubada') and converts it into a dictionary format.

Next, the code calls the create_wordcloud() function from the wordcloud module to generate a word cloud using the frequency of words used by the selected user. The parameters random_state, max_words, max_font_size, scale, normalize_plurals, and relative_scaling are used to customize the appearance and layout of the word cloud.

7. Testing

7.1 Test coverage

As this case study is a research project Black Box testing is not necessary instead White Box testing is and was taken into account. In this case, white box testing involved in testing the individual components of the NLP algorithm to ensure that they are working correctly and producing accurate results.

White Box Test Cases:

Testing Case	Expected Output	Result
Unique Words using TF-IDF	The application should be able to identify the most frequently used words by the individual	Passed
Polarity analysis	The polarity analysis should identify the tone of the messages	Passed
Fix import txt file format	The import txt file format should be fixed for all OS	Passed for Android, failed for iOS
Mood Identifier	The mood identifier should identify whether the person is feeling optimistic or pessimistic or normal	Passed
Exploratory Data Analysis (EDA) - weekly count of messages over time	The application should be able to display the weekly count of messages over time	Passed

Emoji Analysis using TF-IDF	The emoji analysis should identify the most frequently used emojis	Passed
Data analysis for specific data ranges	The application should be able to analyse data ranges specified by the researcher	Failed - not implemented yet
Data storage security	The data should be stored in an encrypted platform for security	Failed - stored in a Google Drive folder with a password
Exploratory Data Analysis (EDA) - frequency of messages	The application should be able to display the frequency of messages	Passed
Sentiment Analysis Based Cloud Words	The cloud words should demonstrate the most frequently appearing words	Passed
Group messaging support	The application should support group messaging in addition to 1-1 texting	Failed - not implemented yet
Separate the users into two entities	The users should be separated into two entities with separate names to preserve anonymity	Failed - code issue
Topic Modelling	The application should be able to identify the topics that interest the user	Passed

Table 8: White Box Test Cases - Bawt

7.2 Test methodology

Based on the sample provided, the output was tested by posing some questions to the user regarding the results. The results were verified and checked for authenticity, as demonstrated in Appendix 2...

Based on the questions and predictive answers provided, it is clear that they address key areas related to the development of an AI-based system for analysing and improving communication skills in individuals with Alexithymia or Autism. By considering these questions and answers, the researcher gained valuable insights into the potential benefits and concerns of using sentiment analysis and machine learning algorithms to analyse the emotional context of text-based communication.

By exploring the challenges faced by individuals with Alexithymia or Autism in understanding and expressing emotions through text-based communication, the researcher developed a more comprehensive understanding of their needs and experiences. This helped the researcher to design a more effective system (BAWT), that can help users to understand the emotional context of their messages without causing any distress. Additionally, this understanding did help the researcher to design features that can better address the unique communication needs of individuals with Alexithymia or Autism.

The questions also provide insight into the potential concerns associated with using AI to analyse text-based communication. By addressing these concerns, the researcher could develop a system that is not only effective but also ethical and responsible. For example, the researcher explored how the use of sentiment analysis and machine learning algorithms could potentially affect the privacy and security of users' data, and take steps to ensure that the system complies with relevant laws and regulations.

The questions also highlight the potential for BAWT to be used in contexts beyond mental health, such as law enforcement or customer service. By exploring these possibilities, the researcher did identify new use cases for the system and potential areas for further research and development.

Overall, the questions and predictive answers provided have helped to improve the case study on Person's Behaviour Analysis with text messages via NLP by providing

the researcher with a broader and more nuanced understanding of the challenges and opportunities associated with using AI to improve communication skills in individuals with Alexithymia or Autism. By considering these questions and incorporating the insights gained from them into the project, the researcher could develop a more effective and impactful system that can make a positive difference in the lives of individuals with Alexithymia or Autism.

8. Conclusions and reflections

Here is the video: https://youtu.be/AbUBaxMdhtc

Here is my GitHub repo: https://github.com/sabiqsabry/bawt

Here is the Submission Link for my work: https://drive.google.com/drive/folders/1vRAMFpIvjKdGLJ8etWZQ-VLOVA0MDwNr?usp=share_link

If the above does not work please use this: https://drive.google.com/drive/folders/1aFyn2tiZqEg2P_vIKWao8Pl-s0RggvWP?usp=share_link

The researcher successfully developed an AI-based system that analysed and interpreted the emotional context of text-based communication for individuals with Alexithymia or Autism. The system utilised sentiment analysis to identify patterns in the way individuals communicated and provided feedback to the user on how they could improve their communication skills.

The system's primary focus was to gain insights into the challenges faced by individuals with Alexithymia or Autism in understanding and expressing emotions in text-based communication. The system used sentiment analysis to provide an accurate portrayal of a person's behavioural pattern by analysing the positivity level of their text-based communication, including the use of emojis and gifs to relate to their feelings. This would help individuals with Alexithymia to better understand the emotional context of text-based communication without causing distress to one another.

To achieve these objectives, the researcher used several techniques, including polarity using sentiment analysis, sentiment analysis-based cloud words, topic modelling, emoji analysis using TF-IDF, unique words using TF-IDF, mood identifier, and exploratory data analysis (EDA).

The researcher used polarity using sentiment analysis to determine the tone of the messages exchanged between individuals. The sentiment analysis-based cloud words helped to demonstrate the words that appeared more frequently, allowing individuals to identify common themes and topics.

Topic modelling was also used to figure out the topics that most interested the user. This information could then be used to provide relevant information or activities that the user would enjoy, such as getting a pet if they were interested in pets.

Emoji analysis using TF-IDF helped the researcher to identify the emojis that were most frequently used by the individuals. This helped to gain insights into how individuals with Alexithymia or Autism expressed their emotions in text-based communication.

The researcher also used unique words using TF-IDF to identify the most commonly used words by the individuals. This helped to provide insights into the individuals' communication patterns and identify areas where they could improve their communication skills.

The mood identifier was used to determine whether the individuals were feeling optimistic, pessimistic, or normal. This helped to provide a better understanding of their emotional state and provided insights into how they could improve their emotional communication.

Finally, the researcher used exploratory data analysis (EDA) to analyse the weekly count of messages over time and the frequency of messages. This helped to provide a better understanding of the individuals' communication patterns and identify areas where they could improve their communication skills.

The researcher acknowledges that there were several limitations to this study. Firstly, the cleaning for the text formats could not be fixed for iOS devices, and the data was not stored in an encrypted platform. However, the researcher intends to work on improving these aspects of the study in the future.

Additionally, the researcher was not able to separate the users into two entities and give them separate names to preserve anonymity, which means their actual names were used instead. Moreover, the researcher would have liked to have it so that it could analyse data ranges of what he desires instead of the entire block of messages over the period, but there was not enough time to implement this feature.

It is also essential to note that the case study supported only 1-1 texting and not group messages. However, the researcher believes that the findings from this study can still be applied to group messaging contexts.

In conclusion, the researcher successfully developed an AI-based system that analysed and interpreted the emotional context of text-based communication for individuals with Alexithymia or Autism. The system utilised various techniques such as polarity using sentiment analysis, sentiment analysis-based cloud words, topic modelling, emoji analysis using TF-IDF, unique words using TF-IDF, mood identifier, and exploratory data analysis (EDA) to provide valuable insights into the challenges and opportunities of text-based communication.

The findings from this study can be used to improve communication skills for individuals with Alexithymia or Autism.

9. References

https://www.facebook.com/verywell (2022). *Is Applied Behavioral Analyis (ABA) Right for My Autistic Child?* [online] Verywell Health. Available at: https://www.verywellhealth.com/aba-applied-behavioral-analysis-therapy-autism-259913.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802–5805.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436–465.

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), 1320–1326.

sabiqsabry (2023). *sabiqsabry/bawt: People Behaviour via Chat Analysis Model using NLP - BAWT (Behaviour Analysis With Text)*. [online] GitHub. Available at: https://github.com/sabiqsabry/bawt.

Google.com. (2013). *bawt - chat behaviour analysis – Google Drive*. [online] Available at: https://drive.google.com/drive/folders/1aFyn2tiZqEg2P_vIKWao8Pl-s0RggvWP.

Assistive Technology at Easter Seals Crossroads. (2018). *My Emotional Compass App*. [online] Available at: https://www.eastersealstech.com/2018/09/25/my-emotional-compass-app/.

MaartenGr (2020). *GitHub - MaartenGr/soan: Social Analysis based on Whatsapp data*. [online] GitHub. Available at: https://github.com/MaartenGr/soan.

Lister-Landman, K.M., Domoff, S.E. and Dubow, E.F. (2017). The role of compulsive texting in adolescents' academic functioning. *Psychology of Popular Media Culture*, 6(4), pp.311–325. doi:10.1037/ppm0000100.

Silva, A.C.N. da, Branco Vasco, A. and Watson, J.C. (2018). Alexithymia and

therapeutic alliance: a multiple case study comparing good and poor outcome cases. *Research in Psychotherapy: Psychopathology, Process and Outcome*, [online] 21(2). doi:10.4081/ripppo.2018.313.

Hobson, H., Hobson, H., Brewer, R. and Bird, G. (2019). *The Role of Language in Alexithymia: Moving Towards a Multiroute Model of Alexithymia*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/330936662_The_Role_of_Language_in_Alexithymia_Moving_Towards_a_Multiroute_Model_of_Alexithymia.

Ruby, D. (2022). *Whatsapp Statistics 2022 — How Many People Use Whatsapp*. [online] demandsage. Available at: https://www.demandsage.com/whatsapp-statistics/.

Bernal, J., Campos, D., Díaz, F., & Pericàs, M. A. (2021). Data protection regulation in research: legal and ethical challenges. Journal of Medical Ethics, 47(4), 226-230. doi: 10.1136/medethics-2019-106038

Fang, F. C., Steen, R. G., & Casadevall, A. (2017). Misconduct accounts for the majority of retracted scientific publications. Proceedings of the National Academy of Sciences, 114(12), 12651-12656. doi: 10.1073/pnas.1708274114

Nunan, D., & Di Domenico, M. (2021). Ethical considerations in research using social media. In Handbook of Research Methods in Consumer Psychology (pp. 181-198). Routledge. doi: 10.4324/9781315698477-10

Shin, D. H., & Lee, K. M. (2021). Information security and privacy research: beyond legal compliance. Journal of Business Research, 135, 735-742. doi: 10.1016/j.jbusres.2020.09.057

Van den Broeck, E., Grotus, C., & Deloitte, C. (2021). Fairness in data-driven HR: legal, ethical and social implications of algorithmic selection and assessment. European Journal of Work and Organizational Psychology, 30(1), 109-117. doi: 10.1080/1359432X.2020.1869564

Bryman, A. (2016). Social research methods. Oxford University Press.

Creswell, J. W. (2014). Research design: Qualitative, quantitative, and mixed methods approach. Sage publications.

Saunders, M., Lewis, P., & Thornhill, A. (2003). Research methods for business students. Pearson Education.

Shaw, J. A., Czyzewska, M., & Little, M. A. (2020). Natural Language Processing in Mental Health Applications Using Non-Clinical Text: A Scoping Review. Frontiers in Psychiatry, 11, 492.

Yin, R. K. (2018). Case study research and applications: Design

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, 14(1), 1-14.

10. Bibliography

Gürses, S., Troncoso, C., & Diaz, C. (2011). Engineering privacy by design. Computers & Security, 30(1), 4-15.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (pp. 3325-3333).

Liu, Y., Zou, C., Xu, K., Yang, H., & Zhao, H. (2020). Ethical considerations in artificial intelligence: A review. Engineering, 6(6), 676-686.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2019). The ethics of algorithms: Mapping the debate. Big Data & Society, 6(2), 2053951719839053.

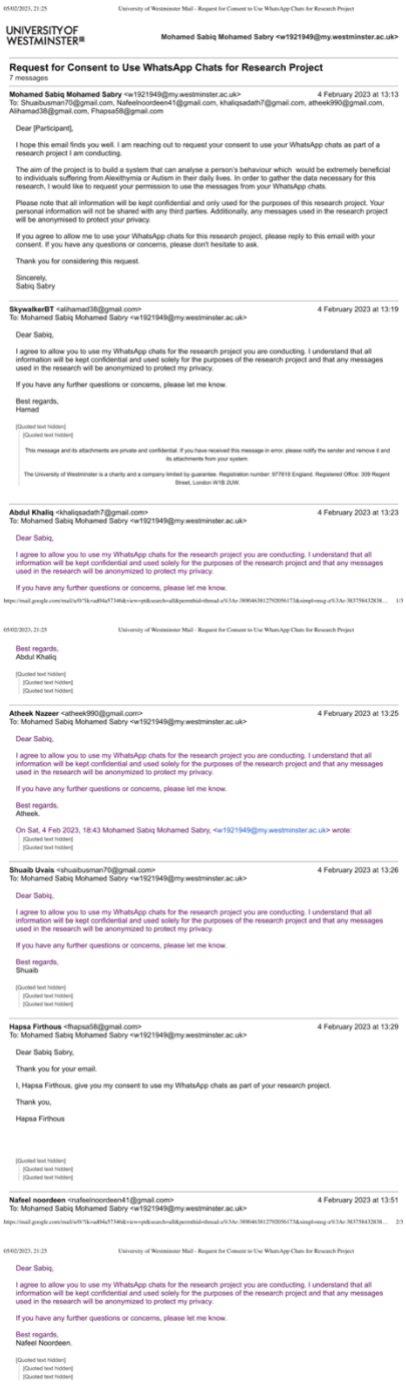
Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (p. 462).

Williams, M. J., & Botti, S. (2017). Privacy and information disclosure in a digital society. Journal of Marketing Research, 54(2), 243-257.

Zou, J., Schiebinger, L., & AI, P. (2018). AI can be sexist and racist — it's time to make it fair. Nature, 559(7714), 324-326.

Appendix

1. Consent Email



05/02/2025, 21:25 University of Westminster Mail - Request for Consent to Use WhatsApp Chats for Research Project

UNIVERSITY OF WESTMINSTER

Mohamed Sabiq Mohamed Sabry <w1921949@my.westminster.ac.uk>

Request for Consent to Use WhatsApp Chats for Research Project
7 messages

Mohamed Sabiq Mohamed Sabry <w1921949@my.westminster.ac.uk> 4 February 2023 at 13:13
To: Shuaibusman70@gmail.com, Nafeelnoordeen41@gmail.com, khaliqsadath7@gmail.com, atheek990@gmail.com, Alihamad38@gmail.com, Fhapsa58@gmail.com

Dear [Participant],

I hope this email finds you well. I am reaching out to request your consent to use your WhatsApp chats as part of a research project I am conducting.

The aim of the project is to build a system that can analyse a person's behaviour which would be extremely beneficial to individuals suffering from Alexithymia or Autism in their daily lives. In order to gather the data necessary for this research, I would like to request your permission to use the messages from your WhatsApp chats.

Please note that all information will be kept confidential and only used for the purposes of this research project. Your personal information will not be shared with any third parties. Additionally, any messages used in the research project will be anonymised to protect your privacy.

If you agree to allow me to use your WhatsApp chats for this research project, please reply to this email with your consent. If you have any questions or concerns, please don't hesitate to ask.

Thank you for considering this request.

Sincerely,
Sabiq Sabry

SkywalkerBT <alihamad38@gmail.com> 4 February 2023 at 13:19
To: Mohamed Sabiq Mohamed Sabry <w1921949@my.westminster.ac.uk>

Dear Sabiq,

I agree to allow you to use my WhatsApp chats for the research project you are conducting. I understand that all information will be kept confidential and used solely for the purposes of the research project and that any messages used in the research will be anonymized to protect my privacy.

If you have any further questions or concerns, please let me know.

Best regards,
Hamad

[Quoted text hidden]
[Quoted text hidden]

This message and its attachments are private and confidential. If you have received this message in error, please notify the sender and remove it and its attachments from your system.

The University of Westminster is a charity and a company limited by guarantee. Registration number: 977818 England. Registered Office: 309 Regent Street, London W1B 2UW.

Abdul Khaliq <khaliqsadath7@gmail.com> 4 February 2023 at 13:23
To: Mohamed Sabiq Mohamed Sabry <w1921949@my.westminster.ac.uk>

Dear Sabiq,

I agree to allow you to use my WhatsApp chats for the research project you are conducting. I understand that all information will be kept confidential and used solely for the purposes of the research project and that any messages used in the research will be anonymized to protect my privacy.

If you have any further questions or concerns, please let me know.

https://mail.google.com/mail/u/0/?ik=ad94a57346&view=pt&search=all&permthid=thread-a%3Ar-3890463812792059173&simpl=msg-a%3Ar-3837584432838… 1/3

05/02/2025, 21:25 University of Westminster Mail - Request for Consent to Use WhatsApp Chats for Research Project

Best regards,
Abdul Khaliq

[Quoted text hidden]
[Quoted text hidden]
[Quoted text hidden]

Atheek Nazeer <atheek990@gmail.com> 4 February 2023 at 13:25
To: Mohamed Sabiq Mohamed Sabry <w1921949@my.westminster.ac.uk>

Dear Sabiq,

I agree to allow you to use my WhatsApp chats for the research project you are conducting. I understand that all information will be kept confidential and used solely for the purposes of the research project and that any messages used in the research will be anonymized to protect my privacy.

If you have any further questions or concerns, please let me know.

Best regards,
Atheek.

On Sat, 4 Feb 2023, 18:43 Mohamed Sabiq Mohamed Sabry, <w1921949@my.westminster.ac.uk> wrote:
[Quoted text hidden]
[Quoted text hidden]

Shuaib Uvais <shuaibusman70@gmail.com> 4 February 2023 at 13:26
To: Mohamed Sabiq Mohamed Sabry <w1921949@my.westminster.ac.uk>

Dear Sabiq,

I agree to allow you to use my WhatsApp chats for the research project you are conducting. I understand that all information will be kept confidential and used solely for the purposes of the research project and that any messages used in the research will be anonymized to protect my privacy.

If you have any further questions or concerns, please let me know.

Best regards,
Shuaib

[Quoted text hidden]
[Quoted text hidden]
[Quoted text hidden]

Hapsa Firthous <fhapsa58@gmail.com> 4 February 2023 at 13:29
To: Mohamed Sabiq Mohamed Sabry <w1921949@my.westminster.ac.uk>

Dear Sabiq Sabry,

Thank you for your email.

I, Hapsa Firthous, give you my consent to use my WhatsApp chats as part of your research project.

Thank you,

Hapsa Firthous

[Quoted text hidden]
[Quoted text hidden]
[Quoted text hidden]

Nafeel noordeen <nafeelnoordeen41@gmail.com> 4 February 2023 at 13:51
To: Mohamed Sabiq Mohamed Sabry <w1921949@my.westminster.ac.uk>

https://mail.google.com/mail/u/0/?ik=ad94a57346&view=pt&search=all&permthid=thread-a%3Ar-3890463812792059173&simpl=msg-a%3Ar-3837584432838… 2/3

05/02/2025, 21:25 University of Westminster Mail - Request for Consent to Use WhatsApp Chats for Research Project

Dear Sabiq,

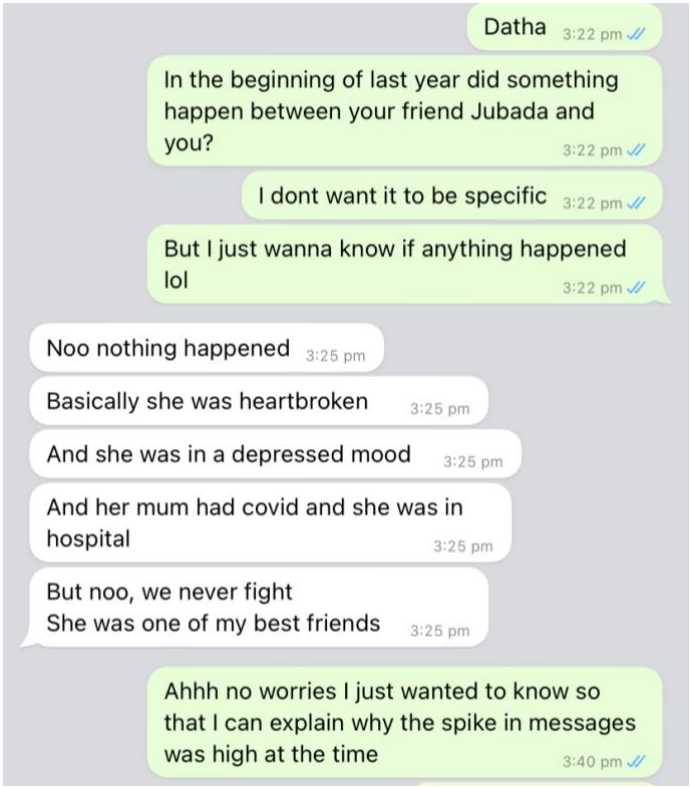
I agree to allow you to use my WhatsApp chats for the research project you are conducting. I understand that all information will be kept confidential and used solely for the purposes of the research project and that any messages used in the research will be anonymized to protect my privacy.

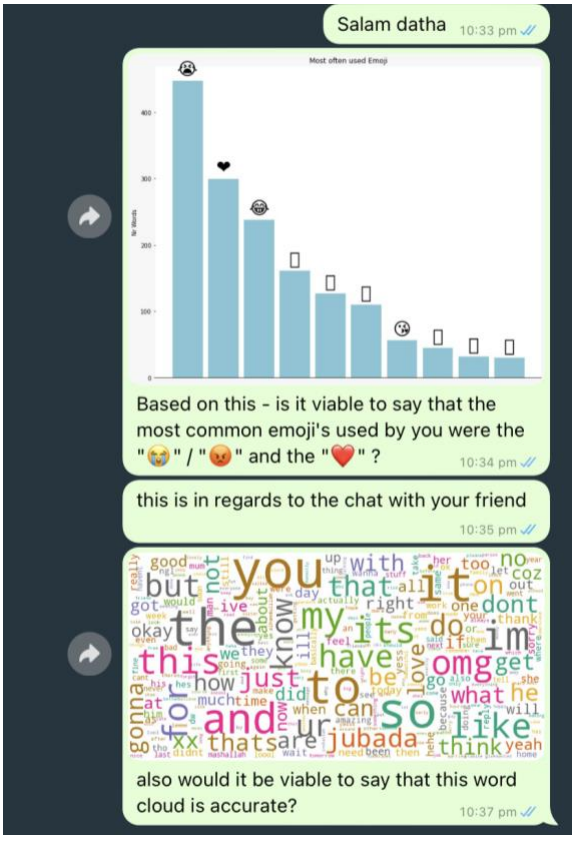
If you have any further questions or concerns, please let me know.

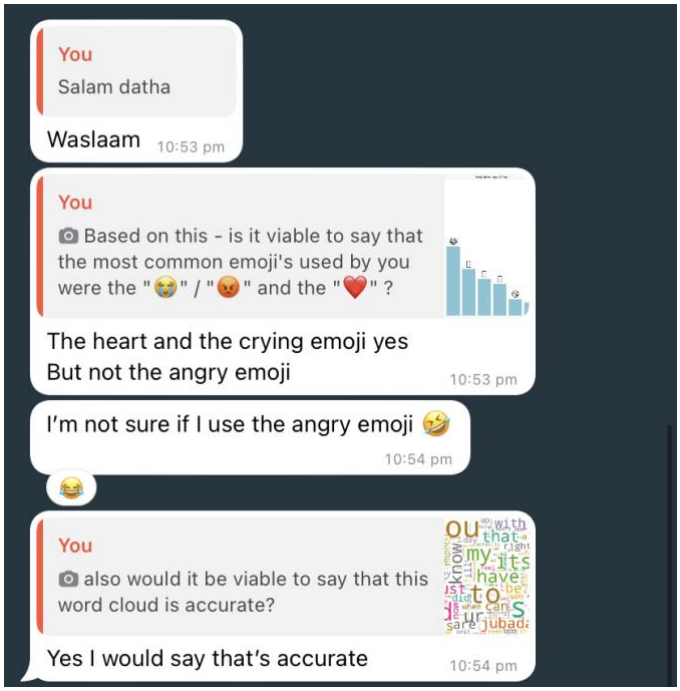
Best regards,
Nafeel Noordeen.

[Quoted text hidden]
[Quoted text hidden]
[Quoted text hidden]

2. Confirmation of Results with User







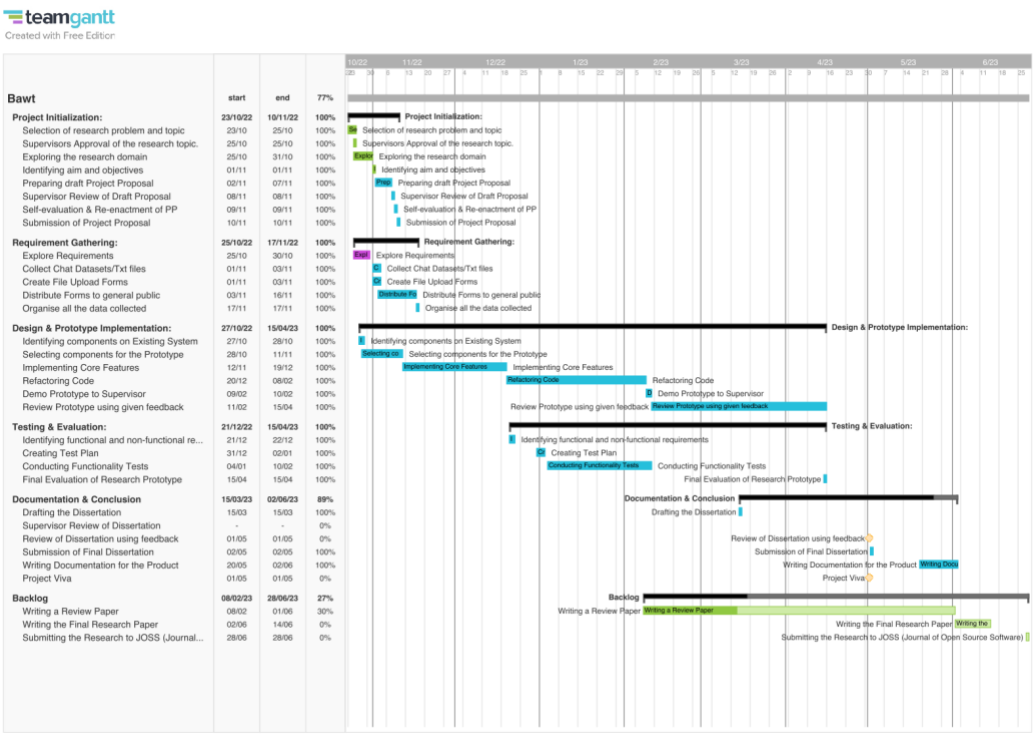
3. Task List

Topics to cover for FYP

Problem: figuring out a person's behavioural pattern using text messages

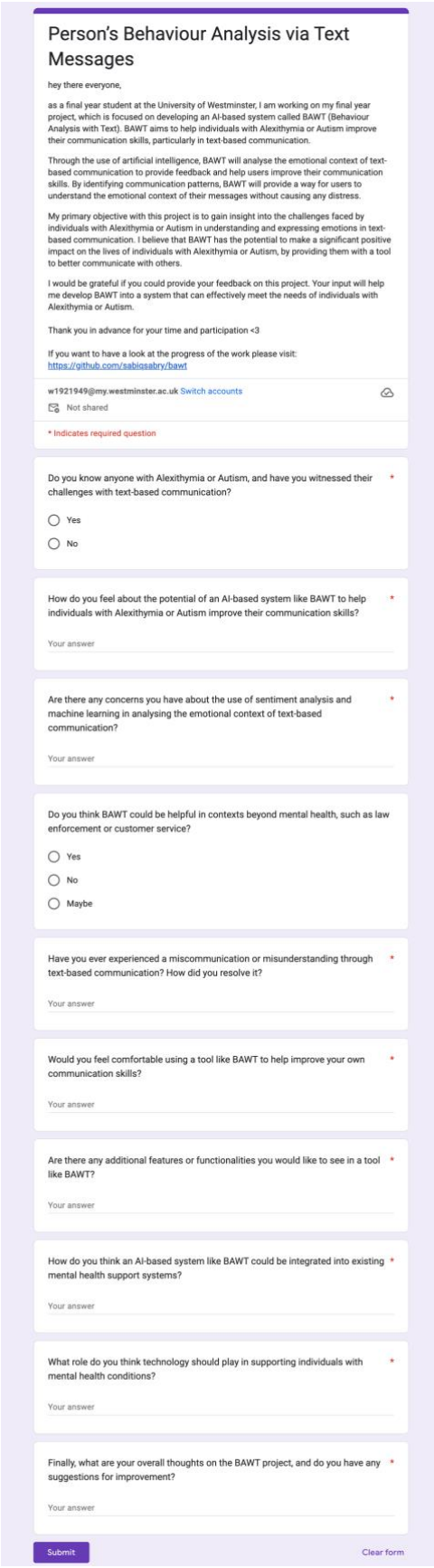
- ✓ Polarity using Sentiment Analysis - to find out the tone of the messages
- ✓ Fix the import txt file format no matter which OS.
- ✓ Sentiment Analysis Based Cloud Words - to demonstrate the which words appear more frequently.
- ✓ Topic Modelling - to figure out what topics most interest the user so that the end individual can do something about - for eg: if the person likes the topic - dog - then the end individual can do something about it by getting the user a dog.
- ✓ Emogi Analysis using TF-IDF
- ✓ Unique Words
  - ✓ using TF-IDF (term frequency-inverse document frequency) to figure out the most used used by the individual
- ✓ Mood Identifier - whether the person is feeling optimistic or pessimistic or normal
- ✓ Exploratory Data Analysis (EDA) -
  - ✓ weekly count of messages over time
  - ✓ frequency of messages

4. Gant Chart:



See full quality on: https://drive.google.com/file/d/1CBMWNdm6XFL9JJ-LSnCOIbKo2hRIX217/view?usp=share_link

5. Bawt Survey



Person's Behaviour Analysis via Text Messages

hey there everyone,

as a final year student at the University of Westminster, I am working on my final year project, which is focused on developing an AI-based system called BAWT (Behaviour Analysis with Text). BAWT aims to help individuals with Alexithymia or Autism improve their communication skills, particularly in text-based communication.

Through the use of artificial intelligence, BAWT will analyse the emotional context of text-based communication to provide feedback and help users improve their communication skills. By identifying communication patterns, BAWT will provide a way for users to understand the emotional context of their messages without causing any distress.

My primary objective with this project is to gain insight into the challenges faced by individuals with Alexithymia or Autism in understanding and expressing emotions in text-based communication. I believe that BAWT has the potential to make a significant positive impact on the lives of individuals with Alexithymia or Autism, by providing them with a tool to better communicate with others.

I would be grateful if you could provide your feedback on this project. Your input will help me develop BAWT into a system that can effectively meet the needs of individuals with Alexithymia or Autism.

Thank you in advance for your time and participation <3

If you want to have a look at the progress of the work please visit:
https://github.com/sabiqsabry/bawt

w1921949@my.westminster.ac.uk Switch accounts

Not shared

* Indicates required question

Do you know anyone with Alexithymia or Autism, and have you witnessed their challenges with text-based communication? *

- Yes
- No

How do you feel about the potential of an AI-based system like BAWT to help individuals with Alexithymia or Autism improve their communication skills? *

Your answer

Are there any concerns you have about the use of sentiment analysis and machine learning in analysing the emotional context of text-based communication? *

Your answer

Do you think BAWT could be helpful in contexts beyond mental health, such as law enforcement or customer service?

- Yes
- No
- Maybe

Have you ever experienced a miscommunication or misunderstanding through text-based communication? How did you resolve it? *

Your answer

Would you feel comfortable using a tool like BAWT to help improve your own communication skills? *

Your answer

Are there any additional features or functionalities you would like to see in a tool like BAWT? *

Your answer

How do you think an AI-based system like BAWT could be integrated into existing mental health support systems? *

Your answer

What role do you think technology should play in supporting individuals with mental health conditions? *

Your answer

Finally, what are your overall thoughts on the BAWT project, and do you have any suggestions for improvement? *

Your answer

Submit Clear form

Check Survey Responses here:
https://drive.google.com/file/d/1iuyyiV6HNhDkcO7ManlbtcaUKtIWVRe6/view?usp=share_link

6. Gathering Requirements

In this case, a library called SoAn was used to gather requirements (Social Analysis). Data from WhatsApp messages was extracted using SoAn, which included word frequency, word clouds, TF-IDF, and Sentiment Analysis.

SoAn was used to collect data from the researcher's and a few colleagues' WhatsApp messages who volunteered to share their chats with specific individuals. The messages were then analysed and aggregated to create a thorough examination of the chats.

The data analysis results were summarised using various word frequency and word cloud visualisations, as well as sentiment analysis results. The findings revealed insights into the types of words and emotions commonly expressed in WhatsApp chats, which can be useful in understanding individuals' communication patterns and preferences.

Because the data collected and analysed was limited to the researcher's own chats and a few voluntarily shared chats, the results may not be representative of all WhatsApp users or conversations. To confirm the findings and ensure generalizability, more research with a larger and more diverse sample size is required of which will be collected in due time to the final representation of the project.

	Functional Requirements
Essential	● Compatibility with OS ● Python and Libraries like Tensorflow, Pandas, Numpy, etc. are supported

	• VSCode • Google Colab/Jupyter • Ability to create figures, architectures, documentations - Google Docs/Canva/Draw.io • Data Backup - Google Drive/Github • Ability to develop ML/DL/NLP models • Research Writing Skills
Desirable	• Support for Linux and Windows • High Performance Hardware - intel i7/M1 or above • Massive RAM and Disk Space to manage datasets and development environments
Luxury	• Advanced Data Visualization tools • UI friendly interface

Table 9: Functional Requirements

	Non-Functional Requirements
Essential	• Performance: the application should be able to efficiently process large datasets and perform intensive resource tasks. • Security: the research project should safeguard user privacy and confidential data. • Reliability: the research project should be stable and should not crash when used normally.

	• Scalability refers to the research project's ability to handle increasing amounts of data and users.
Desirable	• User Experience: the research project should be simple and easy to use. • Flexibility: the research project should be adaptable and integrate with other tools. • Maintenance: the research project should be simple to update and maintain.
Luxury	• Personalization: based on the user's preferences, the research project should make tailored recommendations.

Table 10: Non-Functional Requirements

7. External Links

Here is the video: https://youtu.be/AbUBaxMdhtc

Here is my GitHub repo: https://github.com/sabiqsabry/bawt

Here is the Google Drive Link for my work: https://drive.google.com/drive/folders/1aFyn2tiZqEg2P_vIKWao8Pl-s0RggvWP?usp=share_link