Intention Mining from Social Network Data: A Fuzzy Logic Model Based on the Theory of Planned Behaviour

Qadri Mishael

Supervisor: Prof. Aladdin Ayesh
Dr. Iryna Yevseyeva

Cyber Technology Institute
De Montfort University

This dissertation is submitted for the degree of
Doctor of Philosophy

Faculty of Computing,
Engineering and Media

September 2023

To my loving parents, Khawlah and Izzat , whose unwavering love, support, and sacrifice have been the foundation upon which I have built my dreams. Thank you for instilling in me the value of hard work, perseverance, and the pursuit of knowledge. This thesis is a testament to your enduring love and guidance.

To my wonderful wife, Rana, who has been my rock and source of inspiration throughout this journey. Your patience, understanding, and unconditional love have been a constant reminder of the beauty and strength that lies within our partnership. This accomplishment is as much yours as it is mine.

To my precious daughter, Qamar, who has brought immeasurable joy and light into our lives. May this thesis serve as a reminder that you can achieve anything you set your heart and mind to and that your parents will always be by your side, cheering you on every step of the way.

This thesis is dedicated to each one of you with all my love and gratitude.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Qadri Mishael
September 2023

Acknowledgements

Acknowledgements:

First and foremost, I would like to express my deepest gratitude to my supervisory team, Prof. Aladdin Ayesh and Dr Iryna Yevseyeva, for their unwavering guidance, support, and encouragement throughout my research. Their expertise and profound understanding of the subject matter have been invaluable in helping me to navigate the complexities of this thesis. I am genuinely grateful for their patience, constructive feedback, and the countless hours they have dedicated to helping me shape my ideas and sharpen my analytical skills.

I also sincerely thank Dr Lipika Deka and Dr Mamadou Bamba for their support and assistance during my research journey. Their insightful suggestions, willingness to share their knowledge, and camaraderie have contributed immensely to the success of this thesis. I am fortunate to have had the opportunity to learn from and collaborate with such esteemed colleagues.

Furthermore, I thank my friends and family, including my loving parents Izzat and Khawlah, my devoted wife Rana, my precious daughter Qamar, my supportive brothers Majdi and Omar, and my caring sisters Manal and Duja, for their constant support, understanding, and encouragement throughout this challenging endeavour. Their belief in my abilities and unwavering love and care have given me the strength and motivation to persevere during the most trying times.

Lastly, I am grateful to all those who have contributed to my academic growth and personal development in various capacities. Through the collective wisdom, kindness, and support of these individuals, I have reached this milestone in my academic journey.

Abstract

The pervasive growth of social networks has led to an unprecedented wealth of information generated by users, providing vast opportunities for understanding and predicting user intentions. This research aims to investigate the impact of sentiment analysis on modelling social network users' intentions, focusing on the Theory of Planned Behaviour (TPB) and Fuzzy Logic (FL).

Drawing upon publicly available online datasets and data collected from social media platforms such as Twitter, this thesis employs advanced sentiment analysis and fuzzy logic techniques to create a robust model for estimating user intentions. By examining users' attitudes, subjective norms, and perceived behavioural control, valuable insights into the factors influencing their social network behaviour are uncovered. The study employs advanced classifiers such as the Decision Tree, Naive Bayes, and Support Vector Machine to validate the efficacy of features in predicting user intentions.

Furthermore, the thesis underscores the significance of sentiment analysis, utilising the NRC Valence, Arousal, and Dominance (VAD) Lexicon to evaluate users' opinions and emotions towards various topics, products, and services. The seamless integration of sentiment analysis and fuzzy logic into the methodology for modelling the TPB ensures accurate and reliable results.

The contribution of this research lies in its ability to provide a more comprehensive understanding of the impact of sentiment analysis on social network users' intentions and to demonstrate the feasibility of using fuzzy logic in different applications, such as recommendation systems. The results of this research will interest a wide range of stakeholders, including researchers, businesses, and policymakers. In conclusion, this thesis presents a novel and comprehensive methodology for understanding and predicting user intentions within social networks, ultimately facilitating the development of tailored and effective marketing and advertising strategies.

Publications

- Mishael, Q., Ayesh, A., & Yevseyeva, I. (2019). Users intention based on Twitter features using text analytics. In Intelligent Data Engineering and Automated Learning–IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part I 20 (pp. 121-128). Springer International Publishing.
- Mishael, Q., & Ayesh, A. (2020). Investigating classification techniques with feature selection for intention mining from Twitter feed. arXiv preprint arXiv:2001.10380.

Table of contents

List of figures

List of tables

List of Algorithms

List of Abbreviations

The next list describes several abbreviations that will be later used within the body of the thesis

A_n Social Network attributes

ANEW Affective Norms of English Words

APIs Application Provider Interfaces

At Tweet attributes

ATB Attiude Toward Behaviour

Au User attributes

BOW Bag-of-Words

C_{Deg} Degree of Centrality

CB Betweenness Centrality

CCo Degree of Clustering Coefficient

CIMM Consumption Intention Mining Model

CNN Convolutional Neural Network

CRF Conditional Random Fields

DT Decision Trees

FFS Forward-Feature Selection Algorithm

FL Fuzzy Logic

Fn Network Features

FNN Feedforward Neural Network

FRBS Fuzzy-Rule-Based System

Ft Tweets Textual Features

Fu User Features

HMI Human-Machine Interaction

IG Information Gain

KNIME Konstanz Information Miner

LOOCV Leave-One-Out Cross Validation

N_{fee} Network for Followees

N_{fol} Network for Followers

N_{hash} Network for Hashtags

N_{men} Networks Mentions

N_{ret} Network of retweet

N_t Network of Tweet

NB Naive Bayes

NLP Natural Language Processing

NN Neural Networks

NRC National Research Council Canada

NRC-VAD NRC Valence, Arousal, and Dominance Lexicon

PBC Precieved Behaviour Control

POS Part of Speech

PR Page Rank

RBF Radial Basis function

RF Random Forest

RFE Recursive Feature Elimination

SGD Stochastic Gradient Descent

SMO Sequential Minimal Optimization

SMO The Sequential Minimal Optimization

SN Subjective Norm

SVM Support Vector Machines

TF Term Frequency

TPB Theory of Planned Behaviour

TSK Takagi-Sugeno-Kang

VADER Valence Aware Dictionary and sEntiment Reasoner

VT Tweet Vector

Chapter 1

Introduction

Nowadays, online social networks have become an integral aspect of people's lives. Users of social media want to exchange information by posting about their everyday activities, feelings, views, hobbies, or ambitions. The postings' content types range from text to photos, video clips, and even URLs.

Researchers made an effort to extract human intention despite the fact that doing so depends on the reader's comprehension of social network posts. The researchers recognized intention detection in social networks as a valuable source of information for understanding customers' needs for online businesses[116, 79]. Moreover, it would support building social network platforms with enhanced user services.

The present thesis aims to address the gap in the current literature related to understanding social network users' intentions based on the theory of planned behaviour (TPB) model and how it can be modelled through the application of fuzzy logic (FL). The growing significance of social network platforms in our daily lives has led to a need for effective methods to understand and predict users' behaviour on these platforms. The main focus of this research is to develop a model that can effectively predict the users' intention to engage with social media platforms, such as Facebook and Twitter, through sentiment analysis.

This thesis first presents a comprehensive review of the existing literature related to the topic of social network user behaviour, including the TPB and FL. The research then moves on to the methodology section, which outlines the data collection process and the use of publicly available online datasets in addition to data collected from social media platforms such as Twitter.

The findings of this research will contribute significantly to the existing literature on social network user behaviour. Through the application of FL, this research will demonstrate how social network users' intentions can be effectively modelled and predicted. Moreover, this research will also provide recommendations for future studies on the topic, which will pave the way for the further advancement of this field.

In conclusion, this PhD thesis provides a detailed examination of the topic of social network users' intentions and offers a novel approach to modelling and predicting their behaviour through the application of fuzzy logic and sentiment analysis. The research aims to contribute to the advancement of our understanding of social network user behaviour and to provide practical implications for the design and development of future social media platforms.

1.1 Motivation

The rapid growth of online social networks has opened up a vast amount of opportunities for individuals and organizations to connect and share information. However, with the ever-increasing amount of information available, it has become increasingly challenging for users to navigate and make sense of the information. The vast amount of information can lead to information overload, making it difficult for users to make informed decisions. In this context, understanding the intentions of online social network users has become a critical issue, particularly in the context of sentiment analysis.

Intention is a critical construct in understanding human behaviour, and predicting intentions are of significant importance in social science research. The TPB is one of the most widely-used frameworks for predicting intentions, and it has been applied successfully in many fields, including health, psychology, and marketing. However, the use of TPB in predicting social media intentions is still limited, and there is a need for a more accurate and effective approach to predicting social media users' intentions.

One approach to predicting intentions on social media is through the use of sentiment analysis. Sentiment analysis is a powerful tool that can be used to extract and analyse sentiment from social media data, such as tweets, to gain insights into user opinions and attitudes. However, traditional sentiment analysis methods are limited in their ability to capture the nuances of sentiment expressed in informal language and may not accurately capture the affective dimensions associated with different intentions.

To address this gap in the literature, this study proposes a fuzzy model of intention based on the TPB, which combines the NRC Valence, Arousal, and Dominance (VAD) Lexicon (NRC-VAD)-based sentiment analysis approach with FL. The NRC-VAD provides a standardized measure of the affective norms associated with English words, allowing for the identification and extraction of emotions, opinions, and attitudes expressed in text data. This approach enables the sentiment analysis model to accurately capture the nuances of sentiment expressed in informal languages, such as tweets related to various social media behaviours and provides valuable insights into the attitudes and opinions of users towards different intentions.

Moreover, the proposed approach can be applied to different types of tweet attribute sets (textual, user, and network) to develop a fuzzy model of intention that incorporates the different factors of the TPB. This allows for a comprehensive analysis of user sentiment and intention on social networks, which can inform the design of more effective and efficient recommendation systems and other technologies.

1.2 Aims and Objectives

The aim of this research is to investigate the modelling of social network users' intentions based on TPB using FL in the context of sentiment analysis. The study of user intentions has been a key topic in the field of psychology and sociology for many years, with a significant body of research focused on understanding how individuals make decisions and what influences their behaviour. In recent years, there has been growing interest in understanding user intentions in the context of online social networks. This is particularly relevant in the context of sentiment analysis, as understanding user intentions can provide valuable insights into how users perceive and evaluate information.

Traditional methods for sentiment analysis have been shown to be limited in their ability to capture the complex and nuanced nature of human emotions and opinions, particularly in the context of online social networks, where users express their opinions and emotions through short and informal messages. This highlights the need for a more flexible and interpretable approach to sentiment analysis. FL has emerged as a promising approach in this regard, providing a framework for modelling complex and uncertain data.

Furthermore, this research aims to contribute to the academic literature by expanding the understanding of the TPB and its application in the field of sentiment analysis. The study will be the first to apply FL in the context of sentiment analysis for predicting social network users' intentions based on the TPB.

1.3 Research Questions

To achieve the aims and objectives outlined in this chapter, the following research questions will be addressed in the course of this research:

1. How can the theory of planned behaviour be used to model social network users' intention to perform specific actions?

2. How can fuzzy logic be used to handle the uncertainty and vagueness in users' decision-making processes in the context of modelling social network users' intentions?

3. How can sentiment analysis be used to model users' intentions on social networks?

4. How effective is the proposed approach in modelling social network users' intentions compared to existing methods?

The answers to these research questions will provide a comprehensive understanding of the users' sentiments and intentions on online social networks and will be crucial in developing more effective and efficient recommendation systems.

1.4 Research Contribution

The proposed research aims to make a significant contribution to the field of sentiment analysis and social network user intention prediction by introducing a novel approach that combines the TPB with FL. This study is particularly relevant in the context of social networks, where the analysis of user sentiment and intention can provide insights into their behaviour and preferences.

The use of FL in this context is a novel approach that has not been widely explored in previous research, making this study the first of its kind. This approach is particularly relevant to the research questions, as it allows for the incorporation of uncertain and ambiguous data that is typical of social media content.

In addition to the novel approach proposed in this study, the research will also contribute to the development of sentiment analysis and social network user intention prediction by expanding the understanding of the TPB and its application in these areas. By exploring the relationship between social network users' attitudes, subjective norms, perceived behavioural control and their intention to engage in specific behaviours, valuable insights into the factors that influence user behaviour on social networks could be gained.

Moreover, the proposed approach can be applied to different types of tweet attribute sets (textual, user, and network) to develop a fuzzy model of intention that incorporates the different factors of the TPB. This allows for a comprehensive analysis of user sentiment and intention on social networks, which can inform the design of more effective and efficient recommendation systems and other technologies.

In conclusion, this research aims to make a valuable contribution to the fields of sentiment analysis and social network user intention prediction by introducing a novel approach that combines the TPB with FL. The results of this research will be of interest to researchers, practitioners, and businesses alike, as it has the potential to inform the development of new technologies and services that support the effective use of social networks.

1.5 Chapters Outline

The rest of this report is organised as follows:

- **Chapter 2**: This chapter provides an overview of the relevant concepts and technologies used in sentiment analysis, intention mining, data mining, and FL in order to inform the research design and methodology for this thesis.

- **Chapter 3**: This chapter provides an overview of the research methodology used, which is a combination of both qualitative and quantitative approaches, and includes data collection, feature selection and supervised learning, sentiment analysis, FL modelling, and model validation and evaluation.

- **Chapter 4**: This chapter details the methodology used for collecting and analysing Twitter data to estimate user intentions, including data collection, database structure, data cleaning and pre-processing, data quality assessment, and NLP techniques used to extract and analyse features.
- **Chapter 5**: This chapter explains the use of feature selection, supervised learning and sentiment analysis approaches for intention mining.
- **Chapter 6**: The chapter explains the proposed FL model for intention estimation based on TPB.
- **Chapter 7**: This chapter presents the intention mining experiments that have been done, such as collecting social data, building a social corpus, the classification attempts for the users' intention, and the FL model for the intention based on TPB.
- **Chapter 8**: The final chapter to conclude the thesis and directions to future works.

Chapter 2

Literature Review

The use of social networks has become ubiquitous in today's society, providing a rich source of information on individuals' attitudes, opinions, and behaviours. Understanding the factors that influence behaviour is crucial in a wide range of fields, including health care, consumer research, and marketing. The theory of planned behaviour (TPB) is a widely used model that aims to predict and explain human behaviour by considering the individual's intention to perform the behaviour, which is, in turn, influenced by attitudes towards the behaviour, subjective norms, and perceived behavioural control.

Recent research has shown that sentiment analysis and FL can be used in combination with the TPB to gain a more detailed and nuanced understanding of the factors that influence behaviour. Sentiment analysis is a technique used to automatically identify and extract subjective information from text, such as opinions, evaluations, appraisals, and emotions. FL, on the other hand, is a mathematical method that can be used to model the vagueness and uncertainty of natural languages.

The aim of this literature review chapter is to provide an overview of the current state of research on the use of sentiment analysis and FL to extract attitudes, subjective norms, and perceived behavioural control from social media posts and to use these variables to predict and explain behaviour using the TPB model. The chapter will review the relevant literature on this topic, discuss the potential benefits and limitations of this approach, and identify areas for future research.

2.1 Overview of Intention Modelling Theories

The researchers presented different definitions of user intention, and various models were provided to present intention based on the data source and the target system [28, 39, 107, 121, 142]. This research adapts Ajzen et al. [6] model of human behaviour named "Theory of Planned Behaviour", which is presented in Psychology. This theory will be used to present the user intention in the social network.

Intention modelling is a crucial aspect of understanding and predicting human behaviour in various research fields such as social networks, healthcare systems, marketing, and consumer research [9, 31, 36, 107, 95, 133, 146]. The TPB model is one of the most widely used models for understanding and predicting behaviour by considering the individual's intention to perform the behaviour as the key factor. Intention, in this context, refers to an individual's motivation and willingness to engage in a specific behaviour. Research in social networks, for example, has focused on understanding the intention behind individuals' online behaviours, such as their likelihood to engage in online activities such as purchasing, sharing or commenting on a specific product. In the field of healthcare, researchers have used intention modelling to predict and understand individuals' health behaviours, such as their likelihood to engage in regular physical activity or to comply with a treatment regimen. In marketing, understanding consumers' buying intentions is crucial to effective product development and promotion. Researchers have used intention modelling to understand the factors that influence consumers' buying decisions, such as their attitudes, subjective norms, and perceived behavioural control. In consumer research, researchers have used intention modelling to understand the factors that influence consumers' purchase decisions, such as their attitudes, subjective norms, and perceived behavioural control.

2.1.1 Theory of Reasoned Actions

The Theory of Reasoned Action (TRA) is a psychological model that was first developed in the 1970s by Martin Fishbein and Icek Ajzen. The model aims to explain and predict human behaviour by considering the individual's attitude towards the behaviour and the perceived social pressure to engage in it. It postulates that an individual's behaviour is determined by their intention to perform that behaviour.

Later on, Ajzen expanded the model and modified it to become the Theory of Planned Behaviour in the 1980s. The main distinction between the two models is that the TPB added a third variable, perceived behavioural control, which represents an individual's belief in their capability to carry out the behaviour. This variable was included to take into account situations where an individual may have a positive attitude and social pressure to engage in behaviour but still may not engage in that behaviour because they do not feel capable of doing so.

In summary, The Theory of Reasoned Action (TRA) is a psychological model that was first introduced in the 1970s; that aims to explain and predict human behaviour by considering the individual's attitude towards the behaviour and the perceived social pressure to engage in it. Later, Ajzen expanded the model and modified it to become the Theory of Planned Behaviour (TPB) in the 1980s. This expansion added a third variable, perceived behavioural control, which represents an individual's belief in their capability to carry out the behaviour. This variable was included to take into account situations where an individual may have a positive attitude and social pressure to engage in behaviour but still may not engage in that behaviour because they do not feel capable of doing so.

2.1.2 Theory of Planned Behaviour

The theory of planned behaviour (TPB) is a widely accepted psychological model that was developed to predict and explain human behaviour across a range of contexts. Icek Ajzen first proposed the model in the 1980s [6].

The TPB model, presented in Figure 2.1, posits that an individual's behaviour is determined by their intention to perform that behaviour and that this intention is influenced by three key factors: Attitudes Towards the Behaviour (ATB), subjective norms (SN), and perceived behavioural control (PBC).

ATB are an individual's evaluation of the behaviour in question. This includes their beliefs about the consequences of the behaviour, such as whether it is good or bad, desirable or undesirable, and whether it will lead to positive or negative outcomes. Research has shown that ATB is a strong predictor of behaviour, as individuals are more likely to engage in behaviours that they believe are good or desirable [6, 21, 50, 55, 130].

SNs are an individual's perceptions of the social pressure to engage in the behaviour. This includes the opinions and expectations of important people in their life, such as family, friends, and colleagues. Research has shown that SN is a strong predictor of behaviour, as individuals are more likely to engage in behaviours that they perceive as socially acceptable or desirable [6, 55, 65, 107].

PBC is an individual's beliefs about their ability to perform the behaviour. This includes the presence of resources and perceived barriers to the behaviour, such as time, money, and skill. Research has shown that PBC is a strong predictor of behaviour, as individuals are more likely to engage in behaviours that they believe they are able to perform [6, 37, 55, 96].

Overall, the three factors of the TPB model (ATB, SN, and PBC) are considered key predictors of behaviour and have been found to have a strong explanatory power. Studies have found that the TPB is a robust and reliable model for predicting behaviour, with strong explanatory power [21, 50, 65, 107, 130, 133]. The combination of these factors can provide valuable insights into the factors that influence behaviour and can be used to understand and predict human behaviour in a wide range of fields.

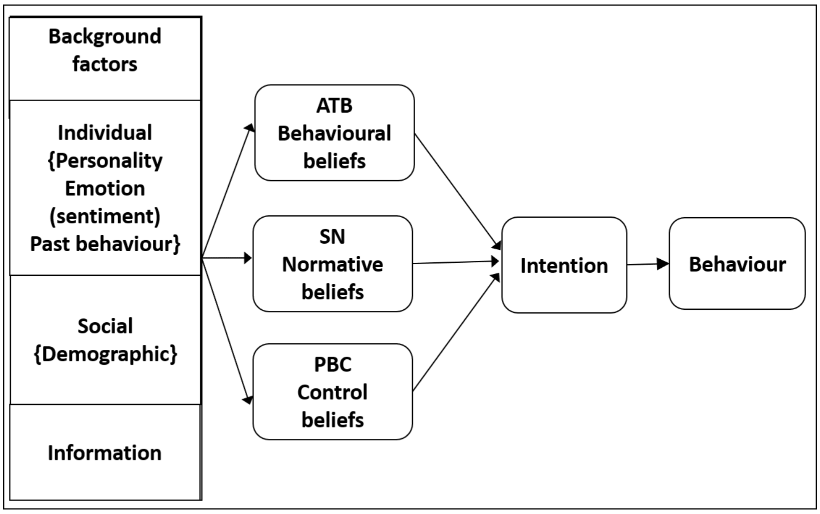


Fig. 2.1 Theory of Planned Behaviour Model [55].

In summary, this section of the literature review explored the two main models used for modelling intention. The next sections cover the discussion on the use of data mining and machine learning techniques, sentiment analysis techniques and FL to extract attitudes, subjective norms, and perceived behavioural control from social media posts and other text sources. Examples of these models will be discussed, and their potential benefits and limitations will be analysed. Additionally, this section will also examine the ways in which intention modelling has been applied in different research fields, such as social networks, healthcare systems, marketing, and consumer research. It will provide an in-depth understanding of the current state of the field and identify areas for future research.

Examples of intention modelling in these fields include:

- In social network research, studies have used intention modelling to predict individuals' likelihood to engage in online behaviours, such as sharing or commenting on a specific product [28, 121, 130, 133].
- In the healthcare field, researchers have used intention modelling to predict and understand individuals' health behaviours, such as their likelihood to engage in regular physical activity or to comply with a treatment regimen [31, 36, 37, 96].
- In marketing, researchers have used intention modelling to understand consumers' buying intentions and the factors that influence their buying decisions [9, 39, 65, 95].
- In consumer research, studies have used intention modelling to understand the factors that influence consumers' purchase decisions [21, 39, 50, 107, 146].

Overall, this section of the literature review will provide a comprehensive overview of the current state of intention modelling research and its applications in various fields. It will also highlight the potential benefits and limitations of using sentiment analysis and FL in intention modelling and will identify areas for future research.

2.2 Overview of Intention Mining in Research

This section of the literature review will provide an overview of the current state of research on intention mining, focusing on the literature on mining intention in human-machine interaction, mining intention for online web queries, and user intention mining in social networks. It will review the various models and approaches used in these studies and will discuss the potential benefits and limitations of using these techniques in understanding and predicting human behaviour.

A number of papers [28, 29, 42, 45, 110, 133, 138], focused on discovering user intention from social networks and Web pages. Users use social media to express their desires, wishes, likes and dislikes on social networks. However, data on social networks may not be presented in an appropriate format for mining, such as language informality in social posts, misspellings, emot-icons, hashtags, and multiple data formats such as images, audio and videos.

In general, mining user intention studies have been carried out with different aims. In Dai et al.[39], the aim of the study is mining commercial consumption intention. However, Change et al.[28] study mining user intention to identify adopting services by social network users. Nevertheless, Das et al.[42] study mining the intention aimed at information diffusion in social networks. At the same time, Oh et al. [110] study intention mining on developing travel destination recommendations. The intention mining problem is considered a classification problem according to Chen et al.[35], and Zhang et al[148]. Hidden Markov Model(HMM) algorithm used by different researchers to build intention models, for instance, Khodabandelou et al. [80] built Map Mind model based on HMM to find user's intention from computer logs, assuming the problem as a dynamic problem to find the hidden goals of the users. The studies of mining users' intentions on social networks have been carried out with different aims. For example, Dai et al.[39], and Ding et al.[45] worked on mining commercial consumption intention, and Chang et al[28] tried to extract users' intention to adopt social network services, or Oh et al[110] research on intention develop travel destination recommendations to the social network users. The following are some of the studies that targeted intention mining and its applications in computer science.

2.2.1 Mining Intention for Human-Machine Interaction

Exploring the intention of users goes back to the Lumiere project[73]. The research aim was to develop a software system assistance for Microsoft Office 97. The user actions and queries were monitored to capture and model the user's goals i.e intentions. Bayesian user model networks were used to create models; they applied the time pause as a factor in their research. The mentioned work focused on explicit intent extraction through the listen to the users' queries.

Lin et.al.[89] also proposed using Bayesian Belief Networks to learn the user intention model. The model was built by learning user preferences in emails to provide multimedia files as attachments. They considered the linguistic features to capture the user's intention, divided into words, syntax, patterns and synonyms. Also, they used media agents to learn the preferences and record the user's actions for a long period of time. Their work suffers from the long computation time due to the complexity of applying the Bayesian Belief Networks since it increases the records count in the dataset.

Chen et al. [35] worked on the user interaction with the web browsing environment to model users' intention from surfing the web through log files. They used a modified Naive Bayes classifier algorithm to model user-intended actions. They aimed to predict future actions based on extracting features from the user's typed sentences and viewed content using the Information Gain algorithm. They compared their work to [89] and argued that their algorithm is less complex but low in accuracy.

2.2.2 Mining Intention for Online Web Queries

Online users' web queries were studied to extract users' intentions by analysing these queries. The research presented by Dai et al.[39] focused on user commercial intent mining. They used data that can be found in search queries and URL web page contents. They divided the intention into commercial and non-commercial intentions. The researchers applied a supervised learning algorithm to learn the best combination query results that related to the commercial intention and then used Support Vector Machines (SVM). They worked on discovering the user's intentions after sequencing sessions. Antwarg et al.[8] focused on user attributes to model user intention. They wanted to predict the goals that the user wanted to achieve. Their model used both the user actions and the user attributes. The model was represented by HMM tree leaf and combined with Conditional Random Fields (CRF) and applying pruning to avoid overfitting. They argued that the use of HMM is due to the ability of the model to predict action by observing the set of stochastic processes that produce it. While using, CRF has a better effect on accuracy due to the ability to add overlap features. Their work considered that users have fixed set attributes while in their model could not cope with changing the attributes with different goals.

Another research focused on users' intention on web queries by Budalakoti et al. [22]. Their goal was to model user intention from queries on the web, especially Yahoo! Answers. They classify the queries as if it is set to seek expert aid or for social interaction through rating suggestion for the answers. They considered the objectivity of each user by adopting the Random Surfer model and Page-rank algorithm.

Vineet et al. [137] used Quora and Yahoo! Answers to mine user purchase intention and predict what users want to buy. They aimed to find the linguistic features along with statistical features of purchase intention expressions to classify data over 'bag-of-words' to extract features at two levels of text granularity as word and phrase-based features and grammatical dependency-based features depending on WordNet. Zhang et al. [148] worked on capturing user intention from online medical queries. They proposed a neural network model that utilizes feature-level correlations and models semantic transitions in text queries. The intention application was only in the medical field.

2.2.3 User Intention Mining in Social Networks

Chen et al. [33] worked on identifying user intents that can be found in posts on forums. They built their Co-Class transfer learning classifier algorithm based on the Information Gain method given by [144] for feature selection and Expectation Maximization algorithm.

Ding et al.[45], a Consumption Intention Mining Model (CIMM) has been proposed. The researchers focused on mining implicit intent from Twitter. They used Convolution Neural Network (CNN) to identify the consumption. They formulate consumption intention mining as a binary classification problem by deciding if the sentence contains user consumption intention or not. They compared their work with two other methods, 'Bag-of-words' combined with Support Vector Machine SVM and word embedding combined SVM.

Another approach, Wang et al.[138], proposed a graph-based semi-supervised learning technique for mining user intent by classifying the tweets into six categories. They argued that using their intent keyword would lead to better results than using a 'Bag of Words' since tweets usually are noisy and contain both intent and non-intent-related words, but this would limit the variety of the user intent that would be extracted.

Park et al.[112] worked on recommending mobile applications by predicting user intent from microblogging posts. They proposed parallel corpora for text in microblogging to translate implicit intent text into explicit intent text. Kim et al. [81] worked on classifying travellers' reviews to categorise their travelling intents, used a set of eight intents words in the travelling domain, and created an intent corpus.

2.3 Overview of Sentiment analysis

Sentiment analysis is a method of analysing and extracting subjective information from text-based data, also known as opinion mining [16, 30, 90]. It is used to determine the emotional tone and the overall attitude of a piece of text, such as a social media post [16].

2.3.1 Techniques for Sentiment Analysis

Sentiment analysis can be performed using various techniques, including lexicon-based methods, machine learning, and hybrid techniques [16, 24, 97].

Lexicon Based Techniques

One of the most commonly used techniques for sentiment analysis is sentiment lexicons [97, 120]. Sentiment lexicons provide a list of words and their associated sentiment scores, which can be used to extract sentiment from text. The sentiment score is usually a binary value (positive or negative) or a numerical score that ranges from -1 to 1. There are three main techniques for creating a lexicon:

1. Manual: This involves manually assigning sentiment scores to words. This is the most accurate method, but it is also the most time-consuming [16, 141].

2. Dictionary-based: This involves using a dictionary to find synonyms and antonyms of words. The sentiment scores of the synonyms and antonyms are then used to infer the sentiment score of the original word. This approach is faster and easier than the manual approach, but it is less accurate, as the

dictionary may not contain all of the words that are used in the text being analysed [16, 97].

3. Corpus-based: This involves using a corpus of text to find words that co-occur frequently. The sentiment scores of the words that co-occur frequently are then used to infer the sentiment score of the original word. This approach is the most accurate of the three, but it is also the most computationally expensive [16, 97].

The main advantage of lexicon-based sentiment analysis is that it is relatively simple to implement and can be used to analyse text in any language, as long as a lexicon for that language is available. Additionally, lexicon-based sentiment analysis can be used to analyse text quickly.

However, lexicon-based sentiment analysis also has some disadvantages. One disadvantage is that it can be inaccurate, as it may not be able to capture the nuances of sentiment expressed in the text. Additionally, lexicon-based sentiment analysis can be domain-specific, meaning that it may not be accurate for text from a different domain.

Section 2.3.2 provides a review and an analyse some of the commonly used sentiment lexicons dictionary based, including their features, strengths, and limitations. The discussion will justify the selection of the (NRC-VAD) list in this research, which provides a standardised measure of affective norms associated with English words and allows for identifying and extracting emotions, opinions, and attitudes expressed in text data.

Machine Learning Techniques

Machine learning techniques are used to classify text into positive, negative, or neutral categories. These techniques can be divided into four main categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning [16].

- **Supervised learning** is the most common approach to machine learning for sentiment analysis. In supervised learning, a model is trained on a dataset of labelled text, where each text is labelled with its sentiment (positive, negative, or neutral). The model learns to predict the sentiment of new text based on the patterns it has learned from the training data. Some examples of supervised learning algorithms for sentiment analysis include:

  - Naive Bayes: A simple algorithm that is fast and easy to implement.
  - Decision trees: A more complex algorithm that can handle more complex data.
  - Support vector machines: A more complex algorithm that can handle more complex data and is more accurate than Naive Bayes or decision trees.

- **Unsupervised learning** is used when there is no labelled data available. In unsupervised learning, the model learns to cluster the text into groups without any prior knowledge of the sentiment of the text. The model then assigns a sentiment to each group based on the characteristics of the group. Some examples of unsupervised learning algorithms for sentiment analysis include:
  - K-means clustering: A simple algorithm that divides the text into k clusters based on the similarity of the text.
  - Latent Dirichlet allocation (LDA): A more complex algorithm that models the text as a mixture of topics, and assigns a sentiment to each topic.

- **Semi-supervised learning** is a hybrid of supervised and unsupervised learning. In semi-supervised learning, the model is trained on a dataset of labelled text and a dataset of unlabelled text. The model learns to predict the sentiment of new text based on the patterns it has learned from the labelled data and the unlabelled data. Some examples of semi-supervised learning algorithms for sentiment analysis include:
  - Label propagation: An algorithm that propagates the labels from the labelled data to the unlabelled data.
  - Self-training: An algorithm that trains a model on the labelled data and then uses the model to label the unlabelled data.

- **Reinforcement learning** is a type of machine learning that is used to train an agent to take actions in an environment in order to maximize a reward. In sentiment analysis, the agent can be trained to take actions such as classifying text as positive, negative, or neutral. Some examples of reinforcement learning algorithms for sentiment analysis include:

- Q-learning: An algorithm that learns to map states to actions that maximize the expected reward.
- Policy gradient: An algorithm that learns to directly optimize the policy for taking actions.

The advantages of machine learning sentiment analysis include its accuracy and flexibility. However, it can be more complex and require more data to train.

The challenges of machine learning sentiment analysis include the need for labelled data, the need for a large training dataset, and the need to choose the right algorithm for the task.

Hybrid Based Techniques

Hybrid approaches combine lexicon-based and machine learning approaches to improve sentiment analysis performance. They can achieve better accuracy than either approach alone by leveraging the strengths of both approaches. The main reason for using a hybrid approach is to inherit the high accuracy of machine learning and the stability of lexicon-based approaches. There are two main ways to combine lexicon-based and machine learning approaches:

1. Feature-based approaches use the lexicon to extract features for the machine learning classifier. For example, the lexicon can be used to identify sentiment words, and the scores of these words can be used as features for the classifier [16].

2. Decision-based approaches use the lexicon to make decisions about the sentiment of the text. For example, the lexicon can be used to identify a set of rules that can be used to classify text into positive, negative, or neutral categories [16].

The choice of which approach to use depends on the specific application. Feature-based approaches are more flexible and can be used with a variety of machine learning classifiers. Decision-based approaches are more efficient and can be used in real-time applications. Here are some examples of hybrid approaches for sentiment analysis:

Hybrid Approach for Sentiment Analysis *HARN* [43] is a hybrid approach that combines lexicon-based and machine learning approaches. HARN first uses a lexicon to identify sentiment words, and then it uses a machine learning classifier to classify the text into positive, negative, or neutral categories [16]. Linguistic Inquiry and Word Count *LIWC* [113] is a lexicon-based approach that uses a dictionary of words to identify sentiment words. LIWC has been used in a variety of studies to analyse the sentiment of text. SentiStrength is a machine learning-based approach that uses a corpus of text to train a classifier [129, 128]. SentiStrength has been shown to be effective in classifying sentiment in a variety of domains. Hybrid approaches have advantages over lexicon-based and machine learning approaches alone. First, they can achieve better accuracy [16]. This is because they combine the strengths of both approaches. Lexicon-based approaches are good at identifying sentiment words, while machine learning approaches are good at learning the nuances of sentiment.

Second, hybrid approaches are more flexible than lexicon-based approaches. This is because they can be used with a variety of machine learning classifiers. Lexicon-based approaches are limited to the sentiment scores that are provided in the lexicon. However, hybrid approaches can use any machine learning classifier, which gives them more flexibility [16].

Hybrid approaches also have some disadvantages. First, they can be more complex than either lexicon-based or machine learning approaches alone. This is because they need to combine the two approaches. However, this complexity can be mitigated by using a simple machine learning classifier.

Second, hybrid approaches can require more data to train the machine learning classifier. This is because the machine learning classifier needs to learn the nuances of sentiment from the data. However, this requirement can be mitigated by using a transfer learning approach, where the machine learning classifier is pre-trained on a large dataset of text.

Other Techniques

In addition to the aforementioned categories of sentiment analysis techniques, there exist alternative approaches that do not neatly align with either the machine learning or lexicon-based paradigms. One notable alternative is ontology-based sentiment analysis [82]. This approach offers a distinctive methodology for assessing sentiment within textual data. In essence, an ontology serves as a meticulously structured, machine-readable representation of domain-specific knowledge, encapsulating concepts, their interconnections, and attributes pertinent to that domain.

The primary distinction between ontology-based and lexicon-based sentiment analysis lies in their textual interpretation and processing methods. Unlike lexicon-based techniques, ontology-based approaches adopt a more nuanced and context-aware perspective when deciphering sentiment in text. Instead of assigning sentiment scores to entire documents or sentences, ontology-based techniques dissect text into specific aspects or facets, associating sentiment scores with each aspect. This approach facilitates a more intricate and detailed understanding of sentiments expressed in complex statements.

Lexicon-based methods are undeniably valuable and widely employed, especially when a quick and broad assessment of sentiment is sufficient. In contrast, ontology-based techniques shine when there is a need for a finer level of granularity, domain-specific knowledge incorporation, and the dissection of textual content into distinct aspects, enabling a more comprehensive exploration of sentiment within intricate textual data. This nuanced approach equips sentiment analysts with a powerful tool to capture subtleties and domain-specific nuances in sentiment expression.

2.3.2 Lexicon Dictionaries

One of the key elements of sentiment analysis is the use of lexicons, which are dictionaries of words that are annotated with sentiment scores or other related attributes. This section discusses some of the most commonly used lexicons in sentiment analysis.

1. Affective Norms of English Words (ANEW) List:
   The ANEW list is a widely used lexicon for sentiment analysis. It provides a standardized measure of the affective norms associated with English words, based on ratings provided by a large sample of English-speaking participants [19, 108, 140]. The ANEW list includes a wide range of words, including both positive and negative sentiment words, and is based on ratings of affective valence, arousal, and dominance. The ANEW list is a valuable resource for sentiment analysis, as it allows for the identification and extraction of emotions, opinions, and attitudes expressed in text data.

- Example: The ANEW list was used in different research works [104, 133, 134]. The lexicon-based method involved using the ANEW list to score the sentiment of individual words in tweets and aggregating the scores of constituent words to calculate the sentiment score for the tweet.

2. NRC Valence, Arousal, and Dominance (VAD) Lexicon:
   The NRC-VAD lexicon is a comprehensive resource for sentiment analysis that provides a standardized measure of the affective norms associated with English words based on ratings provided by a large sample of English-speaking participants [99]. The NRC-VAD lexicon covers 20,007 positive and negative sentiment words and is based on affective valence, arousal, and dominance ratings. The NRC-VAD lexicon is a valuable resource for sentiment analysis, as it identifies and extracts emotions, opinions, and attitudes expressed in text data.

   - Example: The NRC-VAD lexicon has been used in various research works [76, 87, 88], and in the proposed approach for sentiment analysis of tweets, as discussed in chapter 5. The lexicon-based method involved using the NRC-VAD lexicon to score the sentiment of individual words in tweets and aggregating the scores of constituent words to calculate the sentiment score for the tweet. This approach allows for a more accurate representation of sentiment in informal languages, such as tweets, and helps better understand social media users' emotions, opinions, and attitudes.

3. SentiWordNet:
   SentiWordNet is a lexicon that assigns sentiment scores to synsets in WordNet, a lexical database of English. The sentiment scores in SentiWordNet are based on a combination of three factors: positivity, negativity, and objectivity. SentiWordNet has been widely used in sentiment analysis and has been shown to be effective in capturing the nuances of sentiment expressed in text data [10, 53, 102].

   - Example: SentiWordNet has been used in numerous research studies, including a study by Montejo-Ráez et al. [103] on the sentiment of microblogging data. The study used SentiWordNet as a lexical resource to improve the accuracy of sentiment polarity classification on Twitter.

4. VADER (Valence Aware Dictionary and sEntiment Reasoner):
   VADER is a rule-based sentiment analysis tool that uses a lexicon and a set of rules to determine the sentiment of text data. The lexicon used by VADER includes over 7,500 words and phrases annotated with sentiment scores. VADER is particularly well-suited to analysing social media text data, as it is designed to handle the nuances of informal language and the use of emoticons and slang [75].

   - Example: VADER has been used in numerous research studies, including a study by Palomino et al. [111] used the VADER lexicon, along with several other sentiment analysis tools, to analyse a dataset of hotel reviews.

Table 2.1 summarize the features of the lexicons dictionaries used in sentiment analysis.

Table 2.1 Comparison between Lexicon Dictionaries Features

Feature	ANEW	VADER	SentiWordNet	NRC-VAD
Main Purpose	Affective norms	Sentiment analysis	Sentiment analysis	Affective norms
Sentiment Scores	Valence, Arousal, Dominance	Positive, Negative, Neutral	Positive, Negative, Objective	Valence, Arousal, Dominance
Sentiment Dimensions	3 (VAD)	3 (PN, Neu)	3 (PN, Obj)	3 (VAD)
Sentiment Scale	1-9 (each dimension)	-1 to +1 (Compound)	0-1 (each dimension)	0-1 (each dimension)
Scope	~1,000 words	~7,500 tokens (words, phrases)	155,000+ synsets (words)	~20,000 words
Language Coverage	English	English	English	English
Domain Coverage	General	General (incl. social media)	General	General
Emoticons/Emojis	Not included	Included	Not included	Not included
Idioms/Phrases	Not included	Included	Not included	Not included

In the context of this thesis, NRC-VAD list has been chosen for sentiment analysis. The NRC-VAD list provides a standardized measure of affective norms with comprehensive coverage of English words. Leveraging the standardized measure of affective norms provided by the NRC-VAD list, the sentiment analysis model can effectively capture the subtle nuances of sentiment in informal languages like tweets. The NRC-VAD list's widespread use in various research studies underscores its dependability and significance. Consequently, employing the NRC-VAD list aligns with the objectives of the research questions, enabling me to accurately analyze the sentiment found in tweets concerning (New Year's resolutions) dataset.

2.3.3 The Relationship Between Sentiment Analysis and the TPB model

When considering the TPB model, sentiment analysis can provide valuable insight into the attitudes and subjective norms of social media users towards particular topics or products. The attitude towards a behaviour represents the positive or negative evaluation of that behaviour, which can be extracted from social media posts using sentiment analysis. This extracted attitude can then be used to predict the likelihood of that behaviour being performed.

Furthermore, subjective norm refers to the perceived social pressure to perform or not to perform a behaviour. By extracting the opinions of others expressed in social media posts using sentiment analysis, the subjective norm of the user can be predicted. Additionally, perceived behavioural control, which represents the perception of the ease or difficulty of performing a behaviour, can be predicted by extracting the expression of the easiness or difficulty of performing the behaviour in social media posts using sentiment analysis.

Therefore, sentiment analysis is a valuable tool for intention mining, as it provides insights into the attitudes and emotions of social media users, which are crucial in fields such as marketing. Combining sentiment analysis with other data mining and machine learning techniques, such as text mining, information gain feature selection, and neural networks can extract valuable insights from social media data.

However, sentiment analysis is not without its limitations. Factors such as sarcasm, irony, and negation can affect the accuracy of sentiment analysis, leading to incorrect text classification. Therefore, it is crucial to consider the limitations of sentiment analysis and use a combination of techniques to ensure an accurate interpretation of the results.

Regarding the most suitable lexicon for sentiment analysis, the (NRC-VAD) list stands out as a reliable and widely used resource for measuring the affective norms associated with English words. Its comprehensive coverage of words, including positive and negative sentiment words, and ratings of affective valence, arousal, and dominance, make it a valuable resource for sentiment analysis. Hence, in this research, the (NRC-VAD) list have been chosen as the primary lexicon for sentiment analysis to provide reliable and accurate results.

2.4 Overview of Machine Learning and Data Mining

This section delves into the various techniques used in classifying social media posts and their relationship to understanding the intentions and behaviour of users. The utilization of machine learning algorithms and data mining methods are particularly relevant in this context, as they enable the efficient analysis of vast amounts of data found on social media platforms. Additionally, this section explores the application of sentiment analysis in categorizing social media posts. The discussion will also delve into the use of supervised machine learning methods such as Support Vector Machines (SVMs), Neural Networks (NNs), and Decision Trees (DTs) for classifying social media posts. Furthermore, the section will examine feature selection techniques, such as Information Gain and Recursive Feature Elimination (RFE) how they can enhance the performance of these classifiers. The goal of this section is to provide a comprehensive overview of the current state-of-the-art in classification techniques and how they can be applied to the intention mining of social media posts.

2.4.1 Feature Selection

Information gain feature selection (IG) is a technique used in machine learning and data mining to select the most informative features for a given dataset. The technique is based on the concept of information theory and is used to identify the features that have the most impact on the outcome of a given problem. The idea behind IG is to measure the amount of information provided by each feature in a dataset. This is done by calculating the mutual information between the feature and the target variable. The feature that provides the most information about the target variable is considered the most informative and is selected for use in the model.

The process of information gain feature selection can be broken down into several steps:

1. Calculate the entropy of the target variable: The entropy of a variable is a measure of the amount of uncertainty associated with the variable. The entropy of the target variable is calculated to determine the amount of

uncertainty present in the dataset. Entropy of the target variable:

$$H(Y) = -\sum_{i=1}^{n} p(y) \log_2 p(y) \tag{2.1}$$

Where:

- $H(Y)$ is the entropy of the target variable.
- $p(y)$ is the probability of each class in the target variable.

2. Calculate the conditional entropy of each feature: The conditional entropy of a feature is a measure of the amount of uncertainty associated with the feature given the target variable. The conditional entropy of each feature is calculated to determine the amount of information provided by the feature about the target variable. Conditional entropy of a feature X:

$$H(Y|X) = -\sum_{i=1}^{n} p(x,y) \log_2 p(y|x) \tag{2.2}$$

Where:

- $H(Y|X)$ is the conditional entropy of the feature.
- $p(x,y)$ is the joint probability of the feature and the target variable.
- $p(y|x)$ is the conditional probability of the target variable given the feature.

3. Calculate the information gain for each feature: The information gain for a feature is calculated by subtracting the conditional entropy of the feature from the entropy of the target variable. This provides a measure of the amount of information provided by the feature about the target variable. The information Gain for a feature X:

$$IG(X) = H(Y) - H(Y|X) \tag{2.3}$$

Where:

- $IG(X)$ is the Information Gain for the feature.
- $H(Y)$ is the entropy of the target variable.

- $H(Y|X)$ is the conditional entropy of the feature.

4. Select the feature with the highest information gain: The feature with the highest information gain is considered the most informative and is selected for use in the model.

Information gain feature selection is a useful technique for identifying the most informative features in a dataset and can be applied to a wide range of machine learning and data mining problems. However, it should be known that this method only considers the mutual information between the features and the target variable and does not take into account other factors such as computation cost or model interpretability.

Recursive Feature Elimination (RFE) is a backward selection method, meaning it starts with all available features and iteratively eliminates the least important ones until a desired number of features is reached or a predetermined performance metric is achieved [52, 66]. RFE assigns importance scores to each feature and ranks them based on their contribution to the model's performance. The intuition behind RFE is that by removing less informative features, the model's performance can be improved. The RFE algorithm can be summarized in the following steps:

1. Initialization: Begin with the full set of features, denoted by F.

2. Model Training: Train a machine learning model a classifier using the features in F and evaluate its performance using a predefined metric (e.g., accuracy, mean squared error).

3. Feature Ranking: Calculate feature importance scores based on the trained model. The specific method for calculating importance scores depends on the model used.

4. Feature Elimination: Identify and remove the feature with the lowest importance score from F.

5. Convergence Check: Check if the desired number of features has been reached or if a predefined performance threshold has been achieved. If not, return to step 2 with the reduced set of features F.

6. Final Model Training: Once the stopping criterion is met, train the final model using the selected features.

7. Model Evaluation: Evaluate the final model's performance on a separate validation or test dataset to assess its generalization ability.

The process of iteratively removing features continues until the desired number of features is reached or the model's performance stabilizes. RFE thus provides a ranked list of features, allowing practitioners to select the top k features that best contribute to the model's performance. The core of the RFE algorithm involves selecting the feature with the lowest importance score in each iteration. The importance score, $S(f_i)$ for a feature f_i can be defined as:

$$S(f_i) = PerformanceMetric(F - f_i) - PerformanceMetric(F) \tag{2.4}$$

Where:

- F represents the current set of features.
- f_i represents the feature to be eliminated.
- $PerformanceMetric(F)$ is the evaluation metric of the model using the features F.
- $PerformanceMetric(F - f_i)$ is the evaluation metric of the model after removing feature f_i from F.

The feature with the lowest importance score, $S(f_i)$, is eliminated in each iteration.

In Summary, both RFE and IG are effective feature selection techniques, but they have some key differences. RFE is a backward selection method that starts with all the features in the dataset, while IG is a forward selection method that starts with no features. RFE is based on the feature importance score, as determined by a given estimator, while IG is based on the reduction in the uncertainty of the class label, given the feature.

RFE has been shown to be effective in improving the performance of various classifiers, such as Support Vector Machines (SVMs) and Random Forest (RF), while IG has been shown to be effective in improving the performance of Naive Bayes (NB) and Decision Trees (DT) classifiers. Additionally, RFE can be computationally expensive when working with large datasets, while IG is relatively fast.

2.4.2 Classification Techniques

Having discussed feature selection techniques in the previous section, this section will delve further into the classification methods that can be applied to the intention mining of social media posts. It will examine the use of supervised machine learning techniques such as Naive base (NB), Support Vector Machines (SVMs), Neural Networks (NNs), and Decision Trees (DTs) for classifying social media posts. This section aims to build on the previous section by providing an in-depth examination of the various classification techniques and their applicability to the task of intention mining in social media.

Naive Base

Naive Bayes classification is a machine learning technique that is based on the Bayes theorem and the assumption of independence between features. It is a simple and efficient algorithm that is widely used for classification problems, particularly in natural language processing and text classification [86].

The algorithm makes a prediction based on the probability of a given class given the features. It calculates the probability of the class given the features using the Bayes theorem, which is given by the following equation:

$$P(y|x) = P(x|y) * P(y)/P(x) \tag{2.5}$$

where:

- $P(y|x)$ is the probability of the class y given the features x.
- $P(x|y)$ is the likelihood of the features x given the class y.
- $P(y)$ is the prior probability of the class y.
- $P(x)$ is the evidence of the features x.

The algorithm then makes a prediction based on the class with the highest probability.

One of the main advantages of Naive Bayes is its simplicity and efficiency. It requires relatively little training data and is relatively insensitive to irrelevant features. It also performs well when the features are independent of each other, which is a common assumption in many classification problems.

However, one of the main disadvantages of Naive Bayes is the assumption of independence between features, which is often not true in real-world datasets. This can lead to a decrease in accuracy and can cause the algorithm to make sub-optimal predictions. Additionally, the algorithm has trouble dealing with continuous or numerical features and it is not able to handle the interactions between features.

Decision Tree

Decision tree classification is a machine learning technique that is based on a tree-like model of decisions and their possible consequences. It is a powerful algorithm for both classification and regression problems that can handle both categorical and numerical data.

The algorithm works by recursively partitioning the data into subsets based on the values of the input features. It starts at the root of the tree and at each internal node, it selects the feature that best splits the data into subsets. The feature is chosen based on a criterion such as information gain or Gini impurity. The process continues recursively until the subsets are pure, i.e., they contain only observations belonging to one class.

Decision tree classification uses different formulas to determine the best feature to split the data at each internal node. Two commonly used formulas are:

- Information gain formula, mentioned in equation 2.3.
- Gini impurity:

$$Gini(D) = 1 - \sum_{i=1}^{n} p(i|D)^2 \tag{2.6}$$

  Where:

  - $Gini(D)$ is the Gini impurity of the target variable (class).
  - $p(i|D)$ is the proportion of observations in class i for a given dataset D.

The feature with the highest information gain or the lowest Gini impurity is chosen as the feature to split the data at the current node. The process is repeated recursively on the subsets created by the split until the subsets are pure.

One of the main advantages of decision tree classification is that it is easy to understand and interpret, as it represents the decisions and their possible consequences in a clear and visual way. It also doesn't require any assumptions about the distribution of the data, which makes it a non-parametric method. Additionally, it is able to handle both categorical and numerical features, and it is able to handle interactions between features.

However, one of the main disadvantages of decision tree classification is that it is prone to overfitting, especially when it has a large number of features or a small number of observations. This can lead to a decision tree that performs well on the training data but poorly on the test data. Additionally, it can be sensitive to small changes in the data and can be unstable, meaning that small changes in the data can lead to large changes in the tree.

Support Vector Machine

Support Vector Machine (SVM) classification is a machine learning technique that is based on the concept of finding the optimal hyperplane, also known as the maximum margin hyperplane, that separates the data into different classes. The distance between the hyperplane and the closest observations from each class is called the margin.

One of the main advantages of SVM classification is that it is able to handle both linear and non-linear problems, and it is able to handle high-dimensional data. It also has a regularization parameter, which helps to avoid overfitting, and it allows the user to specify different kernel functions for non-linear problems.

However, one of the main disadvantages of SVM classification is that it can be sensitive to the choice of kernel function and parameters, and it can be sensitive to the scale of the data. Additionally, it can be computationally expensive for large datasets, especially when using non-linear kernel functions.

LibSVM is a library for support vector classification and regression that provides an efficient implementation of the SVM algorithm. It offers several kernel functions to handle non-linear problems, including the linear, polynomial and radial basis function (RBF) kernels. The linear kernel is the simplest kernel function. The polynomial kernel is a non-linear kernel function that maps the input data into a higher-dimensional space using polynomial functions. RBF kernel is another non-linear kernel function that maps the input data into a higher-dimensional space using radial functions.

LibSVM also allows users to specify different parameters such as the cost and gamma to control the trade-off between accuracy and complexity. The cost parameter controls the trade-off between maximizing the margin and minimizing the misclassification errors, while the gamma parameter controls the width of RBF kernel.

LibSVM also offers different solvers for the optimization problem, such as the Sequential Minimal Optimization (SMO) algorithm, and the use of these solvers can help to solve the problem more efficiently and with less computational cost.

The mathematical formulation of the classification problem using SVM using linear kernel is be represented in the following equation:

$$min \frac{1}{2}||w||^2 \tag{2.7}$$

$$subject\ to\ y_i(w.x_i + b) \quad \geq 1, i = 1, 2, ..., n \tag{2.8}$$

Where:

- w is the normal vector of the hyperplane.
- x_i is the vector representation of an observation.
- y_i is the class label of an observation.
- b is the bias term.
- $||w||$ is the Euclidean norm of the normal vector w.

The optimization function for non-linear problem is represented in the following equation when using a RBF kernel function:

$$min \frac{1}{2}||w||^2 \tag{2.9}$$

$$subject\ to\ y_i(w.\varphi(x_i) + b) \quad \geq 1, i = 1, 2, ..., n \tag{2.10}$$

Where:

- w is the normal vector of the hyperplane.
- x_i is the vector representation of an observation.
- y_i is the class label of an observation.
- b is the bias term.

- $\varphi(x)$ is the kernel function
- $||w||$ is the Euclidean norm of the normal vector w.

Feed-forward Neural Network

A Neural Network (NN) is a machine learning algorithm that is modelled after the structure and function of the human brain. It is composed of layers of interconnected nodes, called neurons, which process and transmit information. NNs can be used for a wide range of tasks, including classification, regression, and prediction.

The Feed-forward Neural Network (ANN) is one of the most commonly used types of NNs for classification problems. It is composed of an input layer, one or more hidden layers, and an output layer. The input layer receives the input data, the hidden layers process the data, and the output layer produces the output.

The ANN classification algorithm is based on feed-forward learning, where the input data is fed through the network, and the output is produced. The feed-forward process starts at the input layer, where each neuron receives input data and applies an activation function to it. The output of the neurons in the input layer becomes the input to the neurons in the next hidden layer. The process is repeated until the output layer is reached. The output of the neurons in the output layer corresponds to the predicted class label. The algorithm uses backpropagation to adjust the weights and biases of the neurons based on the error between the predicted output and the actual output. The backpropagation process starts at the output layer, where the error is calculated using a loss function such as the cross-entropy loss. The error is then propagated backwards through the network, and the weights and biases of the neurons are adjusted to minimize the error.

The ANN algorithm can be represented mathematically using the following steps:

1. Initialize the weights and biases of the neurons in the network randomly.
2. Feed the input data through the network using the feed-forward process.
3. Calculate the error between the predicted output and the actual output using a loss function, such as the cross-entropy loss. The cross-entropy loss

for a single sample can be represented as:

$$L = -(y * \log(a) + (1-y) * \log(1-a)) \tag{2.11}$$

where:

- L is the error.
- y is the actual output.
- a is the predicted output.

4. Use backpropagation to adjust the weights and biases of the neurons based on the error. The update rule for the weights can be represented as:

$$w = w - \alpha * \frac{\delta L}{\delta w} \tag{2.12}$$

where:

- w is the weight vector of the neuron.
- α is the learning rate.
- $\frac{\delta L}{\delta w}$ is the gradient of the loss function with respect to the weight vector.

The update rule for the biases can be represented as:

$$b = b - \alpha * \frac{\delta L}{\delta b} \tag{2.13}$$

where:

- b is the bias of the neuron.
- α is the learning rate.
- $\frac{\delta L}{\delta b}$ s the gradient of the loss function with respect to the bias.

5. Repeat steps 2 to 4 for a number of iterations until the error is minimized.

6. Use the trained network to classify new input data.

The ANN algorithm is a powerful tool for classification problems, it can handle a wide range of data types and it can learn complex relationships between inputs and outputs. However, it can be computationally expensive, especially when applied to large datasets with many features. Additionally, the choice of architecture, the activation function, and the loss function can affect the performance of the algorithm.

It is also worth noting that there are variations of the NN such as Convolutional Neural Network (CNN) which is used for image classification, and Recurrent Neural Network (RNN) which is used for sequence data.

2.5 Overview of Fuzzy Logic

FL is a mathematical logic that deals with reasoning that is approximate, rather than fixed and precise. It allows for the representation of the concept of partial truth, where the truth value of a statement can range between completely true and completely false. The main idea behind fuzzy logic is to use fuzzy sets, which are sets with a continuum of membership values between 0 and 1, to represent the degree of truth of a statement [147].

FL models are mathematical models that use FL to represent and reason about systems that are subject to uncertainty and imprecision. In the field of artificial intelligence, possibly the easiest way to represent human knowledge is to transform it into natural language expressions in the format of IF-THEN rules [134]. These rules are based on natural language representations and models, which are based on fuzzy sets and fuzzy logic [118]. These models can be used in a wide range of applications, such as control systems, decision-making, image processing, and natural language processing. There are three well-known types of fuzzy rule-based systems, Mamdani, Takagi-Sugeno and Tsukamoto [118, 134, 109]. The first two types of fuzzy rule-based systems apply to regression problems, as the output from such systems is a real value, and the third type generally applies to classification problems, as the output is a discrete value.

2.5.1 Mamdani Model

In 1975, Mamdani and Assilian's influential work [93] introduced the first rule-based controller powered by a fuzzy inference mechanism. The system is generally called a fuzzy-rule-based system (FRBS) [94, 51]. Mamdani FRBS, also known as the Mamdani-Assilian model has been developed by researchers for different application problems.

The Mamdani Fuzzy Model is based on the fuzzy **if-then** rule. The Mamdani style fuzzy inference process is performed in three steps: Fuzzification of input variables, Rule evaluation (inference), and Defuzzification [117, 134].

Fuzzification: This is the process of converting crisp inputs into fuzzy sets. The inputs are mapped to membership functions, which are mathematical functions that assign a degree of membership between 0 and 1 to each input value. The degree of membership represents the degree of truth of the statement that the input belongs to the fuzzy set.

Inference: This is the process of applying a set of fuzzy if-then rules to the fuzzified inputs to infer the fuzzy output. The following equation represents the fuzzy if-then rule:

$$if\, x\, is\, A\, then\, y\, is\, B \tag{2.14}$$

Where x is the input variable, A is the fuzzy set associated with the input variable, y is the output variable, and B is the fuzzy set associated with the output variable. The inference process uses the minimum operator (also known as the "*and*" operator) to combine the fuzzy sets. The minimum operator is represented by $\wedge$ and is defined as:

$$y = min(x1, x2, ..., xn) \tag{2.15}$$

The output of the inference process is a fuzzy set that represents the degree of truth of the statement that the output variable has a certain value.

Defuzzification: This is the process of converting the fuzzy output into a crisp output. The most common method of defuzzification is the centre of gravity method, which is represented by the following equation:

$$y = \frac{(\sum_{i=1}^{n} x_i * y_i))}{(\sum_{i=1}^{n} x_i)}, \text{ where } 0 \leq x_i \leq 1 \text{ and } \sum_{i=1}^{n} x_i > 0 \tag{2.16}$$

Where x_i is the membership value of the i-th fuzzy set, and y_i is the corresponding crisp value. The centre of gravity method calculates the weighted average of the crisp values, where the weights are the membership values of the fuzzy sets.

The Mamdani Fuzzy Model is widely used in control systems, decision-making, and other applications that require the representation of uncertain and imprecise data. It is a simple and intuitive model that allows for the modelling of complex systems by using a set of fuzzy if-then rules. However, it has some limitations, such as the need for a large number of rules to model complex systems and the difficulty of dealing with non-linear systems. Despite these limitations, the Mamdani Fuzzy Model remains a popular choice for many applications, and it can be used to enhance the prediction of the user's behaviour and intention by incorporating it with TPB and sentiment analysis.

2.5.2 Takagi-Sugeno Model

The fuzzy model proposed by Takagi and Sugeno also known as the Takagi-Sugeno-Kang (TSK) model is a type of fuzzy logic model that is based on the fuzzy if-then rule [126, 127]. The main feature of the model is to express the local dynamics of each fuzzy implication rule by a linear system model. The overall fuzzy model of the system is achieved by the fuzzy blending of the linear system models. The model is composed of three main stages: fuzzification, inference, and defuzzification.

Fuzzification: This is the process of converting crisp inputs into fuzzy sets, it is similar to the Mamdani fuzzification process.

Inference: This is the process of applying a set of fuzzy if-then rules to the fuzzified inputs to infer the crisp output. The following equation represents the fuzzy if-then rule:

$$if\, x\, a\, is\, A\, then\, y = f(x) \tag{2.17}$$

Where x is the input variable, A is the fuzzy set associated with the input variable, y is the output variable and $f(x)$ is a linear function of the input variables.

Defuzzification: This step is not needed as the output of the inference is a crisp value, unlike the Mamdani model where the output is a fuzzy set.

The Takagi-Sugeno Fuzzy Model is widely used in control systems and other applications that require the modelling of linear systems. It is a more sophisticated model than the Mamdani model, as it can handle non-linear systems by using a linear function for the output.

Additionally, it's also more computationally efficient as it does not require a defuzzification step. However, it also has some limitations, such as the need to have accurate linear models of the system and the difficulty of dealing with non-linear systems, especially when the linear models are not accurate.

2.5.3 Tsukamoto Model

The Tsukamoto Fuzzy Model, also known as the Tsukamoto-Fuzzy Model, is a type of fuzzy logic model that is based on the fuzzy if-then rule. The model is composed of three main stages: Fuzzification, inference, and defuzzification.

Fuzzification: This is the process of converting crisp inputs into fuzzy sets, similar to Mamdani and Takagi-Sugeno models fuzzification process.

Inference: This is the process of applying a set of fuzzy if-then rules to the fuzzified inputs to infer the fuzzy output, see equation 2.14. The inference process uses the product operator (also known as the "*or*" operator) to combine the fuzzy sets. The product operator is represented by $\vee$ and is defined as:

$$y = max(x_1, x_2, ..., x_n) \tag{2.18}$$

The output of the inference process is a fuzzy set that represents the degree of truth of the statement that the output variable has a certain value.

Defuzzification: This is the process of converting the fuzzy output into a crisp output. The Tsukamoto Fuzzy Model uses the centroid defuzzification method, which is represented by the equation 2.16.

The Tsukamoto Fuzzy Model is widely used in in systems that have a continuous output, unlike Mamdani and Takagi-Sugeno models. However, it has some limitations, such as the need for a large number of rules to model complex systems.

In summary, the Mamdani, Takagi-Sugeno, and Tsukamoto Fuzzy Models are all types of fuzzy logic models that can be used to represent the intention of social media posts based on the TPB model. The Mamdani model is well-suited for simple systems with small data sets, the Takagi-Sugeno model is suitable for linear systems, and the Tsukamoto model is ideal for systems with continuous outputs. It is worth noting that the choice of model will depend on the complexity of the system, the type of data, and the resources available for the modelling process. The use of fuzzy logic models in combination with TPB and sentiment analysis can provide valuable insights into the intentions of social media users by representing the uncertain and imprecise nature of human behaviour.

2.6 Linking FL and TPB to Sentiment Analysis

Fuzzy logic can be used to model the uncertainty and imprecision of the variables in the TPB model, as it allows for the representation of partial truth values. For example, the attitude variable in the TPB model can be represented as a fuzzy set, where the membership value represents the degree of positive or negative attitude towards the behaviour. Similarly, fuzzy logic can be used to model the uncertainty and imprecision of the sentiment analysis, as it allows for the representation of the degree of positivity or negativity of the sentiment. By incorporating fuzzy logic in the TPB and sentiment analysis, it can enhance the prediction of the user's behaviour and intention. This research focuses on the use of the sentiment analysis to extract the TPB model factors ATB, SN, and PBC, which are used to estimate intention in the model. For the ATB, the measure of valence is used to represent the user's attitude regarding behaviour or action, which can be detected from his/her post. Linguistic representation values for valence are set to "positive or negative". Furthermore, user emotion strength (arousal) or the degree of user excitement toward the behaviour is another factor that is considered for evaluating the ATB. Therefore, the arousal of the text is measured for each user and introduced the following linguistic values "calm, excited and contented ". A set of rules is used to extract ATB linguistics based on the valence and arousal values, while the values of ATB are set to "highly positive, positive, low positive, high negative, negative, low negative". The linguistic values are used in this research for the fuzzy logic rule-based system to build how positive or negative the user is toward specific actions within his posts by estimating a crisp value.

2.7 Summary

This chapter provides a comprehensive overview of the relevant literature on sentiment analysis, the TPB, and data mining. The chapter explores how sentiment analysis can be used to extract the user's attitude and emotion towards a particular behaviour or action and how the TPB model can be applied to predict the user's intention to perform that behaviour.

In addition, the chapter provides an overview of the current state of intention mining research, highlighting the various approaches and techniques that have been used to predict user intention. The literature review revealed that the application of sentiment analysis and the TPB model to intention mining had shown promising results, as it provides a way to extract relevant factors that influence intention from user-generated content.

Overall, the chapter contributes to the research questions of this thesis by providing a foundation of knowledge on sentiment analysis, the TPB model, and data mining tools. It sets the stage for the development of the intention mining model, which will be explored in the subsequent chapters.

Given the prolonged extent of this chapter, a comprehensive discourse on social network platforms, social network analysis, and data mining tools have been appended . The discourse will expound on diverse kinds of online social network platforms, such as Facebook, Twitter, and LinkedIn, and how individuals utilize them to foster connections and disseminate information. Subsequently, a thorough investigation of social network analysis will be conducted, which involves employing statistical and computational methods to comprehend the structure and dynamics of social networks. This will encompass an examination of various metrics and techniques deployed in analysing social networks, including centrality measures, community detection, and network visualization. The introduction of data mining tools such as KNIME, WEKA, Orange, and RapidMiner will also be explored, emphasizing their respective strengths and weaknesses concerning data preprocessing, visualization, modelling, evaluation, and deployment. This exposition, which finds its place in Appendix A, has been included due to its perceived importance.

In section A.1, the exploration of various aspects of online social networks with regard to intention mining shall be embarked upon. Commencing with an assessment of the various online social networks platforms, such as Facebook, Twitter, and LinkedIn, and how users employ them to connect and exchange information. An in-depth examination of social network analysis will then be conducted, delving into the process of utilizing statistical and computational methodologies to comprehend the framework and dynamics of social networks. This will encompass a discussion of various metrics and techniques deployed in analysing social networks, such as centrality measures, community detection, and network visualization. The primary objective of this section is to provide a comprehensive overview of the current state-of-the-art in online social network research and its relevance to the task of intention mining.

Furthermore, section A.2 presents a critical comparison of the four data mining tools, taking into account their functionality, user-friendliness, scalability, integration capabilities, and cost. Each tool exhibits its unique strengths and weaknesses, and researchers must assess them judiciously based on their individual needs and research objectives. KNIME has been identified as a superb option for researchers due to its versatility, comprehensive array of data analysis functions, user-friendly interface, and seamless integration with other tools.

Chapter 3

Research Methodology

The methodology chapter provides a comprehensive overview of the research approach and the methods employed to achieve the research aims and objectives. It outlines the steps involved in conducting the study, from the collection and analysis of data to the creation of the FL model for intention. The methodology adopted in this research draws upon both qualitative and quantitative approaches, making use of data mining techniques, sentiment analysis and FL to model the intentions of social network users. This chapter is structured to provide a clear and concise description of the research methodology and the techniques used to carry out the study. The chapter will begin with a brief overview of the research methodology, followed by an explanation of the combination of qualitative and quantitative approaches. The subsequent sections will cover the data collection, feature selection and supervised learning, sentiment analysis, FL model, and data analysis and validation processes in detail.

3.1 Overview of the Research Methodology

This section of the methodology chapter outlines the overall approach that will be taken in conducting this research. The aim of this research is to model the intention of users in a social network by combining data collection, supervised learning, sentiment analysis, and FL. The methodology of this research will involve a combination of qualitative and quantitative approaches. This will include conducting a thorough literature review to gain an understanding of existing theories and models related to the intention of social network users, the application of FL in sentiment analysis, and data collection from various social media platforms. The collected data will then be used to determine the sentiment of users towards products or services and to evaluate their intention to purchase or use them. FL will be used to model the sentiment analysis and identify the influence of users' intentions towards the products or services. The results of this stage will provide an in-depth understanding of the relationships between users' intention, their attitudes, and the social norms that influence their behaviour. Finally, the results of the study will be validated and compared with existing studies to evaluate the validity and reliability of the proposed model.

3.2 Qualitative and Quantitative Approaches

In this research, the methodology will involve a combination of qualitative and quantitative approaches. The qualitative approach involved conducting a thorough literature review to gain an understanding of the existing theories and models related to the intention of social network users and the application of FL in sentiment analysis, see chapter2. This will provide the necessary background information to support the research and aid in the development of the proposed model. The quantitative approach will involve collecting and analysing data to test the proposed model. The data collection methodology will utilize publicly available online datasets and information gathered from various social media platforms such as Twitter and Facebook. The aim is to gather comprehensive and relevant data that will aid in the successful execution of the research objectives. The methodology adopted for the collection of data from social media platforms will be systematic and will adhere to ethical considerations such as data privacy and security. The collected data will then be used to determine the sentiment of users towards products or services and to evaluate their intention to purchase or use them. FL will be used to model the sentiment analysis and identify the influence of users' intentions towards the products or services. The results of this stage will provide an in-depth understanding of the relationships between users' intention, their attitudes, and the social norms that influence their behaviour. The combination of qualitative and quantitative approaches provides a comprehensive and robust methodology for this research. The literature review will provide the necessary background information to support the research, while the collection and analysis of data will provide the empirical evidence necessary to validate the proposed model.

3.3 Data Collection

The Data Collection section of the methodology chapter aims to provide an overview of the process of gathering and processing data for the research. The success of this research heavily relies on the availability and quality of data, as the focus of the study is the examination of social network users. Therefore, it is crucial to have a dataset that contains adequate information about these users.

In this study, big data will be utilized, and data mining analysis tools will be applied to process and analyse the data in order to achieve the research objectives. Chapter 4 will provide a comprehensive overview of the relevant online datasets discussed in previous literature, as well as the data mining tools that will be used in this study. This chapter will also delve into the datasets employed in this research, including the process of constructing the corpus.

The process of constructing the corpus is of utmost importance as it forms the foundation of the analysis. The corpus will be constructed from publicly available online datasets, as well as from social media platforms such as Twitter. Careful consideration will be given to ethical considerations such as data privacy and security during the data collection process.

3.4 Feature Selection and Supervised Learning

The next crucial step in the methodology for building the intention model is the implementation of feature selection and supervised learning, which will be discussed in detail in Chapter 5. The objective of this stage is to extract features from Twitter users that accurately represent the sentiment analysis variables. To achieve this, data mining tools and techniques will be employed to analyse the social network data and identify the most pertinent features for sentiment analysis. Through this process, the research aims to construct a robust and reliable model that accurately reflects the sentiment of users and their intention towards products or services.

3.5 Sentiment Analysis

Sentiment Analysis is the cornerstone of the methodology used in this research, and it is described in detail in Chapter 5. The extracted sentiment analysis variables from the previous step of feature selection and supervised learning will be used to model the TPB. This will entail the utilization of various techniques, including sentiment analysis and natural language processing, to examine the sentiment expressed in the tweets and determine the intended behaviour of the users. The results of this stage will provide a deeper understanding of the interplay between the users' intentions, attitudes, and the social norms that shape their behaviour.

3.6 Fuzzy Logic Model

The final step of the methodology will be outlined in chapter 6. The intention of the users in the social network will be determined by creating a FL model using the sentiment analysis variables modelled in the previous section. This process will involve the application of FL techniques to analyse the sentiment variables and gain an in-depth understanding of the users' intentions. The validity and effectiveness of the FL model will be ensured through appropriate testing and validation techniques and algorithms. The results of this stage will provide insights into the relationships between the users' intention, their attitudes, and the social norms that influence their behaviour.

3.7 Model Validation

The model validation and evaluation methodology will be described in detail in chapter 7. This will focus on the evaluation and validation of the FL model for intention, which was created in the previous section using sentiment analysis variables. This process will involve the use of appropriate techniques and algorithms to test the accuracy and effectiveness of the model and to ensure its validity and reliability. The results of this evaluation will provide insights into the strengths and weaknesses of the model and will help to identify areas for improvement and further research. In this section, the findings of the study will be compared with existing studies to evaluate the validity and reliability of the proposed model.

3.8 Summary

The methodology chapter of this thesis outlines the research approach that will be employed in order to achieve the research aims and objectives and to answer the research questions. The chapter provides a brief overview of the research methodology and explains the combination of qualitative and quantitative approaches that will be used. The chapter also includes detailed sections on each step of the methodology, including data collection, feature selection and supervised learning, sentiment analysis, and the FL model. Each section provides an introduction to the step, explains its importance in the research, and refers to the relevant chapter in the thesis where the step will be discussed in more detail.

Chapter 4

Structuring a Corpus of Online Social Data

This chapter provides an overview of the process of building a corpus of online social data for research purposes. Specifically, it outlines the techniques and best practices for collecting, cleaning, and preprocessing social network data, as well as extracting relevant attributes and structuring the corpus database. The chapter also includes a critical analysis of different attribute sets for social network research, considering both their strengths and limitations.

The chapter begins with an overview of available online datasets and previous literature reviews of such datasets. It then delves into the process of building the corpus, including collecting online social data, hydrating and expanding datasets, and extracting data attributes related to tweets, user profiles, and network data. The chapter also provides an overview of the database structure and details the processes of data cleaning, preprocessing, and quality assessment. Overall, this chapter offers a comprehensive guide for researchers seeking to build and analyse large-scale datasets of online social data. By providing a systematic and rigorous approach to corpus building, it aims to enhance the value and reliability of social network research and facilitate new insights into the complexities of social interactions and communication in the digital age.

4.1 Overview of Online Datasets

This section provides a comprehensive overview of the available online datasets that were used in this research. The aim of this section is to give an understanding of the sources of data that were employed to support the research objectives. The section will describe the various online datasets, their relevance to the research, and the methods used to obtain them. The available online datasets can provide valuable insights into social network users and their intentions, which are essential for building the FL model for intention. This section will provide a foundation for the subsequent sections.

4.1.1 Available Online Datasets

A number of datasets, which have been considered for this research, are available online. These datasets were studied for their suitability in relation to the research problem. A table4.1 lists some of these datasets, including MovieTweetings, Amazon, AffectiveTweets, NewYearResolution 2015, Sentiment140, Dataset Tumblr, Kaggle Twitter US Airline Sentiment, and Gender classifier data.

Table 4.1 Online Datasets

Dataset	**Type**	Source	**Related papers**
MovieTweetings	Movies Recommendation	github	[46]
Amazon	product data	SNAP	[71]
AffectiveTweets	Twitter emotion	waikato	[20]
Sentiment140	Twitter sentiment	waikato	[61]
Dataset Tumblr	Tumblr sentiment	Kaggle	[3]
Kaggle Twitter US Airline Sentiment	Twitter sentiment	Kaggle	[48]
NewYearResolution2015	Twitter sentiment	data.world	[98]
Gender Classifier data	Twitter gender classification	data.world	[38]

- **MovieTweetings**: The MovieTweetings dataset is a collection of movie ratings and reviews obtained from Twitter users intended for use in the development and evaluation of recommendation systems and sentiment analysis algorithms.

  The dataset was created by Simon Dooms and Koen Verstrepen and was released in 2013. It contains ratings and reviews of over 20,000 movies, as well as associated metadata such as the movie title, release year, and director. The MovieTweetings dataset includes two main data files, which are:

  - Rating data file: contains ratings of movies on a scale from 0 to 10, as well as additional information such as the Twitter user ID, movie title, and release year.
  - Review data file: contains reviews of movies written by Twitter users, as well as additional information such as the Twitter user ID, movie title, and release year.

  The MovieTweetings dataset has been used in a variety of research applications, such as movie recommendation, sentiment analysis, and opinion mining. The dataset is particularly useful for developing and evaluating recommendation systems that suggest movies to users based on their past ratings and reviews.

  One limitation of the MovieTweetings dataset is that it only includes ratings and reviews obtained from Twitter users, which may not represent the broader population of moviegoers. Additionally, the dataset may contain biases and anomalies, such as incomplete or inaccurate user profiles or repeated or irregular ratings and reviews.

- **Amazon**: The Amazon dataset is a widely used dataset in the field of product review sentiment analysis. The dataset contains customer reviews of products sold on Amazon.com, as well as associated metadata, such as the product category, rating, and helpfulness votes. The Amazon dataset was created by Julian McAuley and Jure Leskovec at Stanford University in 2014 and is intended for use in the development and evaluation of machine learning models for sentiment analysis and recommendation systems.

  The dataset contains over 142 million reviews of products sold on Amazon.com, covering 24 categories, including books, electronics, home and kitchen, and toys and games. The reviews are written in natural language and range in length from a few words to several paragraphs. The dataset also includes metadata about each review, such as the product ID, reviewer ID, rating, and helpfulness votes [72].

  Despite its size and usefulness, the Amazon dataset has some limitations and challenges that should be considered. For instance, the dataset may contain biased or fake reviews, as it relies on self-reported and unverified reviews by customers. Additionally, the dataset may not capture the full range of opinions and sentiments expressed in product reviews, as it only includes a limited set of categories and products.

- **AffectiveTweets**: The AffectiveTweets dataset is a collection of tweets and their associated emotions and intensities. The dataset was created by Saif Mohammad and Felipe Bravo-Marquez at the National Research Council Canada in 2017 and is intended for use in the development and evaluation of machine learning models for emotion detection and sentiment analysis.

  The AffectiveTweets dataset contains 6,788 tweets, each labelled with one of eight emotions: anger, fear, joy, love, sadness, surprise, thankfulness, and neutral. Each tweet is also labelled with a continuous score indicating the intensity of the emotion, ranging from 0 (no emotion) to 1 (maximum intensity). The tweets in the dataset were collected using the Twitter Application Programming Interface (API) by searching for relevant keywords and hashtags, such as "happy", "sad", "angry", and "fear". The tweets were then manually labelled by multiple human annotators, who were instructed to assign the appropriate emotion and intensity scores based on the content of the tweet[100, 101].

  One limitation of the AffectiveTweets dataset is that it only includes English-language tweets, which may limit its applicability to other languages or cultures. Additionally, the dataset is relatively small compared to other sentiment analysis datasets and may not capture the full range of emotions and sentiments expressed in the social network.

- **NewYearResolution 2015**: A dataset collected from Twitter for sentiment analysis of users' New Year's Resolution hashtag of the year 2015. The tweets, dated between December 2014 and February 2015, feature demographic and geographical data of users in addition to resolution categories and are categorised based on resolutions in ten unique categories (Humour, Personal Growth Health and Fitness, Career, Time Management/Organisation, Family/Friends/Relationships, Recreation and Leisure, Philanthropic, Finance, Education/Training), each of which is labelled manually and combined with the topic of the post [98].

- **Sentiment140**: The Sentiment140 dataset is a widely used dataset in the field of sentiment analysis that contains 1.6 million tweets and their associated sentiments, labelled as positive or negative. Alec Go, Richa Bhayani, and Lei Huang created the dataset at Stanford University in 2009 and designed it to provide a benchmark dataset for researchers to develop and evaluate machine learning models for sentiment analysis [60–64].

The tweets in the dataset were collected using the Twitter API by searching for positive and negative emot-icons, such as ":)", ":D", ":(", and ":/". The tweets were then randomly sampled and labelled by human annotators, with a majority vote used to determine the final sentiment label. The dataset contains an equal number of positive and negative tweets, each with a maximum length of 140 characters. The Sentiment140 dataset has three fields: polarity, tweet ID, and date, where:

- The polarity field indicates the sentiment label of the tweet, either "0" for negative sentiment or "4" for positive sentiment.
- The tweet ID field contains a unique identifier for each tweet.
- The date field indicates the date and time when the tweet was posted in the format "EEE, dd MMM yyyy HH:mm:ss Z".

Despite its popularity and usefulness, the Sentiment140 dataset has some limitations and challenges that should be considered. For instance, the dataset only contains tweets from 2009, which may not reflect the current trends and topics in the social network. Additionally, the dataset may have some inherent biases, as it was labelled by human annotators who may have different interpretations of sentiment and language.

- **Dataset Tumblr**: The Dataset Tumblr is a collection of posts obtained from the social network platform Tumblr, intended for use in the development and evaluation of machine learning models for sentiment analysis, image analysis, and other applications[3].

  The Dataset Tumblr includes posts containing both text and images, as well as posts containing only text. The dataset was manually analysed to identify sentiment, emotion, and other features, and a codebook was created to classify and count the content of each post.

  The Dataset Tumblr has been used in a variety of research applications, such as sentiment analysis, emotion detection, and image classification.

The dataset is particularly useful for developing and evaluating machine learning models that analyse and understand the content and context of posts on social network platforms.

One limitation of the Dataset Tumblr is that it only includes posts obtained from Tumblr, which may not represent the broader population of social network users. Additionally, the dataset may contain biases and anomalies, such as incomplete or inaccurate metadata or repeated or irregular posts. Despite these limitations, the Tumblr dataset is a valuable resource for researchers and practitioners who are interested in sentiment analysis, image analysis, and other machine-learning applications. The dataset provides a diverse collection of posts containing both text and images and can be used to develop and evaluate machine learning models and algorithms for a wide range of social network applications.

- **Kaggle Twitter US Airline Sentiment**: The Kaggle Twitter US Airline Sentiment dataset is a collection of tweets that express sentiments towards major US airlines, such as American Airlines, Delta, United, US Airways, and Virgin America. The dataset was created by Kaggle, an online community for data scientists, and released in 2015.

  The Kaggle Twitter US Airline Sentiment dataset is structured of 7702 users and 14641 tweets, each labelled with one of three sentiment classes: positive, negative, or neutral. The tweets were collected using the Twitter API in February 2015, during which a major snowstorm disrupted flight operations across the US, providing a rich and varied set of sentiments towards airlines [48].

  The Kaggle Twitter US Airline Sentiment dataset includes several data fields, including the text of the tweet, the airline being mentioned, the user's location, and the date and time of the tweet. The dataset also includes metadata about the Twitter users, such as their screen names and follower counts.

  The Kaggle Twitter US Airline Sentiment dataset has been widely used in research to evaluate the performance of various sentiment analysis algorithms, such as Naive Bayes, Support Vector Machines, and Random Forests. The dataset had also been used to develop and evaluate natural language processing and machine learning models that aimed to identify and quantify the sentiment of tweets related to airlines.

One limitation of the Kaggle Twitter US Airline Sentiment dataset is that it only covers a limited time period and may not reflect the broader population of Twitter users or the general sentiment towards airlines. Additionally, the dataset may contain biases and inaccuracies, such as incomplete or inaccurate user profiles or repeated or irregular tweets.

Despite these limitations, the Kaggle Twitter US Airline Sentiment dataset is a valuable resource for researchers and practitioners who are interested in sentiment analysis and natural language processing. The dataset provides a well-labelled and diverse collection of tweets and sentiments towards major US airlines and can be used to benchmark and compare the performance of different machine learning models and sentiment analysis algorithms.

- **Gender classifier data**: The Twitter Gender Classifier dataset is a collection of tweets and associated metadata that is labelled with the gender of the tweet author, intended for use in the development and evaluation of machine learning models for gender classification and text analysis.

  The dataset was created by Ben Hamner, a data scientist, and released in 2014. It contains over 20,000 tweets that were randomly selected from a larger collection of tweets and have been manually labelled with the gender of the tweet author, which is male, female, or brand.

  The Twitter Gender Classifier dataset includes several data fields, including the text of the tweet, the user ID, the user's name and screen name, the user's location, and the date and time of the tweet. The dataset also includes metadata about the Twitter users, such as their follower counts and the number of tweets they have posted.

  The Twitter Gender Classifier dataset has been used in a variety of research applications, such as text classification, natural language processing, and social network analysis. The dataset is particularly useful for developing and evaluating machine learning models that classify the gender of Twitter users based on their tweets and associated metadata.

  One limitation of the Twitter Gender Classifier dataset is that it only covers a limited time period and may not represent the broader population of

Twitter users. Additionally, the dataset may contain biases and inaccuracies, such as incomplete or inaccurate user profiles or repeated or irregular tweets.

Despite these limitations, the Twitter Gender Classifier dataset is a valuable resource for researchers and practitioners who are interested in text analysis, natural language processing, and social network analysis. The dataset provides a well-labelled and diverse collection of tweets and associated metadata and can be used to develop and evaluate machine learning models and algorithms for a wide range of applications related to gender classification and text analysis.

4.1.2 Previous Literature Review of Online Datasets

In recent years, online datasets have played an increasingly important role in the field of data analysis and research. Many research studies have utilized various online datasets to investigate a wide range of topics, including sentiment analysis, text classification, and user behaviour analysis. This section provides a review of the previous literature that has used the online datasets listed in Table4.1 in their research.

MovieTweetings: MovieTweetings is a dataset consisting of ratings on movies that were contained in well-structured tweets on Twitter. This dataset has been used in research studies that focus on sentiment analysis and text classification. For example, [83] utilized the MovieTweetings dataset to develop a sentiment analysis model for movie ratings, while [145] proposed a method for modelling long-tailed user behaviour and used it to train and test the proposed model.

Amazon: The Amazon dataset features product reviews and metadata from Amazon. It has been widely used in research studies that focus on sentiment analysis, product recommendation, and user behaviour analysis. For instance, [23] utilized the Amazon dataset to develop a sentiment analysis model for product reviews, while [139] investigated the effectiveness of different recommendation algorithms for product recommendations.

AffectiveTweets: AffectiveTweets is a WEKA package for analysing the emotion and sentiment of English-written tweets. It has been used in research studies that focus on sentiment analysis and emotion recognition. For example, [101] utilized the AffectiveTweets dataset to develop a sentiment analysis model for English-written tweets, while [100] investigated the effectiveness of different emotion recognition approaches for English-written tweets.

NewYearResolution 2015: The NewYearResolution 2015 dataset was collected from Twitter for sentiment analysis of users' New Year's Resolution hashtags of the year 2015. Although there seems to be no published research paper that used the "NewYearResolution 2015" dataset, this dataset can be valuable for various research questions related to public sentiment and behaviour. For example, researchers can use this dataset to identify popular themes and categories of New Year resolutions, as well as to analyse the sentiment and language used in the tweets. They can also develop and validate machine learning models for sentiment analysis, topic modelling, or other related tasks, using this dataset as a benchmark or training set. However, it is important to note that the "NewYearResolution 2015" dataset has some limitations and challenges that need to be considered. For instance, the dataset is relatively small in size and limited in scope, as it only covers tweets posted during a specific time period and related to a specific topic. Additionally, the dataset may have some inherent bias, as it was collected and annotated by human coders who may have different perspectives and interpretations. Despite these limitations, the "NewYearResolution 2015" dataset can be a valuable resource for researchers who are interested in studying public sentiment and behaviour related to New Year resolutions or other related topics. Future research can build on the insights and findings from this dataset or expand its coverage and diversity to include other social network platforms, time periods, and topics.

Sentiment140: The Sentiment140 dataset is a widely used dataset for sentiment analysis. This dataset has been utilized in numerous research studies that focus on sentiment analysis and text classification. For example, [41] utilized the Sentiment140 dataset to develop a deep learning-based sentiment analysis model, while [68] investigated the effectiveness of different feature representation techniques for sentiment analysis.

Dataset Tumblr: The Dataset Tumblr is a collection of Tumblr meta-data of posts and bloggers, collected via a bootstrapping method and built-in Structured Query Language (SQL) format. It has been used in research studies that focus on user behaviour analysis and text classification. For instance, [18] used a corpus of Tumblr posts and analysed the sentiment expressed in posts containing text, images, and their combination to identify the relationships between different modalities and sentiment expressions, while [2] utilized the dataset to develop a text classification model for Tumblr posts.

Kaggle Twitter US Airline Sentiment: The Kaggle Twitter US Airline Sentiment dataset is structured with 7702 users and 14641 tweets for sentiment analysis job about the problems of each major U.S. airline. It has been used in research studies that focus on sentiment analysis and text classification. For example, [114] utilized the Kaggle Twitter US Airline Sentiment dataset to develop a sentiment analysis model for airline tweets, while [131] investigated the effectiveness of different text classification approaches for airline tweets.

Gender classifier data: The Gender classifier data is a dataset of 20000 records used to train a CrowdFlower AI gender predictor. It has been used in research studies that focus on gender classification and text classification. For example, [54] utilized the Gender classifier data to investigate the effectiveness of different machine learning algorithms for gender classification, while [135] utilized the dataset to develop a text classification model for gender classification.

In conclusion, the use of online datasets in research provides several advantages. Firstly, these datasets are widely available, making it easy for researchers to access and analyse large amounts of data. Secondly, many of these datasets are well-structured and preprocessed, which saves researchers time and effort in data preparation. Additionally, many of these datasets have been used in previous studies, providing a basis for comparison and evaluation of new research findings. Furthermore, the use of these datasets allows for the replication and generalization of research results, providing valuable insights into the robustness of findings.

In the context of this study, the utilization of online datasets such as MovieTweetings, Amazon, AffectiveTweets, NewYearResolution 2015, Sentiment140, Tumblr, Kaggle Twitter US Airline Sentiment, and Gender classifier data will provide a rich source of data for the analysis and investigation of the research problem. These datasets have been widely used in previous studies and have been shown to be effective in providing insights into various aspects of user behaviour, sentiment analysis, and text classification. The use of these datasets will enable the study to build on the previous literature and contribute to the advancement of knowledge in these areas.

4.2 Building the Corpus

The construction of a corpus of tweets from users is a pivotal undertaking in this research, which seeks to address several critical research questions of the utmost importance. Specifically, to explore the use of the TPB in modelling the intention of social network users to perform specific actions. Additionally, this research seeks to investigate the application of FL in handling the inherent uncertainty and vagueness associated with users' decision-making processes as it pertains to modelling social network users' intentions. Lastly, this research endeavours to employ sentiment analysis to assess users' opinions and emotions towards different products, services, and topics in relation to modelling users' intentions on social networks.

To achieve these objectives, a corpus of tweets from users is constructed. This corpus is meticulously preprocessed and structured to facilitate sophisticated analysis. Specifically, any superfluous information, such as URLs or hashtags are removed, cleaned to standardize the text. Once this is completed, the corpus is structured by separating the tweets into individual documents. Subsequently, further text processing is performed, such as stemming and lemmatization, to ensure that words are represented in a consistent form.

Following the construction of the corpus, machine learning models are trained and evaluated that incorporate the TPB and FL to model the intention of social network users. This entails training models to predict the likelihood of users performing specific actions based on their attitudes, subjective norms, and perceived behavioural control. Additionally, sentiment analysis are deployed to evaluate users' intentions and emotions towards different products, services, and topics in relation to modelling users' intentions on social networks.

Ultimately, the aim of this work is to evaluate the efficacy of the proposed approach in modelling social network users' intentions when compared to existing methodologies. This requires a comparison between the developed models and the baseline models, and evaluate their performance using standard evaluation metrics. By constructing a comprehensive corpus and deploying advanced analytical tools, new insights into social network users' intentions are offered. Which contributes to the development of more effective models for predicting and understanding user behaviour on social networks.

4.2.1 Collecting Online Social Data

The collection of social network data is an important step in research that involves the analysis of user intentions and behaviour. In particular, microblogging platforms like Twitter offer a wealth of data that can be analysed to gain insights into users' intentions, opinions, and emotions. However, the collection of this data must be done in a responsible and ethical manner, taking into account issues related to data privacy and transparency.

Twitter is the data source of choice for this research, and the online datasets collected from Twitter are published in anonymized form based on the platform's Developer Agreement and Policy[1]. As a result, the tweets are concealed in indexes, and the datasets include tweet IDs or/and user IDs, making it a challenging task to reuse them. To overcome this issue, Twitter was crawled using Application Provider Interfaces (APIs) to build the dataset.

Twitter APIs make collecting a large number of tweets relatively easy, as they support different programming languages and data mining tools such as R, Python, Java, KNIME, and others tools. To retrieve published posts, specific requirements should be provided, such as defining explicit terms included in the post to be retrieved, the user that published the post, the location of the user, or setting the language of the retrieved posts.

In the first stage of the research, KNIME Twitter API was used with a connection node to collect data. This task required an authentication Key and Access Token. A number of search nodes were added to retrieve a large set of tweets with different search terms and matching specific constraints. To ensure that the collected tweets are relevant to the research topic, specific search terms were defined, such as "wish," "want," "need," "look for," "request," "like," and "going to get." These terms were the core of the dataset domain, which was later expanded in subsequent stages of the research.

In addition, the language of the tweets was set to English, and any non-English tweets were removed, as this research focused on the intention of English-expressed posts. This helped to limit the retrieved results, with 6698 tweets collected in the English language in 15 minutes.

Table 4.2 shows the distribution of the collected tweets according to the search words vector. Each post is assumed to represent a single document. Taking

[1] https://developer.twitter.com/en/developer-terms/agreement-and-policy#ii-restrictions-on-use-of-licensed-materials [Last accessed September 2020]

Table 4.2 The Distribution of the Collected Dataset from Tweeter According to Words Vector

Key_Words	**Number of tweets**
wish	698 (10.4%)
want	1000 (14.9%)
need	1000 (14.9%)
look for	1000 (14.9%)
request	1000 (14.9%)
like	1000 (14.9%)
going to get	1000 (14.9%)

into consideration that the post on Twitter cannot exceed 140 characters, as the rules of the Twitter platform state at the time of collecting the data. However, this number increased later to 280 by Twitter. Based on this assumption, the intention within the post is considered to represent the user's intention and all the words of the post leading to the intention.

Once the tweets were collected, the next step was to process and label them. Each tweet was represented in a document format and labelled by using a string search algorithm combined with manual labelling by human annotators. The tweets were divided into a polarity class set and a target class set, where Yes indicates that the document contains intention words, and No indicates the opposite.

Table 4.3 The Distribution of Class Labelling Occurrence Dataset from Twitter

Class_Label	**Occurrence**
YES	3284 (52.2%)
NO	3005 (47.8%)

In addition to data processing and labelling, it is important to consider the ethical implications of collecting data from social network platforms. The collection of personal information from social network users raises issues around privacy and transparency, which must be taken into account when collecting and analysing data. To address these concerns, the collected tweets were anonymized, with only tweet IDs and user IDs being retained. Additionally, measures were taken to ensure that users' personal information was protected from unauthorized access.

Ultimately, the collection of social network data is a valuable tool for researchers to gain insights into user behaviour and intentions. By using appropriate methods to collect and analyse this data, researchers can contribute to the development of new theories and frameworks for understanding social network users' intentions while also providing valuable insights for businesses and organizations seeking to engage with their target audiences online.

Furthermore, the collection and analysis of social network data is an important field of research in its own right. It is an interdisciplinary field that combines methods from computer science, linguistics, psychology, and sociology, among others. By contributing to this field, researchers can develop new methods and techniques for analysing social network data, as well as contribute to the broader theoretical and practical understanding of social network users' behaviour.

Overall, the ethical and responsible collection of social network data is an important consideration for any research that involves social network users. By following best practices and taking into account issues related to data privacy and transparency, researchers can ensure that their work is conducted in a way that is respectful of users' rights and interests while also providing valuable insights into the behaviour and intentions of social network users.

In addition to ethical considerations, it is important for researchers to be transparent about their methods and provide detailed descriptions of their datasets.

Moreover, the collection of social network data can also present challenges related to bias and representativeness. For example, social network users may not be representative of the broader population, and there may be biases in the types of data that are shared on social network platforms. It is, therefore, important for researchers to carefully consider these issues and to take steps to address them, such as through the use of sampling techniques and statistical methods.

Finally, it is worth noting that the collection of social network data is a rapidly evolving field, with new tools and techniques being developed all the time. Researchers should stay up-to-date with the latest developments in the field and be willing to adapt their methods and approaches as needed. By doing so, they can ensure that their work remains relevant and impactful in the ever-changing landscape of the social network.

4.2.2 Hydrating Dataset

Hydrating dataset is the process of enriching and expanding a dataset with additional information, such as entities, relationships, and context [47, 85, 92]. The process of hydrating a dataset is an integral component of this research, particularly in the analysis of user intentions and behaviour on social network platforms. Specifically, this study utilizes the "NewyearResolution2015" dataset that was collected from Twitter for the purpose of sentiment analysis of users' New Year's resolutions for the year 2015. Further information regarding this dataset is presented in section four.

It is assumed that the intention of each user is expressed in his or her tweets, which are classified based on their resolutions into ten distinct categories, including humour, personal growth, health and fitness, career, time management/organization, family/friends/relationships, recreation and leisure, philanthropic, finance, and education/training. Each category is manually labelled and combined with the tweet topic, such as fitness and health resolutions, which are indicated by phrases such as "eat healthy food," "go to the gym," and "quit smoking."

Prior to using the dataset, it is necessary to ensure that the users or tweets are still in existence and have not been removed, deleted, or hidden (protected). Consequently, the dataset is cleaned of any unavailable users and tweets, with only the publicly available users and posts retained. The dataset comprises 5011 tweets for 4518 users, and following the cleaning process, only 2965 users remained.

The subsequent step in the hydration process is to retrieve the users' timeline over a period of 60 days, based on the assumption that users will post their intentions in the near future. A total of 6931899 posts are retrieved for all the users in the dataset. Additionally, other profile data is retrieved, such as user descriptions, number of posts, number of followers, and friends. This information is divided into two sets of attributes: tweet attributes (At) and user attributes (Au). Furthermore, there are other attributes that can be extracted from these two sets, such as social network attributes (An).

The ethical considerations of collecting data from social network users are crucial to address, as such data collection raises issues around privacy and transparency. It is important for researchers to ensure that their work is conducted in an ethical and responsible manner, taking into account issues related to data privacy and transparency. By following best practices, researchers can ensure that their work is conducted in a way that is respectful of users' rights and interests while also providing valuable insights into the behaviour and intentions of social network users.

Overall, the process of hydrating a dataset is a critical aspect of research in the field of social networks, as it allows researchers to gain insights into user behaviour and intentions. By using appropriate methods to collect and analyse data, researchers can contribute to the development of new theories and frameworks for understanding social network users' intentions while also providing valuable insights for businesses and organizations seeking to engage with their target audiences online. Additionally, the ethical and responsible collection of social network data is essential for any research that involves social network users, and researchers should be transparent about their methods and provide detailed descriptions of their datasets for the sake of reproducibility.

One thing to note is that the process of hydrating a dataset, especially in the context of social network data, can be complex and time-consuming. It involves not only cleaning the dataset and checking for the availability of users and tweets but also retrieving additional metadata and profile data that can provide valuable insights into user behaviour and intentions. Additionally, the process of categorizing and labelling tweets based on their content can be a challenging task that requires careful consideration and attention to detail.

Furthermore, it is important to note that the assumptions and decisions made during the dataset hydration process can have a significant impact on the quality and reliability of the resulting corpus. Therefore, it is crucial for researchers to document and justify their decisions and assumptions in a transparent manner so that others can understand and potentially replicate their work.

Overall, the process of hydrating a dataset is an essential step in building a corpus for social network research. It involves a range of technical and methodological considerations, as well as ethical and privacy concerns. By following best practices and taking these considerations into account, researchers can ensure that their work is conducted in a responsible and ethical manner while also providing valuable insights into the social network users' intentions and behaviours.

4.2.3 Extract Tweets Data Attributes

This section elaborates on the process of extracting user profile attributes from the hydrated "NewyearResolution2015" dataset. As previously stated, the dataset was collected from Twitter and classified into ten categories based on the users' New Year's resolutions for 2015.

To begin, the tweets' different attributes need to be preprocessed and filtered. The **tweets' text (t_{text})** is the main attribute that represents the tweet post. It can hold various useful parts, such as hashtags, mentions, and links. There are several other attributes that can be extracted from the tweets, including **tweets syntax(t_{syn})**, **tweets hashtags (t_{hash})**, **user mentions (u_{men})**, and **links(t_{link})**.

Tweets' syntax (t_{syn}) is checked by examining certain tweets' text syntax and determining its similarity to other tweets based on certain natural linguistic patterns. **Tweets' hashtags (t_{hash}) list** the hashtags included in the tweet, which can indicate the user's interests and their relation to any intention. **User mentions (u_{men})** in tweets, if present, provide information about a specific connection relation or brand. **Links (t_{link})** in tweets can be analysed to provide an understanding of the user's intentions if it is related to a specific need such as shopping, travel, or a job.

The context of tweets is also an essential indicator of the user's intention. Therefore, the same set of words in different contexts can produce different meanings. From the tweets' context, other attributes can be extracted, such as **tweets topics (t_{topic})**, **tweets sentiments (t_s)**, and checking if the tweet is **geo-enabled**.

Tweets topics (t_{topic}) are crucial for estimating users' intentions on Twitter. Users frequently post about topics that concern them more frequently than any other topic. Hence, tweets' topics can be summarized by two features: **tweets frequency per topic ($topic_{freq}$)** and **top topic ($topic_{top}$)** tweeted by the user. The former is used to extract knowledge about the user's interest in that topic and estimate the user's intention towards that topic. The latter gives an indication of the user's interest and needs. **Tweets sentiments(t_s)** can be analysed to measure the **sentiment strength ($s_{strength}$)** of the user's positive or negative tweets regarding a certain topic, estimating the score of the user ATB. The **user's average sentiment score ($s_{u-score}$)** is used to measure the user's interest in that topic and gives an estimation of the user's intention towards a topic. The **sentiment contradiction rank (s_{u-con})** represents the difference between the user's sentiment across a set of topics and the sentiments of the user's neighbours on the same topics[124]. This value indicates the effect of the user's neighbours' connections on his intention and how the neighbours influence the user. The **sentiment inconstancy (s_{t-inco})** is analysed by comparing the sentiment in the tweet against the sentiment in the embedded URL regarding the topic, which could affect the user ATB in a manner of increment or decrement. The **variance of a tweet sentiment over time (s_{t-var})** helps update the estimation of the user's partial intention over time [44]. Finally, checking if the tweet is geo-enabled helps estimate if the user would be able to in perceived behavioural control and attitude.

It is important to note that this process of extracting user profile attributes is a critical aspect of research in the field of social networks. It allows researchers to gain insights into user behaviour and intentions, which are valuable for developing new theories and frameworks for understanding social network users' intentions. However, researchers should ensure that they conduct their work in an ethical and responsible manner, taking into account issues related to data privacy and transparency.

By following best practices and taking these considerations into account, researchers can ensure that their work is conducted in a responsible and ethical manner while also providing valuable insights into the social network users' intentions and behaviours.

Extracting user profile attributes from the social network data is a crucial step in the research process, as it allows researchers to gain a deeper understanding of the characteristics and behaviours of users on these platforms. The process of extracting such attributes involves analysing various aspects of the social network data, including tweet text, syntax, hashtags, user mentions, links, context attributes, topics, and sentiment.

By carefully examining the tweet text and syntax, researchers can identify natural linguistic patterns and check for similarities between tweets. Hashtags and user mention providing valuable insights into a user's interests and connections, while links can provide an understanding of a user's intentions related to certain needs, such as shopping or travel.

Furthermore, examining the context of tweets can reveal important indicators of user intention. By analysing the topics of tweets, researchers can identify the issues and concerns that users are most interested in, as well as the frequency of their posts on those topics. Sentiment analysis can reveal the user's overall attitude and emotional responses to certain topics and issues, as well as the strength and consistency of their sentiment over time.

In addition to the attributes discussed above, other important factors that can be extracted from the social network data include the frequency of user posts, user descriptions, number of posts, number of followers, and friends. These attributes can provide researchers with valuable insights into the characteristics and behaviours of social network users, which can be used to develop new theories and frameworks for understanding user intentions and behaviours.

However, it is important to note that the process of extracting user profile attributes from social network data is not without its challenges. The vast amount of data available on these platforms can be overwhelming, and the quality and reliability of the resulting corpus can be impacted by the assumptions and decisions made during the data extraction process.

4.2.4 Extract User Profiles Attributes

The previous section explores the process of hydrating a dataset, which involved cleaning the dataset and retrieving metadata about the users in order to gain insights into their behaviour and intentions on the social network. This section discusses the process of extracting user **profile attributes (**A_u**)**, which is an essential step in understanding social network users' behaviour and intentions.

Each user profile on Twitter provides valuable metadata that can be extracted and used as attributes in this research. These attributes include the **user's description (**u_{desc}**)**, which provides a brief overview of the user's interests and personality. It is important to note that this attribute may be missing or incomplete in some cases, so only users with clear and complete descriptions are included in this study.

Another important attribute is the user's posting behaviour, which can be extracted from the **number of posts (**u_{post}**)**, **retweets (**u_{ret}**)**, and **replies (**u_{rep}**)**, as well as **mentions of the user by other users (**u_{men}**)**. Additionally, the **number of followers (**u_{fors}**)** and **followees (**u_{fee}**)** can provide insight into the user's popularity and influence on the social network.

The user's geographic location can also be an important attribute, as it can provide insight into their behaviour and intentions. This attribute is extracted from the **user's tweet's geolocation (u_{geo})** and can be used to estimate the user's perceived behavioural control and attitude.

Temporal attributes (A_{u-temp}) are another important set of attributes that can be extracted from the user's behaviour over time. These attributes include the **average number of tweets per day ($u_{\Delta tpd}$)**, which indicates the user's attitude and perceived behavioural control, as well as the user's **posting time signatures (u_{time})**.

Finally, **demographic attributes (U_{Dem})** can provide valuable insights into the user's behaviour and intentions. These attributes include **gender (U_{gen})**, **age (U_{age})**, **marital status (U_{mart})**, **education (U_{edu})**, **career (U_{career})**, **interests (U_{inter})**, and **impact (U_{imp})**. It is important to note that demographic information can be sensitive, and researchers should take care to use this information responsibly and ethically.

By extracting these user profile attributes and combining them with the tweet data attributes discussed in the previous section, researchers can gain a more comprehensive understanding of the social network users' behaviour and intentions. However, it is important to note that the process of extracting these attributes can be complex and time-consuming, and researchers should take care to document their decisions and assumptions in a transparent manner. By following best practices and taking ethical and privacy concerns into account, researchers can ensure that their work is conducted in a responsible and ethical manner while also providing valuable insights into the social network users' intentions and behaviours.

4.2.5 Extract Network Data Attributes

The next step in the dataset hydration process involves the extraction of network attributes. This is achieved by analysing and examining the profiles of all users in the dataset, with a focus on identifying the users' followers and followees. Several networks can be built for each user, such as **followers (N_{fol})** and **followees (N_{fee})** networks, **mentions (N_{men})** and **hashtags (N_{hash})** networks, as well as the **tweet (N_t)** and **retweet networks (N_{ret})**.The following network attributes are extracted:

- In and out degree of centrality (C_{Deg}): This measures the number of connections a user has in the network and provides valuable insights into the user's influence and importance.

- Average clustering coefficient (*CCo*) of the retweet and mention networks associated with each user[124]: This provides an indication of the level of interaction and engagement between users and can reveal patterns of behaviour and intention.

- Page Rank (*PR*)and betweenness centrality (*CB*) of users in both the retweet and mention networks[124]: These attributes are used to measure the importance and influence of a user in the network and can be used to identify key influencers and opinion leaders.

It is worth noting that the extraction of network attributes is a critical aspect of the dataset hydration process, as it provides valuable insights into the structure and dynamics of social network networks. By analysing and understanding the patterns of interaction and influence between users, researchers can gain a deeper understanding of social network users' intentions and behaviours.

4.2.6 Critical Analysis of Attribute Sets for the social network Research

The first attribute set, the tweet data attributes, is crucial to the research question as it provides insights into user intentions based on the content and sentiment of their tweets. By extracting information such as tweet syntax, hashtags, and user mentions, the researcher gains a deeper understanding of the user's interests and intentions. Additionally, analysing the tweet's context attributes, including topics and sentiment, provides a more nuanced understanding of the user's behaviour and intentions. This level of detail allowed for a more accurate estimation of the user's attitude and perceived behavioural control.

The second attribute set, the user profile attributes, is equally important as it provides information about the user's demographic profile, which is essential to understanding their behaviour and intentions. By extracting information such as user description, posting behaviour, and the number of followers and followees, the researcher can identify patterns and trends that can be used to estimate the user's intention towards a given topic. Furthermore, by analysing temporal attributes, such as the average number of tweets per day and posting time signatures, the researcher can track changes in user behaviour over time, providing valuable insights into how user intentions evolve.

The final attribute set, the network attributes, is critical to understanding the social context in which users operate. By examining users' followers and followees, as well as their retweets and mentions, the researcher can identify influential users and understand how users are connected. Analysing centrality measures, such as in and out degree of centrality, average clustering coefficient, page rank, and betweenness centrality, provides further insights into how users are connected and the degree of influence they hold.

Overall, the combination of these attribute sets provides a comprehensive understanding of user intentions and behaviour on social network platforms. By following best practices and taking into account ethical considerations, researchers can leverage these attribute sets to contribute to the development of new theories and frameworks for understanding social network users' intentions while also providing valuable insights for businesses and organizations seeking to engage with their target audiences online.

4.3 Database Structure

The corpus database is designed to store and manage the large amount of data that is generated during the data hydration process, as well as the data extracted from the various attributes. In order to effectively store and manage this data, a database management system such as MongoDB is used.

MongoDB is a NoSQL document-oriented database that is well-suited for storing large volumes of unstructured data. The database structure is designed to store the extracted attributes in a structured and efficient manner, making it easy to access and analyse the data.

The corpus database is structured into different collections, each of which represents a different attribute set. For example, there is a collection for textual attributes, a collection for user attributes, and a collection for network attributes. Within each collection, the records are organized based on the unique user identifier, which allows for easy retrieval and analysis of data for a particular user.

Each record in the database includes a set of fields that correspond to the different attributes that were extracted from the data. For example, the record for a particular user in the user attributes collection might include fields such as user description, posting behaviour, number of followers, and demographic information. Similarly, the record for a particular tweet in the textual attributes collection might include fields such as tweet text, hashtags, and user mentions.

By structuring the data in this way, the corpus database allows for efficient and effective management of large volumes of data, making it possible to easily analyse user behaviour and intentions on the social network platforms. Additionally, by using MongoDB as the database management system, the corpus database is scalable and can easily handle large amounts of data, allowing researchers to continually expand and update their datasets.

4.3.1 Structure of the Corpus Database

The corpus database structure for this research is organised to represent each user profile as a row, with the columns containing the extracted features of the textual attributes, user attributes, and network attributes. These features are divided into three sets: tweets **textual features** (F_t), **user features** (F_u), and **network features** (F_n). The structure of the corpus database is designed to meet the needs of the research questions, specifically for the purpose of estimating user intentions.

The textual features set includes the **tweets' text (**t_{text}**)** and the different textual features it holds, such as hashtags, mentions, and URL links. Hashtags on Twitter are usually used to demonstrate the user's interest in a certain case or a topic. The **average number of hashtags (**$avg.no.hash$**)** included in a tweet and the list of the **most frequently used hashtags by the user (**$hash_{top}$**)** is considered to indicate the user's attitude and interest. Mentions in tweets, if present, represent a way of communication between users and can provide information about a connection or a relation between users or the user's interest in a certain brand. The **number of mentions (**$mention_{count}$**)** for each user in all posts, the **average number of mentions (**$mention_{avg}$**)** in a user post, and the type of the mention as a person or a **brand (**$mention_{type}$**)** is computed.

The **type of URL links(**$link_{type}$**)** in a tweet can indicate the user's intentions if it is related to a certain need. The type of link is labelled as shopping, travel, or job, manually. In addition to the above textual features, the text syntax similarity of certain tweets to other tweets based on a certain natural lingual pattern is also considered.

Moreover, the context of tweets is a crucial indicator of the user's intention. The same set of words in different contexts can produce different meanings. Therefore, the tweets' context is used to extract features such as topics and sentiment. When looking at Twitter users' posts, it becomes evident that users try to post, retweet, and follow the topics that concern them more frequently than any other topics. Hence, tweets' topics are important for estimating users' intentions. The topic feature can be summarised by two features: the frequency of a topic extracted from the posts a user publishes regarding a certain topic over the number of total posts and the top topic tweeted by the user, which gives an indication of the user's interest and needs.

Tweets are described to have positive or negative sentiments and are usually used prior to any future actions. Thus, analysing the sentiment of a tweet has a great role in understanding the behaviour and needs of the users. The following sentiment features are extracted and used in the research: the sentiment strength of tweets as positive or negative, measured for the user's positive or negative tweets regarding a certain topic, and the user's average sentiment score over a period of time to measure the user interest about that topic, and the sentiment contradiction rank, which represents the difference between a user's sentiment across a set of topics and the sentiments of the user's neighbours on the same topics. This value indicates the effect of the user's neighbours' connections on their intention and how the user is influenced by their neighbours. Sentiment inconstancy is analysed by comparing the sentiment in the tweet against the sentiment in the embedded URL regarding the topic, which could affect the user ATB in a manner of increment or decrement. The variance of a tweet sentiment over time helps in updating the estimation of the user's partial intention over time.

The user attributes set includes the user description, posting behaviour, number of followers and followees, geo-coordinate for the tweets, the similarity of the user profile to others, and temporal attributes. These attributes provide useful information about the user's behaviour, interests, and demographics.

The network attributes set includes the users' network structure, including followers, followees, mentions, hashtags, and retweets. These attributes provide a deeper understanding of the users' connections, interactions, and influence within their social network.

Table 4.4 Tweet’s Textual Attributes Using MongoDB

Attribute	**Data Type**
Avg. no. hashtags	float
Hashtag top 5	array
No. mentions	integer
Avg. no. mentions	float
Mention type	string
Link type	string
Syntax similarity	float
Topics top 5	array
Topic frequency	array
Sentiment strength	float
Avg. sentiment score	float

Table 4.5 Tweet’s User Profile Attributes Using MongoDB

Attribute	**Data Type**
User description	string
No. posts	integer
No. retweets	integer
No. replies	integer
No. followers	integer
No. followees	integer
Geo-coordinate	array
Similarity to others	float
Avg. no. tweets per day	float
Posting time signatures	array
Gender	string
Age	integer
Marital status	string
Education	string
Career	string
Interests	array
Impact	float

To store and manage this vast amount of data, a suitable database management system is required. MongoDB is a popular choice for managing large amounts of unstructured data. In the proposed database structure, each user is represented as a document with attributes as fields. The tables 4.4 to 4.6, show the data types used to represent the fields for each tweet’s attribute set using MongoDB.

Table 4.6 Tweet's Network Attributes

Attribute	**Data Type**
Followers network in-degree	integer
Followers network out-degree	integer
Followees network in-degree	integer
Followees network out-degree	integer
Retweet network clustering coefficient	float
Mention network clustering coefficient	float
Page rank in Retweet network	float
Page rank in Mention network	float
Betweenness centrality in Retweet network	float
Betweenness centrality in Mention network	float

The tweets' textual features are divided into three sets: hashtags, mentions, and URL links. Hashtags represent the user's interests and can be used to estimate their attitude towards a topic. The average number of hashtags included in a tweet and the list of the most frequently used hashtags by the user are considered features. Mentions in tweets provide information about a connection or relation between users and are used to calculate the number of mentions and the average number of mentions in a user post. The type of mention, as a person or a brand, is also computed. The type of URL links in a tweet provides an understanding of the user's intentions related to shopping, travel, or job.

The tweets' context is analysed to extract the topic and sentiment features. Two features summarise the topic feature: topic frequency and top topic tweeted by the user. The sentiment is measured in terms of sentiment strength and user average sentiment score.

The user attributes set provides useful information about the user's behaviour, interests, and demographics. The user description, posting behaviour, number of followers and followees, geo-coordinate for the tweets, and similarity of the user profile to others are considered features. Temporal attributes such as the average number of tweets per day and the posting time signatures are also considered.

The network attributes set provides a deeper understanding of the user's connections, interactions, and influence within their social network. In and out degree of centrality, average clustering coefficient, Page Rank, and betweenness centrality of users in both retweet and mention networks are considered as features.

In summary, the proposed database structure includes three attribute sets: tweet textual features, user attributes, and network attributes. MongoDB is used to store and manage this vast amount of unstructured data. The database structure is designed to meet the research questions of the study, which aims to estimate user intention on Twitter based on their behaviour, interests, and connections within their social network.

4.3.2 Data Cleaning and Preprocessing

The process of preparing data for analysis is essential in ensuring accurate and meaningful results. In this research, the data cleaning and preprocessing stage follows after the data collection and hydration stage, where tweets were extracted from the New Year Resolution 2015 dataset. This stage involves several critical steps that ensure the data is suitable for analysis.

The first step in the data cleaning process involves removing unwanted information from the dataset. This information may include irrelevant text, links, or HTML tags that may affect the quality of the data. For instance, Twitter has a limit of 140 characters per tweet. Therefore, tweets exceeding this limit are truncated, resulting in incomplete tweets. To address this, incomplete tweets are removed from the dataset.

The second step is the correction of typographical errors and spelling. Incomplete words misspelt words, and punctuation errors are corrected. This step ensures the uniformity of data, enhancing the accuracy and reliability of the analysis.

The third step involves the removal of stop words. Stop words are commonly used words such as "and," "the," and "is," which do not add any value to the analysis. Removing stop words reduces the size of the dataset, making it more manageable and enhancing the analysis's accuracy.

The fourth step involves the extraction of features from the data. In this research, the extracted features are divided into three sets: tweet textual features, user attributes, and network attributes. These features provide critical information about the user's interests, behaviour, and demographics and their connections and interactions within their social network.

The fifth step is the normalization of the data. Normalization involves converting the data into a consistent format, making it easier to compare and analyse.

For instance, dates may be represented in different formats, such as "day-month-year" or "month-day-year." Normalizing the data converts all dates to a standard format, such as "year-month-day."

The sixth and final step in the data cleaning and preprocessing process involves the identification and handling of missing data. Missing data can be a result of incomplete records or errors in the data collection process. There are various methods of handling missing data, including imputing the missing data with estimated values, deleting records with missing data, or flagging missing data in the analysis. In this research, the records with missing data were excluded.

All these steps were carried out using various text mining and NLP techniques. In the first preprocessing step, all URLs were removed from the documents to avoid confusion when building the word vectors in the following steps. Then followed by using Part of Speech (POS) tagging, Punctuation Erasure, Diacritic Remover, Number Filter, N Chars Filter, Stop-word Filter, Case Converter Filter, Snowball Stemmer, Dictionary Filter, and Case Converter.

For example, the Punctuation Erasure technique removes all punctuation marks from the text, while the Diacritic Remover removes diacritical marks, such as accents, from the text. The Number Filter removes all numerical digits from the text, while the N Chars Filter removes all words that are less than or equal to N characters in length. The Stop-word Filter removes all common words, such as "the" and "and," which do not add any meaning to the text. The Case Converter Filter converts all the text to lowercase or uppercase to ensure consistency. The Snowball Stemmer reduces words to their root form, which reduces the number of unique words in the text, while the Dictionary Filter removes words that are not in a specified dictionary. These preprocessing techniques ensure that the text is clean and ready for feature extraction and classification, thus improving the accuracy of the results.Algorithm 1 displays the steps of the preprocessing procedure.

4.3.3 Data Quality Assessment

The accuracy and reliability of the data in the corpus database are essential to ensure the validity of the research findings. A data quality assessment was performed to ensure the quality and integrity of the data. This was done using

Data: A collection of tweets
Result: A filtered document text with each document record

1. Convert each tweet text to a separate document.
2. Remove all URLs from the tweets.
3. Apply the following filtering techniques to each document:
   (a) Part of speech (POS) tagger: Identify the part of speech for each word in the tweet (e.g. noun, verb, adjective).
   (b) Punctuation erasure: Remove all punctuation marks from the tweet text.
   (c) Diacritic remover: Remove all diacritic marks (e.g. accents) from the tweet text.
   (d) Number filter: Remove all numbers from the tweet text.
   (e) N Chars filter: Remove all words with less than N characters.
   (f) Stop word filter: Remove all stop words (e.g. "and", "the", "in") from the tweet text.
   (g) Case converter: Convert all words to lower- case.
   (h) Snowball stemmer: Convert each word to its base form (e.g. "running" to "run").
   (i) Dictionary filter: Remove all words that are not found in a pre-defined dictionary.
   (j) Case converter: Convert all words to lowercase (again).
4. Output the filtered document text with each document record.

Algorithm 1: Preprocessing Steps Algorithm)

statistical analysis and NLP techniques for cleaning and preprocessing, while sentiment analysis techniques replaced human annotation.

The first step in the data quality assessment process is to check for completeness. This involves examining the data for missing values and determining the percentage of missing data in the dataset. Missing data can be caused by data collection errors, data processing errors, or a lack of data for some variables. If the percentage of missing data is significant, it can affect the reliability and accuracy of the analysis. To address this, the records with missing data were excluded. The percentage of missing data was found to be less than **5%**, which did not significantly affect the analysis's reliability.

The second step is to check for accuracy. This involves comparing the data in the corpus database to external sources to ensure that the data is correct. For instance, the location data of the users can be verified using geocoding services. Inconsistencies in the data can be identified, and measures are taken to correct them. Inconsistencies in the data were identified and corrected using entity recognition and named entity recognition.

The third step is to check for consistency. This involves examining the data to ensure that it is internally consistent. For instance, data about the same user should be consistent across different variables, such as the number of followers, the number of tweets, and the user description. Inconsistencies can be identified, and measures are taken to correct them, such as cross-checking the data or contacting the user for clarification. To check for consistency, the data was examined to ensure it was internally consistent. Inconsistencies were identified, and measures were taken to correct them. Word embeddings and semantic similarity, were used to ensure consistency.

The fourth step is to check for duplication. This involves identifying duplicate data in the corpus database, such as multiple entries for the same user or multiple entries for the same tweet. Duplicate data can affect the accuracy and reliability of the analysis and should be removed. Duplicate data was identified and removed from the corpus database. Cosine similarity and Jaccard similarity, were used to identify duplicate data.

The fifth and final step is to check for validity. This involves ensuring that the data in the corpus database is valid for the research questions being addressed. For instance, the data should be relevant to the research questions and should have been collected using appropriate methods. Inconsistencies can be identified, and measures are taken to correct them by removing irrelevant data.

Statistical analysis tools were used to evaluate the data quality assessment process. These tools allowed for the quantification of the level of completeness, accuracy, consistency, duplication, and validity of the data. The sentiment analysis techniques were used to determine the accuracy and reliability of the data based on the sentiment expressed in the tweets. The sentiment analysis techniques enabled the identification of any inconsistencies or biases in the data and allowed for the correction of these issues.

Cleaning and preprocessing techniques were also used in the data quality assessment process. These techniques involved removing unwanted information, correcting typographical errors and spellings, removing stop words, extracting features, normalizing the data, and handling missing data. These steps ensured that the dataset was cleaned, organized, and made ready for analysis, ensuring accurate and meaningful results.

Re-assessment of the data quality was also performed periodically to ensure that the quality and integrity of the data were maintained throughout the research. This involved repeating the same data quality assessment process and checking for any new issues that may have arisen during the research.

Finally, documentation was an essential part of the data quality assessment process. The documentation included details on the steps taken to ensure the quality and integrity of the data and any issues that were encountered during the process. This documentation was used to track the progress of the data quality assessment process and to ensure that the research findings were based on high-quality data.

4.4 Summary

In this research, the online dataset and data collection chapter have been constructed with great attention to detail to ensure that the research questions are answered accurately and meaningfully. The chapter begins by discussing the available online datasets and the previous literature review of online datasets to identify the most appropriate dataset for this research. This is followed by a thorough explanation of the building of the corpus, which includes collecting online social data, hydrating datasets, and extracting tweets, user profiles, and network data attributes. The critical analysis of attribute sets for social network research is also presented to identify the essential features of the research.

The database structure section is critical in the methodology chapter as it highlights the importance of data cleaning and preprocessing, data quality assessment, and the structure of the corpus database. Data cleaning and preprocessing ensure that the data is suitable for analysis by removing unwanted information, correcting typographical errors, removing stop words, extracting features, normalizing the data, and handling missing data. This process enhances the accuracy and reliability of the analysis.

The data quality assessment is essential to ensure the quality and integrity of the data are maintained. The assessment process involves checking for completeness, accuracy, consistency, duplication, and validity. In this research, statistical analysis and NLP techniques were used to ensure the accuracy and reliability of the data in the corpus database.

The corpus database structure is designed to meet the research questions of the study, which aims to estimate user intention on Twitter based on their behaviour, interests, and connections within their social network. MongoDB is used to store and manage this vast amount of unstructured data. The database structure is organized to represent each user profile as a row, with the columns containing the extracted features of the textual attributes, user attributes, and network attributes.

In conclusion, the online dataset and data collection chapter provide a comprehensive account of the processes involved in building the corpus database for the research questions. The critical analysis of attribute sets for the social network research and the database structure section ensure that the research obtains accurate and meaningful results. The data cleaning and preprocessing, data quality assessment, and the structure of the corpus database are crucial to ensuring that the research achieves its objectives. The next chapter, Feature selection and Supervised learning will use the dataset to extract the features for the research aims.

Chapter 5

Intent Mining: Data Mining and Sentiment Analysis

Chapter 5 is a crucial component of the research as it focuses on the methodology used to estimate user intention using data mining techniques and sentiment analysis, specifically by presenting feature selection and supervised learning. The chapter builds upon the literature reviewed in Chapter 2 and outlines the selection of appropriate feature selection techniques and classifiers, such as Decision Tree, Naive Bayes, and Support Vector Machine, to validate the features' efficacy in predicting user intention. Additionally, the chapter emphasizes the significance of sentiment analysis in this research and its integration into the methodology of modelling the TPB for accurate and reliable results. Overall, this chapter's content is essential to the thesis's overall objective, which aims to accurately predict and analyse user intention in social networks for effective marketing and advertising strategies.

5.1 Intent Mining Using Data Mining Techniques

Intent mining using data mining techniques is a vital part of this research, designed to identify users' intentions in their Twitter posts. To achieve this, the approach involves the application of various data mining techniques on the collected dataset from Twitter. The preprocessing step is necessary to deal with the sparsity and the noisy nature of social media posts. The approach includes two double-phase methods, each consisting of two parts. The first phase involves feature extraction from the dataset, utilizing the Information Gain (IG) algorithm based on entropy, referred to as Phase-One feature extraction. The second phase, named Phase-Two feature extraction, employs two machine learning algorithms back to back, using the Forward-Feature Selection Algorithm (FFS).

The second part of both approaches is the classification of the dataset based on the extracted features using four machine learning algorithms: Decision Tree, Naive Bayes, Support Vector Machine, and Feed-Forward Learner Neural Network. The results of the approach are presented in chapter 7, where it is shown that the feature set can be significantly reduced without significantly impacting classification results. This approaches outcomes demonstrate the effectiveness of data mining techniques in extracting essential features from Twitter data, aiding in the identification of user intentions.

5.1.1 Feature Selection

In order to identify tweets that contain users' intentions, a data mining approach was applied to the collected dataset from Twitter.

To represent the set of features that classify tweets into two classes, Yes if an intention exists in the text and No if there is no intention, a tweet vector was used. The tweet vector ($\overrightarrow{VT}$) is a binary representation of a word vector of terms in the tweet space, with:

$$\overrightarrow{VT} = \{t_1, t_2, t_3, \ldots, t_n\}$$

The feature vectors were constructed using the Bag of Words (BOW) model, which created a vector of unigrams for the terms that exist in the text based on Part of Speech (PoS) tagging that was done in the preprocessing phase.

The BOW model creates a matrix representation of a text corpus, where each row corresponds to a document, and each column corresponds to a unique word in the corpus. The cells in the matrix are filled with the frequency of each word in the corresponding document. PoS tagging is used to identify the grammatical roles of the words in the text, which helps in determining their relevance and importance in the classification process [119].

For example, the tweet "I want to start exercising every day" would be represented by the tweet vector $\overrightarrow{VT} = \{I, want, start, exercising, every, day\}$, and the BOW representation would be a row in the matrix where the columns for "I", "want", "to", "start", "exercising", "every", and "day" would have a value of **1**, and all other columns would have a value of **0**. This approach was used in the feature selection process to identify the most relevant terms for the classification of tweets into **Yes** or**No**.

Another feature selection technique used in this study was Term Frequency (TF), which calculates the frequency of each term in the text and considers these terms as features to be extracted. This approach has been widely used in text classification tasks [144, 11]. For example, in a tweet such as "I plan to start exercising regularly this year," the terms "plan," "start," "exercising," and "regularly" would be extracted as features using the TF method.

After the feature extraction phase, the feature selection algorithm was applied to the dataset to select the most relevant features for classifying tweets into the two categories of **"Yes"** or "**No**" for the presence of intention. The Information Gain (IG) algorithm based on entropy was used in the first phase of feature selection, which selects the features with the highest information gain for classification. In the second phase, the Forward-Feature Selection (FFS) algorithm was applied to further reduce the feature set and select the most important features for classification.

5.1.2 Hybrid Feature Selection Approach

In this subsection, a hybrid feature selection algorithm approach is introduced based on a threshold of the features to specify the number of features that achieve the maximum score in the form of a term vector. The algorithm consists of two phases: the first phase is a hybrid of feature selection based on two different algorithms, Information Gain (IG) and Forward-Feature Selection (FFS). In the beginning the IG algorithm was used to extract the first set of features, and in the second phase of feature selection, the FFS algorithm was applied. FFS built an empty set of features and added one feature at a time, evaluating each feature using Leave-One-Out Cross Validation ($LOOCV$) error to find the best individual feature. The algorithm's accuracy was measured using four classification algorithms, Decision Trees (DT), Support Vector Machine (SVM), Feed-forward Neural Network (ANN), and Naive Bayes (NB).

The FFS threshold was set at the minimum number of features that give the maximum accuracy score, which varies with each algorithm applied. The selected features were expected to be the words that are frequently used in Twitter feeds.

To measure the accuracy of the work, four classification algorithms DT, SVM, ANN, and NB were used. The hybrid feature selection approach is crucial to reducing the number of features and improving the accuracy of the classification algorithms, which is an essential step in predicting users' intentions on social networks.

5.1.3 Preprocessing Data

The preprocessing steps outlined in Chapter 4 were conducted to ensure that the text is ready for analysis, and that the results of the sentiment analysis and feature selection are accurate and meaningful, and to ensure precision in classifying users' intentions on social media. The collected Twitter dataset was cleaned, which often contains noise in the form of informal language and spelling errors. The outcome of this process is a refined document text, with each document record carefully filtered.

5.1.4 Implementing Data Mining Techniques

Feature Selection: First-Approach

In the first feature selection approach, the focus was on extracting the features from posts that could determine intention through the use of the Information Gain (IG) Algorithm, which predicts features based on the terms in a document, thereby reducing the feature dimension and speeding up the classification process. The tweet vector was represented using a document vector, which in turn represented each tweet's term space as a Boolean value, and feature vector dimensions were reduced using Bag of Words Creator (BOW) and Term of Frequency (TF) techniques. The IG algorithm was used to reduce the feature dimension of the document vector, with the number of features set varying based on the dataset, i.e. based on the text of the tweets. The selected features have the highest mutual information, and all feature term Information Gain greater than zero. Four classification models, including Decision Tree Classifier (DT), Support Vector Machine (SVM), Feed-froward Learner Neural Network (NN), and Naive Bayes (NB), were used to train the reduced features.

This approach shows that IG gives features that mostly describe the users' posts and supports the implementation of extracting the user's intention as in this

Algorithm 2: Phase-One Feature Selection using Information Gain (IG)

Data: Preprocessed tweet data set D, set of candidate features C

Result: Set of features F that maximize the information gain

1. Preprocess the tweet dataset using the techniques in algorithm 1.
2. Create a Tweet Vector (VT) to represent each tweet as a set of terms in the tweet space.
3. Apply Bag of Words Creator (BOW) and Term of Frequency (TF) techniques to create a table of terms and their frequencies in the tweet dataset.
4. Calculate the information gain (IG) of each feature in C, and sort them in descending order by IG.
5. Select the top N features from C, based on the IG threshold.
6. Construct a feature vector for each tweet in D, using the selected features.
7. Train a classification model using the feature vectors and one of the four machine learning algorithms (Decision Tree, Naive Bayes, Support Vector Machine, or Feed-Forward Neural Network).
8. Test the model on a separate test set and evaluate the accuracy and other performance metrics.

study's goal. The textual features were then employed using sentiment analysis techniques. The goal of this approach is to accurately determine intention, and in order to do that, it is necessary to extract the most informative features. By using IG as a machine learning technique, the dimension of the feature vectors could be reduced, while the speed of the classification process could be improved. This allowed for the effective classification of tweets into two categories, with the features set determined by the mutual information of the dataset. By training the reduced features using four well-known classification models, the approach proved to be effective in accurately determining intention. However, it is important to note that the features were not determined based on the context of the posts but through the IG algorithm.

Feature Selection : Second-Approach

The second phase of the proposed approach aims to classify tweets based on two feature selection algorithms. The first part involves the use of the Information Gain (IG) Algorithm to extract features from the first phase, and the second part involves a different feature selection algorithm called Forward-Feature Selection Algorithm (FFS). The goal of this phase is to reduce the dimension of the feature vector by setting a threshold to the minimum number of required features. FFS starts with an empty set of features and adds one feature at a time to the set. Four different classification algorithms, including Naive Bayes, Support Vector Machine, Neural Network, and Decision Tree, are applied in this phase.

The Decision Tree (DT) classifier was set up based on C4.5, where the Gini Index was used as the quality measure for splitting with no pruning. The minimum number of nodes was set to 2, and the split point value was calculated according to the mean value of the two attribute values that separate the two partitions. This algorithm is a heuristic algorithm, meaning a decision is obtained locally and does not guarantee to return the globally optimal solution.

Naive Bayes (NB) was one of the classification models that were used in the second phase of the approach. The basic NB classifier is used to decide the right class of the input data by referring to the highest probability values that are calculated by the trainer classifier using the Bayes formula [32, 86]. The class is calculated as follows:

$$P(c_j|d_i) = \frac{P(d_i|c_j).P(c_j)}{P(d_i)} \tag{5.1}$$

Where:

- $P(d_i)$ is the same for all the classes.

The probability of the word feature wk occurrence in a text document is independent of the word's position and the occurrence of other words in the text document. The probability of $P(c_j|d_i)$ would be calculated based on the number of times a word occurs in a document and the number of words in the document, see equation 5.2.

$$P(c_j|d_i) = P(|d_i|)|d_i|! \prod_{k=1}^{|v|} \frac{P(w_k|c_j)}{n_i k!}^{n_i k} \tag{5.2}$$

Where:

- $P(d_i)$ is the same for all the classes.
- $n_i k$ is the number of times that a word occurs in a document.
- $|d_i|$ is the number of words in a document.

Applying the NB for feature selection in the second phase helped to select the most informative features and achieve high accuracy in the classification task.

ANN is another classification model that was used in the second phase. The ANN algorithm is based on Feed-froward Learning with two inner layers, each having 100 output units. The learning rate was set to 0.1, and the XAVIER initialization weight strategy was used with ReLU Activation Function. The number of training iterations was set to one, and the optimization algorithm used was Stochastic Gradient Descent (SGD). The loss function used was Mean Squared Error.

Support Vector Machine (SVM) was the last classification model that was used in the second phase. The SVM algorithm is based on the LibSVM algorithm with an overlapping penalty set to one. The kernel used was Radial Basis Function (RBF) with Gamma equal to one. The proposed hybrid feature selection approach was able to extract the most informative features that led to high accuracy in the classification task.

The reduction of the features helped to reduce data processing time and affecting the accuracy slightly. The results of the two suggested parts of the hybrid feature selection algorithm are presented in chapter 7, showing the reduction of the features improved the classification accuracy slightly. The use of the IG Algorithm and FFS has proven to be a successful combination, as it produces highly accurate classification models with a reduced number of features. Furthermore, these features are not achieved based on the context of the posts but on the extract feature algorithm IG for documents. This approach helps to identify the most important features used by users in their posts, which supports the implementation of extracting user intention. Algorithm 3 summarizes the two- phased feature selection approach.

Algorithm 3: Phase-Two Feature Selection using Forward-Feature Selection (FFS)

Data: Pre-processed tweet data set D, set of candidate features C
Result: Set of features F that maximize the classification accuracy

1. Initialize the feature set F to the empty set.
2. For each feature in C, calculate the classification accuracy using leave-one-out cross-validation (LOOCV).
3. Select the feature with the highest accuracy, add it to F, and remove it from C.
4. For each remaining feature in C, calculate the accuracy of the feature set F, feature using LOOCV.
5. Select the feature that gives the highest accuracy, adds it to F, and remove it from C.
6. Continue steps 4-5 until a stopping criterion is met (e.g., a maximum number of features is reached).
7. Construct a feature vector for each tweet in D, using the selected features in F.
8. Train a classification model using the feature vectors and one of the four machine learning algorithms (Decision Tree, Naive Bayes, Support Vector Machine, or Feed-Forward Neural Network).
9. Test the model on a separate test set and evaluate the accuracy and other performance metrics.

Discussion

In this chapter, two approaches for feature selection in intent mining on Twitter are presented. The first approach focused on extracting the features that specified intention by applying the Information Gain (IG) Algorithm, which is a machine learning technique used to predict features according to the terms in a document [92]. This approach reduces the dimension of the features and speeds up the classification processes. The tweet vector is used to select the set of features that will be used for classifying tweets into the intention or not. The features were extracted using the bag of words creator (BOW) model, which creates a table containing the terms that exist in the document, and the term of frequency (TF) model, which calculates the frequencies of the terms in the documents set. These features were then used to train four well-known classification models: Decision Tree (DT), Support Vector Machine (SVM), Feed-froward Neural Network (ANN), and Naive Bayes (NB).

However, the second phase of the approach was designed to classify tweets based on two feature selection algorithms. The features threshold was set to a minimum number of features required in the form of a terms vector. This phase consists of two parts. The first part is two levels of feature selection based on two different algorithms. The first level uses the IG Algorithm extracted features, and the second level uses a different feature selection algorithm based on the Forward-Feature Selection Algorithm (FFS). FFS is applied with four algorithms to select features: NB, SVM, ANN, and DT. After setting the features, classification algorithms were again used as DT, SVM, ANN, and NB. Applying the hybrid feature selection approach shows an improvement in the accuracy of intention mining as seen in table 7.2.

When applying the one-phase feature selection using IG by itself, it was noticed that the highest measure in accuracy (F-measure) was for training DT. This may be considered as the selected features holding enough information to give a prediction. However, applying two phases to reduce features may not result in a significant difference in the results, yet it significantly reduces the number of features. The reduction of features reduces data processing time, yet the accuracy is slightly less. This is because the decision trees learning method predicts the values of the target variable by learning simple decision rules inferred from the data features, resulting in a relatively high outcome. It is robust to noisy data, and since it is a heuristic algorithm, a decision is obtained locally and does not guarantee to return the globally optimal solution.

Even though text mining techniques are considered highly recommended and widely used, they do not reflect the context of the tweets in a way that represents the intention based on different factors. One of these factors is labelling. The labelling phase needs to be done by experts or by applying a search for a certain string within the texts of the documents. In other words, by labelling any tweets that have a phrase from the intention vector, which may be considered holding an intention. Hence, using word patterns such as "wish" or "request" might represent an intention, but in a different context, it might not. Most of the previous studies made use of human-judge-labelling to accomplish this.

Another factor to consider is including more search words to retrieve the data from Twitter. More word patterns and terms are needed to be taken into consideration. Therefore, more features of social networks need to be included to improve the accuracy of estimating the user's intention on social networks. In conclusion, the two proposed feature selection approaches have shown promising results in extracting features and classifying tweets into the intention or not and can be further improved by including more search words and taking into consideration different factors that affect the context of tweets, such as labelling and human judgment. However, it is important to note that even with the limitations of the dataset, the proposed approaches have proven to be effective in reducing feature dimensions and improving the accuracy of classification algorithms.

Future work can involve expanding the dataset to include more diverse and representative tweets, as well as exploring other feature selection algorithms and machine learning techniques to further improve accuracy. Additionally, incorporating domain-specific knowledge and including more factors that affect the context of tweets can also enhance the performance of the proposed approaches.

Overall, the results of this study contribute to the growing body of research on intent mining on social networks and provide insights into the effectiveness of different feature selection approaches and classification algorithms for this task. In addition, the proposed approaches provide a starting point for addressing the research questions related to identifying the users' intentions on social networks. With further refinement and improvements, these approaches can be an effective tool for social media analysis and inform decision-making processes in various fields such as marketing and public opinion research.

5.2 Sentiment Analysis and Text Mining for Intent Mining

Sentiment analysis for tweets is the process of identifying the sentiment expressed in a tweet, whether it is positive, negative, or neutral. It is an essential technique in social media mining as it helps to analyse the users' opinions and attitudes towards a particular topic or brand. Sentiment analysis is an integral part of intent mining using data mining techniques. Whether the user has a positive or negative intention towards the topic is determined by analysing the sentiment of a tweet.

The process of sentiment analysis involves several steps. The first step is to preprocess the text by removing stop words, punctuation, URLs, and other irrelevant information. The next step is to tokenize the text, which involves breaking down the text into individual words. After that, the words are analysed to determine their polarity, whether they are positive, negative, or neutral.

There are various methods for sentiment analysis, including lexicon-based methods, machine learning methods, and hybrid methods. Lexicon-based methods involve using a pre-built dictionary of words and their associated polarity scores to determine the sentiment of a tweet. Machine learning methods involve training a classifier to identify the sentiment of a tweet based on a set of labelled data. Hybrid methods combine both lexicon-based and machine-learning techniques to improve the accuracy of sentiment analysis.

In the context of intent mining, sentiment analysis is a crucial step in determining whether the user's intention is positive or negative towards a particular topic. The user's intention towards the topic can accurately be predicted by combining sentiment analysis with other data mining techniques, such as feature selection and classification.

Furthermore, the accuracy of sentiment analysis depends on several factors, such as the quality of the data, the size of the dataset, and the selection of the lexicon or machine learning algorithm. In this research, a combination of lexicon-based and machine-learning techniques is used to ensure the highest possible accuracy.

5.2.1 Importance of NRC-VAD list for the Research Questions

The (NRC-VAD) list is a critical resource for this research, as it provides a standardized measure of affective norms for English words. This enables me to measure the affective valence of words and understand how this valence relates to the sentiment expressed in tweets. The use of the (NRC-VAD) list is essential in quantifying the sentiments expressed in tweets, which can be utilized in creating the fuzzy model for intention based on the TPB model.

The research questions in this thesis focus on identifying the factors that influence the intention to engage in a particular behaviour on social media. One of these factors is the sentiment expressed in tweets related to the behaviour. The use of the (NRC-VAD) list in this research's sentiment analysis approach is crucial in identifying the affective norms of the words related to the behaviour and understanding how these norms influence the sentiment expressed in the tweets.

Moreover, the (NRC-VAD) list helps to improve the accuracy of the developed sentiment analysis model by providing a standardized measure of affective norms for English words. This can be used to train the machine learning techniques employed in the developed model, such as Decision Tree, Naive Bayes, Support Vector Machine, and Artificial Neural Network. The incorporation of feature selection in the previous section further enhances the accuracy of the developed model, allow creating a reliable fuzzy model for intention based on the TPB.

The incorporation of the (NRC-VAD) list in the used sentiment analysis approach is particularly relevant to the research questions. It enables me to classify sentiment accurately based on a comprehensive understanding of the affective dimensions associated with individual words. By using a standardized measure of the affective norms associated with English words, the (NRC-VAD) list allows me to capture the nuances of sentiment expressed in informal languages, such as tweets, and gain insights into the attitudes and opinions of Twitter users towards different intentions.

Overall, the use of the (NRC-VAD) list in this research is justified by its reliability and validity as a measure of affective norms, its comprehensive coverage of English words, and its relevance to this thesis research questions. By incorporating the (NRC-VAD) list into the adopted lexicon-based method for sentiment analysis, the sentiment expressed in tweets related to different intentions are accurately captured and gained valuable insights into the emotions, opinions, and attitudes of users on social media.

5.2.2 Sentiment Analysis for Tweets

Sentiment analysis is a technique that is used to identify and extract the emotions, opinions, and attitudes expressed in text data. It can be applied to various types of text, including social media posts like tweets. This section introduces another approach used to analyse Twitter posts to identify features used in intention mining. The aim is to create intention profiles for Twitter users based on textual features. The features were divided into three sets: tweets textual features, user features, and network contextual features, as discussed in chapter 4. This section will concentrate on analysing textual features of tweets only. In this study, sentiment analysis on tweets was conducted by employing a combination of textual analysis techniques, semantic linguistic models, and sentiment analysis techniques using lexicon-based methods, in conjunction with machine learning techniques.

For this part of the research, feature selection techniques were employed to reduce the feature set, and various classifiers, such as DT, NB, SVM, and ANN were utilized to validate the features for predicting whether a user's data signifies an intention. Ultimately, this section of the thesis has been published as a research paper titled "Users' Intentions Based on Twitter Features Using Text Analytics".[1] at Intelligent Data Engineering and Automated Learning – IDEAL 2019 conference [98].

Problem Definition

The problem that this research aims to address is the identification of features that can be used in intention-mining calculations for Twitter users. The dataset consists of a set of Twitter posts for each user, denoted as $T : t_1, t_2, ..., t_n$, where t_{ij} represents a post i for a user u_j, with m users denoted as $U : u_1, u_2, ..., u_m$.

For each tweet t_i in T, there is a set of features that can be extracted, denoted as $F_{ti} : f_{ti1}, f_{ti2}, ..., f_{tik}$. Similarly, for each user u_j in U, there is a set of features

[1] https://doi.org/10.1007/978-3-030-33607-3_14

that can be extracted, denoted as $F_{uj} : f_{uj_1}, f_{uj_2}, ..., f_{uj_k}$, both of which can be extracted from the Twitter stream.

To minimize the feature set for both F_t and F_u, a use of subsets denoted as $\hat{F}_t$ and $\hat{F}_u$ are proposed, which hold the minimum features of both the user and the posts. These subsets are used to build intentional users' profile posts. Therefore, the problem definition is concerned with identifying the relevant features that can be used to construct intention profiles for Twitter users.

Framework

In this study, a lexicon-based method combined with machine learning techniques are utilized to perform sentiment analysis on tweets. The proposed approach is designed to overcome the limitations of traditional methods by incorporating a dictionary of English words with their associated affective norms(valence, arousal, and dominance), known (NRC-VAD) list.

The lexicon-based method involves using the (NRC-VAD) list as a dictionary to score the sentiment of individual words in the tweets. The (NRC-VAD) list is a database of words with their corresponding emotional valence scores on a scale from 0 to 1, where 0 represents a negative emotion, and 1 represents a positive emotion based on its affective valence, arousal, and dominance.

To use the (NRC-VAD) list for sentiment analysis, each tweet is tokenized by breaking it down into individual words. Then each word is looked up in the (NRC-VAD) list and recorded its emotional valence score. Then the sentiment score is calculated for the tweet by averaging the emotional valence scores of all the words in the preprocessed tweet.

For example, consider the tweet, "I had a great day at the beach today". After preprocessing, this tweet is tokenized into individual words: "great", "day", "beach", "today". Then each word are looked up in the (NRC-VAD) list and recorded its emotional valence score. The emotional valence scores for each word are as follows: great: 0.985, day: 0.719, beach: 0.885, today: 0.806.

To calculate the sentiment score for the tweet, all the emotional valence scores are added up and divided by the number of words in the preprocessed tweet: $(0.985 + 0.719 + 0.885 + 0.806)/4 = 0.848$.

The sentiment score for this tweet is 0.848, which is a positive score.

While the lexicon-based method can be effective in determining the sentiment of a tweet, it may not be accurate in cases where the sentiment is expressed using sarcasm or irony. In addition, the (NRC-VAD) list is an important resource for sentiment analysis, as it provides a standardized measure of affective norms for English words. However, it is not without limitations, as it is based on the affective norms of a limited set of words and may not accurately capture the nuances of sentiment expressed in an informal language such as tweets.

To address these issues, the feature selection approach used in the previous section is incorporated to enhance the accuracy of the sentiment analysis model. Machine learning techniques, including DT, NB, SVM, and ANN, are used to classify tweets as positive, negative, or neutral based on their sentiment. The combination of lexicon-based and machine-learning techniques provided a more accurate and comprehensive analysis of tweet sentiment in this study.

This framework employs a variety of linguistic techniques in order to effectively investigate the Twitter posts of users. These techniques encompass the use of textual analysis, semantic linguistic models [17], and sentiment analysis techniques [13]. To implement these techniques, a proposed framework was devised that was divided into three distinct stages, as depicted in Figure 5.1. The first stage of the framework involved the retrieval of Twitter data using API

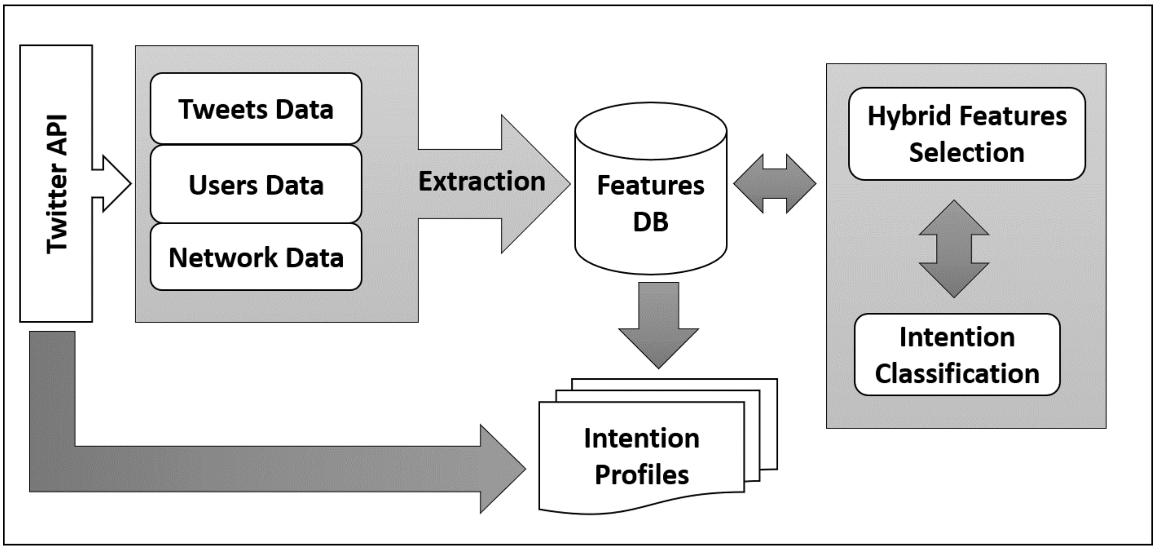


Fig. 5.1 User intention profile system

under certain predefined conditions or queries. The collection of tweets that can be used to extract the three sets of features are the focus as discussed in Section 4. Those features are : tweets textual features (F_t), user features (F_u), and network features (F_n).

The second stage of the framework involved the extraction of textual features from the collected tweets to estimate the user's intention. The features were divided into two categories: topic and sentiment features.

To extract topic features, Latent Dirichlet Allocation (LDA) [17] was used to create ten topics and build model based on words w per topic t and topic t per document d which cover all the posts. Each post was assigned to a topic with the highest confidence score. Each user's posts within a 60-day window interval were used to test the frequency of the specified topic using the created LDA model. A lexicon-based approach was employed using the (NRC-VAD) list for sentiment score. To extract sentiment features, each tweet was tokenized and then each word was looked up in the NRC-VAD list to obtain its VAD scores. These scores were then used to calculate the sentiment strength for each tweet by averaging the valence scores of all the words in the tweet. Moreover, the sentiment average score was calculated for each user by revisiting their 60-day timeline posts and computing the average sentiment strength of their tweets. Using the NRC-VAD list allowed having a reliable and standardized measure of sentiment, as the NRC-VAD list had been extensively validated and provided a broader coverage of English words. This approach also enabled capturing the nuances of sentiment expressed in an informal language such as tweets.

The third stage of the framework involved the extraction of user and network features. The user features included the average number of hashtags included in a tweet, the list of the most frequently used hashtags by the user, the number of mentions for each user in all the posts, the average number of mentions in a user post, and the type of the mentions as a person or a brand. The network features included the follower/followee ratio, the number of likes and retweets, and the number of posts and followers.

In addition, the text syntax similarity of certain tweets to other tweets based on certain natural lingual patterns was analysed to extract user and network features. The type of URL links in a tweet was also considered to understand user intentions if it was related to a certain need. The type of the link was labelled into either shopping, travel, or job manually.

In summary, the proposed approach for sentiment analysis of tweets combined a lexicon-based method using the NRC-VAD list with machine learning techniques. The incorporation of feature selection enhanced the accuracy of the sentiment classification model, which can be applied in a wide range of applications in social media analysis. Algorithm 4 summarizes the proposed framework of using sentiment analysis and tweet textual features for mining intention. By following this algorithm, one can effectively apply sentiment analysis to tweets using a lexicon-based method combined with machine learning techniques, which can be useful in various applications in social media analysis.

Discussion

The importance of sentiment analysis and text mining techniques had been recognized in various fields, including social media analysis. In this section, the approach used for intent mining using sentiment analysis of the tweet textual feature was critically analysed.

The use of the NRC-VAD list for VAD to extract sentiment strength from the tweet textual feature is a commonly used approach. However, the NRC-VAD list for VAD has its limitations, as it is based on a limited set of words that have been rated for valence, arousal, and dominance. Thus, it may not be suitable for analysing social media texts that contain informal language, slang, or new words that are not in the NRC-VAD list. Furthermore, it is worth noting that sentiment analysis is not always reliable, as context plays a crucial role in determining the sentiment of a text. Therefore, the sentiment approach used for intent mining in the tweet textual feature may not be entirely accurate in predicting user intention.

The topic modelling approach used in the sentiment analysis technique is LDA, which is a generative probabilistic model that can create topics based on words per topic and topic per document. The use of LDA is an excellent approach to creating topics and modelling them based on words in each topic. However, selecting the number of topics to create can be challenging, as it is subjective and depends on the researcher's preference. In this approach, ten topics are selected based on the number of predefined topics in the dataset. Additionally, the topics generated using LDA may not always accurately represent the underlying themes in the tweet textual feature. Overall, the sentiment approach used for intent mining in the tweet textual feature has its limitations. The NRC-VAD list for VAD may not be suitable for analysing social media texts that contain informal language, slang, or new words. Additionally, the LDA topic modelling approach used may not accurately represent the underlying themes in the tweet textual feature. Finally, the next chapter introduces the use of the

TPB model in conjunction with FL to improve the accuracy and reliability of the prediction model, especially when dealing with imprecise and incomplete information. This approach can help to overcome the limitations of the current approach and provide a more comprehensive understanding of user intention in social networks.

5.3 Summary

This chapter focuses on the methodological approach employed to discern user intentions by utilising data mining techniques and sentiment analysis, with a particular emphasis on feature selection and supervised learning. This eminent chapter builds upon the academic literature examined in Chapter 2. It elucidates the process of selecting appropriate feature selection techniques and classifiers, such as the esteemed Decision Tree, Naive Bayes, and Support Vector Machine, to corroborate the efficacy of features in predicting user intentions.

Moreover, the chapter accentuates the significance of sentiment analysis within this scholarly research. It explicates its seamless integration into the methodology for modelling the Theory of Planned Behaviour, ensuring accurate and reliable results. In essence, the contents of this chapter are of paramount importance to the overarching objective of the thesis, which aspires to astutely predict and analyse user intentions within social networks, thereby facilitating the creation of efficacious marketing and advertising strategies.

Algorithm 4: Intention Prediction for Tweets Steps Algorithm

Data: A collection of tweets
Result: Intention Prediction Model for users

1. Preprocess the tweets by removing noise, stop words, and stemming, as in algorithm 1.
2. Apply the lexicon-based method by scoring the sentiment of individual words using the NRC-VAD list.
3. Aggregate the scores of all the words in the tweet to calculate the overall sentiment score for each tweet.
4. Employ machine learning techniques, including Decision Tree, Naive Bayes, Support Vector Machine, and Artificial Neural Network, to train and evaluate the sentiment classification model.
5. Use the feature selection approach to identify the most informative features in the tweet data that contribute to the accurate classification of sentiment, as in algorithm 3.
6. Extract tweet textual features (F_t), user features (F_u), and network features (F_n) from the tweet data.
7. Use Latent Dirichlet Allocation (LDA) to extract topics and topic frequency (t_{freq}) for each user.
8. Use the NRC-VAD list to extract sentiment and sentiment strength $(s_{strength})$ of each tweet and compute the user average sentiment score (s_{score}) over a period of time.
9. Combine the tweet textual features, user features, and network features to construct a feature set F.
10. Use feature selection techniques to select the most informative features in F for intention prediction.
11. Train a machine learning model using the selected features to predict user intention.
12. Evaluate the accuracy of the intention prediction model.
13. Interpret the results of the intention prediction while considering the limitations of the feature sets.

Chapter 6

Proposed Fuzzy Logic Model for Intent in Social Networks

In this research, sentiment analysis is used in conjunction with the TPB model introduced in section 2.1.2 to estimate user intention towards a certain behaviour or action. The three sentiment concepts - valence, arousal, and dominance - are used as proxies for the user's attitudes towards the action in question. By combining FL and TPB, a more accurate and reliable model can be built, especially when dealing with imprecise and incomplete information.

The proposed fuzzy intention model employs a set of rules to predict the linguistic values of the TPB factor ATB based on the valence and arousal values of the user's posts. These values are then assigned to "positive, negative, and neutral" based on the results of the sentiment analysis. By estimating a crisp value for the user's attitude, the model can then predict their intention towards the action.

By incorporating sentiment analysis with the TPB model, the proposed fuzzy intention model provides a comprehensive approach to understanding the complex relationship between user attitudes and intentions in social media. This can be particularly useful for businesses and organizations seeking to better understand their customers' attitudes and intentions towards their products or services. By leveraging sentiment analysis in combination with TPB factors such as ATB, SN, and PBC, the proposed model offers a powerful tool for predicting and influencing user behaviour.

The chapter will provide an in-depth discussion of the methods used to construct and evaluate the fuzzy intention model, highlighting its potential applications in various domains.

6.1 TPB and Sentiment Analysis

The sentiment is the contextual polarity attitude of a speaker/writer toward a topic or an event rather than the actual emotions [143]. Sentiment analysis involves the emotional evaluation and investigation of a text. There are three concepts that standardize this evaluation; which are valence, arousal and dominance.

Valence refers to the polarity of emotions as in the "pleasant/unpleasant" or "positive/negative" dimension. Usually, adjectives such as "happy/unhappy", "pleased/annoyed", and "satisfied/unsatisfied" are used to describe the person's level of valence [12]. For the ATB, the measure of valence to represent the user's attitude regarding behaviour or action was used, which can be detected from his/her post. Linguistic representation values for valence are set to "positive, negative, or neutral" [133].

Arousal is used to describe the degree of emotional intensity and reactivity which are provoked by a stimulus. Moreover, arousal is considered a mental activity describing the state of feeling along a single dimension ranging from sleep to frantic excitement and is usually linked to adjectives such as "stimulated/relaxed", "excited/calm", and "wide, awake/sleepy" [125]. Furthermore, user emotion strength (arousal) or the degree of user excitement toward the behaviour is another factor to consider for evaluating the ATB. Therefore, the arousal of the text is measured for each user and introduced the following linguistic values "low, medium and high" [133].

Dominance is the extent to which a word denotes something that is between "weak/submissive" and "strong/dominant", including the in-between "Medium", in other words, the degree of control used by a stimulus [133]. Moreover, dominance is related to the feeling of control and the extent to which an individual feels restricted in his actions and behaviours. Adjectives such as "controlling" refer to the user's dominance over an action [12].

The ATB, SN, and PBC factors in TPB are tightly linked to one or more of these sentiment concepts. The ATB factor, for example, can be evaluated using valence and arousal values. In this research, linguistic values are used in a FL rule-based system to estimate how positive or negative or neutral the user is toward specific actions within their posts.

6.2 Modelling TPB Intention using Sentiment Analysis and Tweets Features

6.2.1 ATB Model

The Attitude Toward Behaviour (ATB) (Attitude, Potency, Activity) model proposed by Friedkin [56] and explained by Tyshchuk [133] measures attitude based on three dimensions: evaluation, potency, and activity. These dimensions can be measured using sentiment, dominance, and arousal values, respectively. The equation to calculate the final score of attitude using the ATB model is:

$$ATB_{Score} = Valence_{Score} * Dominance_{Score} * Arousal_{Value} \tag{6.1}$$

Where:

- Valence (Sentiment) Score: takes the value of -1, 0, or 1 for negative, neutral, and positive valence, respectively. The valence score captures the evaluative dimension of the attitude and reflects the degree to which the user's tweet expresses a positive or negative sentiment [7, 132].

- Dominance Score: represents the dimension of dominance or control in the attitude. It takes the value of -1, 0, or 1 for weak, medium, and strong dominance, respectively. The dominance score captures the degree to which the user's tweet expresses a sense of control or power over the topic of the tweet.

- Arousal Value: represents the potency of the attitude. It takes the value of -1, 0, or 1 for low, neutral, and high arousal, respectively. The arousal value captures the degree of emotional intensity or activation associated with the user's tweet.

The valence, dominance, and arousal scores are combined to calculate the final attitude score using the multiplication operator. This approach assumes that the dimensions of evaluation, dominance, and potency are independent and can be combined multiplicatively. This may not always be the case, as the dimensions of attitude may interact in complex ways, and their effects may not be easily separable.

The final score represents the user's overall attitude towards the topic of the tweet. A positive score indicates a positive attitude, while a negative score indicates a negative attitude. The magnitude of the score reflects the strength of the attitude, with larger magnitudes indicating stronger attitudes.

It is worth noting that the ATB model is a simplification of the complex nature of attitudes and their formation. A variety of factors, including social norms, personal values, and past experiences influence attitudes. While the ATB model provides a useful framework for measuring attitudes based on valence, dominance, and arousal values, it is important to consider these limitations and interpret the results of the model accordingly.

6.2.2 SN Model

Subjective norm (SN) is a critical construct in the field of social psychology that reflects an individual's perception of the social pressure to conform to a specific behaviour [74]. It is defined as a combination of *normative beliefs* (NB) and *motivation to comply* (MC) [133], which can be mathematically represented as

$$SN = f(NB, MC) \tag{6.2}$$

- **Normative beliefs (NB)** refers to the perceived favourable behaviour of important others or the perceived actual behaviour of others. In this context, NB is represented by a combination of the behaviour referred to by each user and the polarity (positive, negative, or neutral) for that behaviour [132, 133]. The equation 6.3 can mathematically represent this as follows:

$$NB = f(BP, AP) \tag{6.3}$$

  Where BP represents *belief propensity*, and AP represents *anticipated social pressure*.

  - The BP is the degree to which a person tends to conform to the beliefs of others and can be calculated based on the number of likes (L), number of followers (F), and number of followees (FO) using the function:

$$BP = f(L, F, FO) \tag{6.4}$$

Where $f()$ function is calculated by:

$$BP = log(L+1) - log(F+FO+1) \tag{6.5}$$

The BP equation measures the degree to which a person tends to conform to the beliefs of others. This representation takes the logarithm of the number of likes, followers, and followees and normalizes them to a range of -1 to 1, where -1 represents non-conformity, and 1 represents full conformity.

The advantage of using the logarithmic representation is that it places greater emphasis on the influence of likes, followers, and followees that are significantly higher or lower than the average while de-emphasizing the influence of those that are closer to the mean. This results in a more accurate representation of the user's tendency to conform to others' beliefs.

– The AP is the degree to which a person expects to receive pressure from others to conform and can be calculated based on the number of retweets (R), number of tweets (T), and number of followers (F) using the function:

$$AP = f(R, T, F) \tag{6.6}$$

Where $f()$ function is calculated by:

$$AP = log(R+T) - log(F+1) \tag{6.7}$$

The AP equation measures the degree to which a person expects to receive pressure from others to conform. This representation takes the logarithm of the number of retweets, tweets, and followers and normalizes them to a range of -1 to 1, where -1 represents no pressure and 1 represents high pressure to conform.

The advantage of using the logarithmic representation is that it places greater emphasis on the influence of retweets, tweets, and followers that are significantly higher or lower than the average while de-emphasizing the influence of those that are closer to the mean. This results in a more accurate representation of the user's expectation to receive pressure from others to conform.

- **Motivation to comply (MC)** reflects an individual's motivation to comply with the perceived expectations of others [6, 91]. It is influenced by factors such as the distance from the normative position (DN) and the level of identification with the group (G), in addition to BP and AP are the same variables used to calculate the NB equation 6.3.

$$MC = f(BP, AP, DN, G) \tag{6.8}$$

  - DN represents the degree to which a person's current behaviour deviates from the normative behaviour and can be calculated based on the sentiment of the tweet or valence (V) and the level of engagement (number of likes, retweets, and replies) (LE) using the function:

$$DN = f(V, LE). \tag{6.9}$$

    where $f()$ is calculated by:

$$DN = \frac{(V - \mu_V)}{\sigma_V} + \frac{(LE - \mu_{LE})}{\sigma_{LE}} + \frac{(V - \mu_{V'})}{\sigma_{V'}} + \frac{(LE - \mu_{LE'})}{\sigma_{LE'}} \tag{6.10}$$

    Where:

    * V is the sentiment or valence of the tweet
    * μ_V is the average sentiment or valence of the user's own tweets
    * σ_V is the standard deviation of the sentiment or valence of the user's own tweets
    * LE is the level of engagement (likes, retweets, and replies) of the tweet
    * μ_{LE} is the average level of engagement of the user's own tweets
    * σ_{LE} is the standard deviation of the level of engagement of the user's own tweets
    * $\mu_{V'}$ is the average sentiment or valence of the user's social network

  * $\sigma_{V'}$ is the standard deviation of the sentiment or valence of the user's social network
  * $\mu_{LE'}$ is the average level of engagement of the user's social network
  * $\sigma_{LE'}$ is the standard deviation of the level of engagement of the user's social network.

  The overall equation for DN is suggested based on the critical reasoning that behaviour deviation can be influenced by multiple factors, including the sentiment or valence of the tweet and the level of engagement, both of which can vary significantly depending on the user's own history and social network. In order to capture these factors in a meaningful way, the equation breaks down the DN score into separate components that quantify the deviation of the tweet's sentiment and engagement from the user's own history, as well as the deviation from the user's social network.

  By weighting each component separately, the equation ensures that the different factors are given appropriate consideration and that the final score reflects the relative importance of each factor. The use of z-score normalization with the standard deviation also allows for fair comparisons across different users and tweets, which may have different baseline levels of engagement or sentiment.

- G represents the degree to which a person identifies with the group that holds the normative beliefs and can be calculated based on the use of specific hashtags or mentions associated with the group (HT), the number of retweets of group-related content (R), and the use of language or jargon specific to the group (LJ) using the function:

$$G = f(HT, R, LJ) \tag{6.11}$$

  where $f()$ is calculated by:

$$G = w1 * HT + w2 * R + w3 * LJ \tag{6.12}$$

  The use of the equation for G is based on the assumption that individuals who use specific hashtags, mentions, or jargon related to a certain group are more likely to identify with that group and adhere to its normative beliefs. Therefore, by calculating G based on these factors, can estimate the degree to which a person identifies with the group and holds the normative beliefs of the group. The weights assigned to each factor in Equation 6.12 reflect the relative importance of each factor in contributing to the overall value of G. The critical reasoning for using this equation lies in the assumption that group identification and adherence to normative beliefs are closely related and can be estimated based on observable factors related to group membership and participation. However, the accuracy of this equation may be limited by the fact that not all individuals who use group-related hashtags or jargon necessarily identify strongly with the group or adhere to its normative beliefs, and there may be other factors that contribute to group identification and adherence.

SN, NB, and MC are all numerical scores calculated based on different features of Twitter tweet objects. These features represent the degree of agreement or conformity with the perceived behaviour of others, the degree of anticipated social pressure to conform, the degree to which a person's current behaviour deviates from the normative behaviour, and the degree to which a person identifies with the group that holds the normative beliefs. Finally, SN is calculated by combining NB and MC using a weighted sum, as represented by the equation:

$$SN = w1 * NB + w2 * MC \tag{6.13}$$

Where w1 and w2 are weights assigned to each variable to reflect their relative importance in determining the user's behaviour.

The mathematical models and functions used to calculate SN, NB, and MC are based on sound theoretical foundations and empirical evidence [133].

6.2.3 PBC Model

Perceived behavioural control (PBC) is a construct that represents an individual's belief in their ability to perform a specific behaviour. PBC is an important determinant of behaviour, as individuals who perceive a higher level of control over behaviour are more likely to perform that behaviour. In the context of social media, PBC can be modelled using a product of values extracted for emotion and arousal [132, 133], represented as PBC in the following equation:

$$PBC = Sentiment * Arousal \tag{6.14}$$

where Emotion is a numerical score representing the emotional valence of the user's tweet (e.g., -1 for negative, 0 for neutral, 1 for positive), and Arousal is a numerical score representing the level of emotional arousal in the user's tweet. The higher the values of Emotion and Arousal, the higher the perceived level of control over the behaviour.

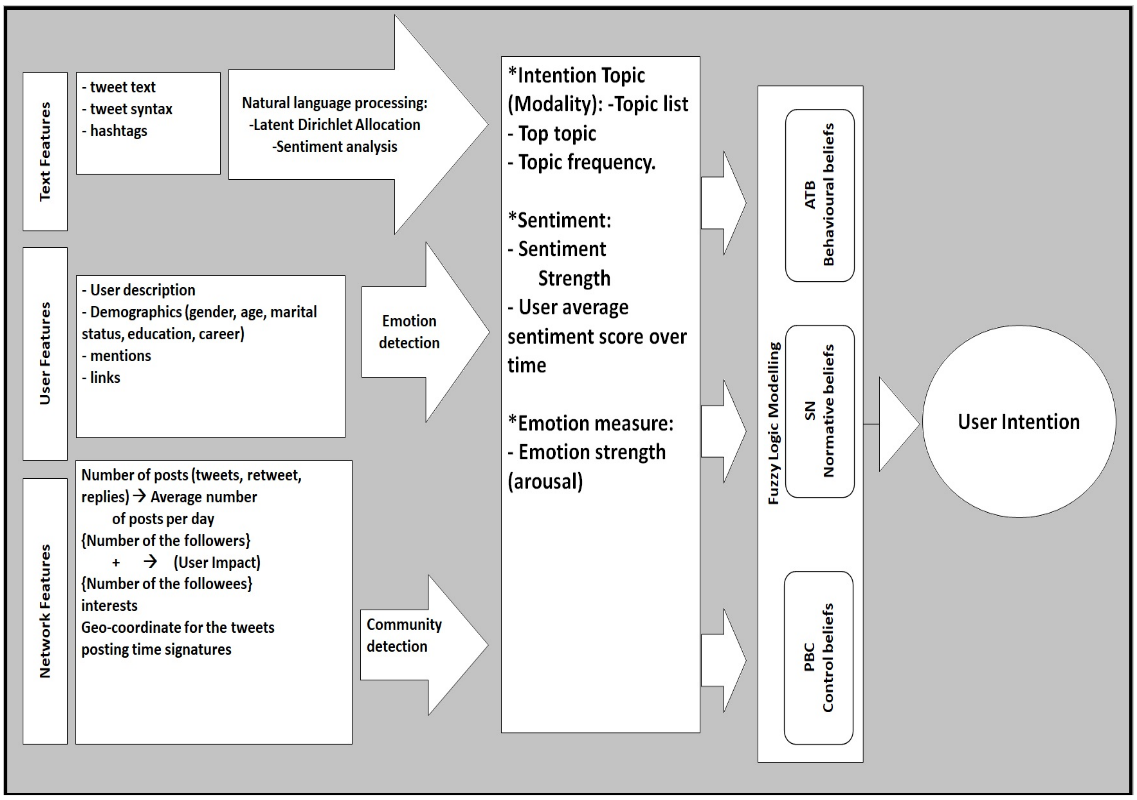


Fig. 6.1 Intention Mining Framework based on TPB Using Sentiment Analysis and Tweet Features

6.3 FL Model for TPB Intention

For the proposed system, the Mamdani fuzzy inference system was used. For each input attribute, the Trapezoidal fuzzy membership function is adopted to convert numerical values into fuzzy linguistic terms due to its popular use in practice [34]. The following section explains the proposed modelling approach for the intention estimation problem in FL based on the TPB.

The first step in building a fuzzy intention model is to define the linguistic variables and fuzzy sets that will be used to represent the input and output variables. In this case, the input variables are valence and arousal, which are used to extract the ATB, SN and PBC factors of TPB. The output variable is the user's intention towards a certain action.

6.3.1 Define Linguistic Values

To define the linguistic variables, TPB was used to define the intention based on ATB, SN, and PBC and the NRC-VAD lexicon, which provides the values of valence and arousal for each word. The range of valence values is from 0 (most negative) to 1 (most positive), while the range of arousal values is from 0 (least arousing) to 1 (most arousing). These values are then mapped to fuzzy linguistic terms using Trapezoidal membership functions. For this research, Intention (I) was defined based on each factor as follows:

- First factor is the Attitude Toward Behaviour (ATB); based on Tyshchuk et al. [133] research, the ATB factor represents the product of the valence sentiment, arousal and dominance. Therefore, the fuzzy set was introduced in the form of $\{positive, negative, netural\}$.

$$ATB = \{positive, neutral, negitive\} \tag{6.15}$$

  - "$positive$": if the input values for valence and arousal and dominance are positive or strong or high, or if two of the values are high and the last is medium or neutral or low.
  - "$negative$": if the input values for valence and arousal and dominance are low, or if two of the values are low, and the last one is neutral or medium or high or strong.
  - "$Neutral$": if the input values for valence and arousal and dominance are medium or neutral, or if two of the values are medium or neutral, and the last one is high or medium.

- The second factor is the Subjective Norm (SN), which reflects the social pressures and influences that affect the individual's behaviour. SN is defined as the degree to which an individual perceives social expectations and pressures to perform a behaviour as favourable socially or unfavourable socially, represented by the linguistic terms $\{low, medium, high\}$.

$$SN = \{low, medium, high\} \tag{6.16}$$

  - "*high*": if the input values for NB and MC are both high or if one of the values is high and the other is medium.
  - "*low*": if the input values for NB and MC are both low or if one of the values is low and the other is medium.
  - "*medium*": if the input values for NB and MC are both medium or if one of the values is low and the other is high.

- Third, the degree of Perceived Behavioural Control(PBC). It represents the individual's perceived level of control over performing a behaviour based on past experiences. It is represented by the linguistic terms "*high*" (experienced) and "*low*" (inexperienced), and "*medium*".

$$PBC = \{low, medium, high\} \tag{6.17}$$

  - "*high*": if the input values for valence and arousal are both high or if one of the values is high and the other is medium.
  - "*low*": if the input values for valence and arousal are both low or if one of the values is low and the other is medium.
  - "*medium*": if the input values for valence and arousal are both medium or if one of the values is low and the other is high.

Furthermore, the high favourable the attitude, the high the degree of pressure of subjective norm, and the more perceived behavioural control towards an action or behaviour, the high probability of individual intention towards that behaviour.

The proposed fuzzy intention model combines these three factors to represent an individual's intention towards a specific behaviour or action. The intention is represented as a vector of the three factors, as follows:

$$\vec{I} \longrightarrow < ATB, SN, PBC > \tag{6.18}$$

Where:

- I refers to an individual's intention to perform a certain behaviour.
- ATB, SN, and PBC are the three key factors that influence an individual's intention to perform a certain behaviour.

The proposed model involves analysing a corpus R of n short statements r_i, where $R = \{r_i | 1 \leq i \leq n\}$, and assigning them to a set of predefined intention classes, a set C of predefined m intention classes, $C = \{c_j | 1 \leq j \leq m\}$.

$$f = \{R \times C\} \Longrightarrow \{low, neutral, hgih\} \tag{6.19}$$

The prediction function f maps the set of short statements documents R to the set of predefined intention classes C, and produces an estimation output indicating whether an intention exists with high, low or medium. Specifically, for each short statement ri in R, the function f assigns it to an intention class c_j in C if one exists and outputs "high" indicating a high presence of intention, low presence of intention, it outputs "low", or neutral intention. This indicates that an intention $\exists c_j \in C$ in $r_i \in R$.

Now to define the linguistic variables of valence, arousal and dominance, the NRC-VAD list lexicon is used, which provides the values of valence and arousal for each word.

- Valence: The linguistic values for valence can be set to "positive", "negative", or "neutral". These values represent the polarity of the emotions associated with a behaviour or action. $Valence = \{positive, neutral, negative\}$
- Arousal: The linguistic values for arousal can be set to "calm" as "low" arousal, "excited" as "high" arousal or "contented" as "medium" arousal. These values represent the degree of emotional intensity associated with a behaviour or action. $Arousal = \{low, medium, high\}$
- Dominance: The linguistic values for dominance can be set to "weak/submissive", "strong/dominant", or "medium". These values represent the extent to which an individual feels in control or restricted in their actions and behaviours. $Dominance = \{weak, medium, strong\}$

Defining these linguistic values can map the sentiment concepts to fuzzy sets and use them as inputs in the proposed FL model for intention prediction.

Finally, the linguistic variables of NB and MC are needed to be defined that are used for representing SN.

- NB: The linguistic values for NB can be set to "low", "medium", or "high". These values refer to the perceived favourable behaviour of important others or the perceived actual behaviour of others. $NB = \{low, medium, high\}$.

- MC: The linguistic values for MC can be set to "low", "medium", or "high". These values reflect an individual's motivation to comply with the perceived expectations of others [6, 91]. $MC = \{low, medium, high\}$.

6.3.2 FL sets based on Sentiment Analysis

Definition: A fuzzy set A in the universe of discourse X is characterized by membership function $\mu_A : X \longrightarrow [0,1]$. A fuzzy set A is represented by the following order pair: $A = \{(x, \mu_A(x)) : \forall x \in X\}$ [69, 57].
The fuzzy operator used to represent the FL rules are (and, or). The (and) operator returns the minimum between values; otherwise will represent an intersection or a conjunction $(\cap)$ between two membership functions. The (or) operator returns the maximum value between values; otherwise will represent the union or disjunction $(\cup)$ between two membership functions.
The Trapezoidal function is a mathematical function that is commonly used in FL, particularly in the Mamdani model. It is a four-sided figure where two sides are parallel, and the other two sides are non-parallel. The Trapezoidal function is defined by four parameters: a, b, c, and d.

The following mathematical model represents the Trapezoidal function:

$$\mu(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a < x \leq b \\ 1 & \text{if } b < x \leq c \\ \frac{d-x}{d-c} & \text{if } c < x < d \\ 0 & \text{if } x \geq d \end{cases} \tag{6.20}$$

Here, a is the left shoulder, b is the left base, c is the right base, and d is the right shoulder. The Trapezoidal function is a piecewise function where the value of the function is 0 before the left shoulder and after the right shoulder. The value of the function gradually increases from 0 to 1 as x moves from the left shoulder to the left base. It remains at 1 as x moves from the left base to the right base and then gradually decreases from 1 to 0 as x moves from the right base to the right shoulder.

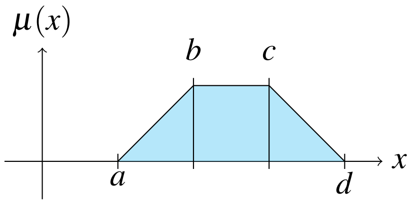


Fig. 6.2 Trapezoidal Membership Function

In the Mamdani model, Trapezoidal functions are used to represent the membership functions of linguistic variables. They are preferred over other types of functions, such as triangular or Gaussian functions, because they offer more flexibility in modelling complex linguistic variables. Trapezoidal functions allow for more gradual transitions between membership grades, which is useful when dealing with linguistic variables that have fuzzy boundaries or overlap with other linguistic variables. Additionally, the Trapezoidal function allows for a wider range of possible shapes and can be adjusted to fit the data more accurately.

To make the membership functions more general, a general representation of the equations with parameters is used. This will allow these functions to be used with any lexicon dictionary as long as the appropriate parameters for that specific dictionary is determined.

The parameters for valence are defined as follows:

A_{neg}: Left shoulder of the Negative Valence fuzzy set

B_{neg}: Base start of the Negative Valence fuzzy set

C_{neg}: Base end of the Negative Valence fuzzy set

D_{neg}: Right shoulder of the Negative Valence fuzzy set

A_{neu}: Left shoulder of the Neutral Valence fuzzy set

B_{neu}: Base start of the Neutral Valence fuzzy set

C_{neu}: Base end of the Neutral Valence fuzzy set

D_{neu}: Right shoulder of the Neutral Valence fuzzy set

A_{pos}: Left shoulder of the Positive Valence fuzzy set

B_{pos}: Base start of the Positive Valence fuzzy set

C_{pos}: Base end of the Positive Valence fuzzy set

D_{pos}: Right shoulder of the Positive Valence fuzzy set

The generalized membership functions will be as follows:

$$\mu(x)neg = \begin{cases} 0 & \text{if } x \leq A_{neg} \\ \frac{x-A_{neg}}{B_{neg}-A_{neg}} & \text{if } A_{neg} \leq x \leq B_{neg} \\ 1 & \text{if } B_{neg} \leq x \leq C_{neg} \\ \frac{D_{neg}-x}{D_{neg}-C_{neg}} & \text{if } C_{neg} \leq x \leq D_{neg} \\ 0 & \text{if } x \geq D_{neg} \end{cases} \tag{6.21}$$

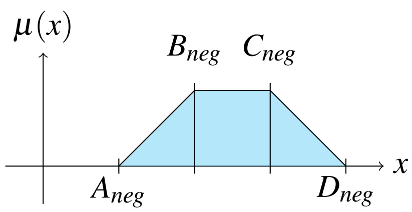


Fig. 6.3 Trapezoidal Membership Function for Negative Valence

$$\mu(x)neu = \begin{cases} 0 & \text{if } x \leq A_{neu} \\ \frac{x-A_{neu}}{B_{neu}-A_{neu}} & \text{if } A_{neu} \leq x \leq B_{neu} \\ 1 & \text{if } B_{neu} \leq x \leq C_{neu} \\ \frac{D_{neu}-x}{D_{neu}-C_{neu}} & \text{if } C_{neu} \leq x \leq D_{neu} \\ 0 & \text{if } x \geq D_{neu} \end{cases} \tag{6.22}$$

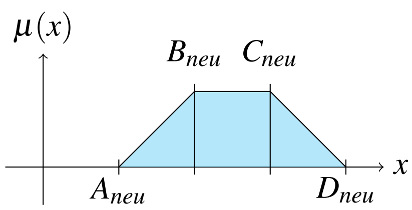


Fig. 6.4 Trapezoidal Membership Function for Neutral Valence

$$\mu(x)pos = \begin{cases} 0 & \text{if } x \leq A_{pos} \\ \frac{x-A_{pos}}{B_{pos}-A_{pos}} & \text{if } A_{pos} \leq x \leq B_{pos} \\ 1 & \text{if } B_{pos} \leq x \leq C_{pos} \\ \frac{D_{pos}-x}{D_{pos}-C_{pos}} & \text{if } C_{pos} \leq x \leq D_{pos} \\ 0 & \text{if } x \geq D_{pos} \end{cases} \tag{6.23}$$

Now, these generalized membership functions can be used with any lexicon

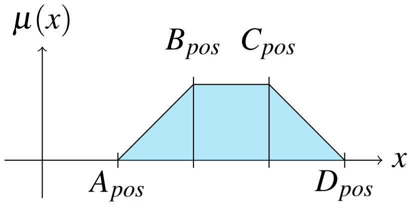


Fig. 6.5 Trapezoidal Membership Function for Positive Valence

dictionary by adjusting the parameters A, B, C, and D for each fuzzy set (Negative, Neutral, and Positive) according to the dictionary's valence range and desired membership ranges.

For example, if a new lexicon dictionary with a valence range of $[0,1]$ was used, the parameters could be adjusted accordingly to create membership functions for that range:

- Determine the new left and right shoulders for each fuzzy set based on the new valence range and desired membership overlap.

- Determine the new base start and end for each fuzzy set to ensure smooth transitions between fuzzy sets.

Figure 6.6, displays an example of trapezoidal membership function for neutral valence, with left and right shoulders at $A_{neu} = 0.32$ and $D_{neu} = 0.73$, respectively, and a base between $B_{neu} = 0.48$ and $C_{neu} = 0.63$.

Once these parameters had been set, the generalized membership functions would be ready for use with the new lexicon dictionary. Furthermore, the membership functions for arousal and dominance will follow the same definition based on the used lexicon dictionary. To ensure that fuzzy sets have been created for both arousal (including Low, Medium, and High) and dominance (consisting of Weak, Medium, and Strong).

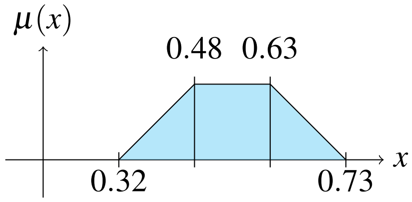


Fig. 6.6 Trapezoidal Membership Function Example for Neutral Valence

Having a set of Valence (V) where $V = \{(x, \mu_V(x)) : x \in X\}$ based on the range of the valence $[0, 1]$ in NRC-VAD list, giving the following membership functions:

- *Negative* Valence fuzzy set membership function as (μ_{neg}) Trapezoidal shape with left and right shoulders at 0 and 0.42, respectively, and a base between 0 and 0.33:

$$\mu(x)_{neg} = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } 0 \leq x \leq 0.33 \\ \frac{0.42-x}{0.42-0.33} & \text{if } 0.33 \leq x \leq 0.42 \\ 0 & \text{if } x \geq 0.42 \end{cases} \tag{6.24}$$

- *Neutral* Valence fuzzy set membership function as (μ_{neu}) Trapezoidal membership function with left and right shoulders at 0.32 and 0.73, respectively, and a base between 0.48 and 0.63:

$$\mu(x)_{neu} = \begin{cases} 0 & \text{if } x \leq 0.32 \\ \frac{x-0.32}{0.48-0.32} & \text{if } 0.32 \leq x \leq 0.48 \\ 1 & \text{if } 0.48 \leq x \leq 0.63 \\ \frac{0.73-x}{6.42-0.63} & \text{if } 0.63 \leq x \leq 0.73 \\ 0 & \text{if } x \geq 0.73 \end{cases} \tag{6.25}$$

- *Positive* Valence fuzzy set membership function as (μ_{pos}) Trapezoidal membership function with left and right shoulders at 0.65 and 1, respectively, and a base between 0.73 and 1:

$$\mu(x)_{pos} = \begin{cases} 0 & \text{if } x \leq 0.65 \\ \frac{x-0.65}{0.73-0.65} & \text{if } 0.65 \leq x \leq 0.73 \\ 1 & \text{if } 0.73 \leq x \leq 1 \\ 0 & \text{if } x \geq 1 \end{cases} \tag{6.26}$$

The fuzzy set of the valence (x) is described as follows:

$$\begin{aligned}(x,\mu) = \{&\{\mu(x_0)_{pos}/(x_0)_{pos}, \mu(x_0)_{neu}/(x_0)_{neu}, \mu(x_0)_{neg}/(x_0)_{neg}\} \\ &\{\mu(x_1)_{pos}/(x_1)_{pos}, \mu(x_1)_{neu}/(x_1)_{neu}, \mu(x_1)_{neg}/(x_1)_{neg}\}, \ldots\}\end{aligned} \tag{6.27}$$

Having a set of Arousal (A) where $A = \{(y, \mu_A(y)) : \forall y \in Y\}$ based on the range of the values in the NRC-VAD list $[0,1]$. The arousal fuzzy set (A) membership functions are given as follows:

- Low: Trapezoidal membership function with left and right shoulders at 0 and 0.37, respectively, and a base between 0 and 0.25.

$$\mu(y)_{low} = \begin{cases} 0 & \text{if} y \leq 0 \\ 1 & \text{if } 0 \leq y \leq 0.25 \\ \frac{0.37-y}{0.37-0.25} & \text{if } 0.25 \leq y \leq 0.37 \\ 0 & \text{if } y \geq 0.37 \end{cases} \tag{6.28}$$

- Medium: Trapezoidal membership function with left and right shoulders at 0.26 and 0.70, respectively, and a base between 0.41 and 0.57.

$$\mu(y)_{medium} = \begin{cases} 0 & \text{if } y \leq 0.26 \\ \frac{y-0.26}{0.41-0.26} & \text{if } 0.26 \leq y \leq 0.41 \\ 1 & \text{if } 0.41 \leq y \leq 0.57 \\ \frac{0.7-y}{0.7-0.57} & \text{if } 0.57 \leq y \leq 0.7 \\ 0 & \text{if } y \geq 0.7 \end{cases} \tag{6.29}$$

- High: Trapezoidal membership function with left and right shoulders at 0.57 and 1, respectively, and a base between 0.73 and 1.

$$\mu(y)_{high} = \begin{cases} 0 & \text{if } y \leq 0.57 \\ \frac{y-0.57}{0.73-0.57} & \text{if } 0.57 \leq y \leq 0.73 \\ 1 & \text{if } 0.73 \leq y \leq 1 \\ 0 & \text{if } y \geq 1 \end{cases} \tag{6.30}$$

The fuzzy set of Arousal (y) is described as follows:

$$\begin{aligned} (y,\mu) = \{\{\mu(y_0)_{low}/(y_0)_{low}, \mu(y_0)_{medium}/(y_0)_{medium}, \\ \mu(y_0)_{high}/(y_0)_{high}\},, \{\{\mu(y_1)_{low}/(y_1)_{low}, \\ \mu(y_1)_{medium}/(y_1)_{medium}, \\ \mu(y_1)_{high}/(y_1)_{high}\}, \ldots\} \end{aligned} \tag{6.31}$$

Having a set of Dominance (D) where $D = \{(w, \mu_D(w)) : \forall w \in W\}$ based on the range of the dominance values in the NRC-VAD list between $[0,1]$, giving the following membership functions for (D):

- *Weak*: Trapezoidal membership function with left and right shoulders at 0 and 0.35, respectively, and the base is between 0 and 0.041.

$$\mu(w)_{weak} = \begin{cases} 0 & \text{if } w \leq 0 \\ 1 & \text{if } 0 \leq w \leq 0.041 \\ \frac{0.35-w}{0.35-0.041} & \text{if } 0.041 \leq w \leq 0.35 \\ 0 & \text{if } w \geq 0.35 \end{cases} \tag{6.32}$$

- *Medium*: Trapezoidal membership function with left and right shoulders at

0.25 and 0.75, respectively, and the base is between 0.46 and 0.54.

$$\mu(w)_{medium} = \begin{cases} 0 & \text{if } w \leq 0.25 \\ \frac{w-0.25}{0.46-0.25} & \text{if } 0.25 \leq w \leq 0.46 \\ 1 & \text{if } 0.46 \leq w \leq 0.54 \\ \frac{0.75-w}{0.75-0.54} & \text{if } 0.54 \leq w \leq 0.75 \\ 0 & \text{if } w \geq 0.75 \end{cases} \tag{6.33}$$

- *Strong*: Trapezoidal membership function with left and right shoulders at 0.66 and 1, respectively, and the base is between 0.95 and 1.

$$\mu(w)_{strong} = \begin{cases} 0 & \text{if } w \leq 0.66 \\ \frac{w-0.66}{0.95-0.66} & \text{if } 0.66 \leq w \leq 0.95 \\ 1 & \text{if } 0.95 \leq w \leq 1 \\ 0 & \text{if } w \geq 1 \end{cases} \tag{6.34}$$

The fuzzy set of Dominance (w) is described as follows:

$$\begin{aligned} (w, \mu) = \{\{\mu(w_0)_{Strong}/(w_0)_{Strong}, \\ \mu(w_0)_{Medium}/(w_0)_{Medium}, \\ \mu(w_0)_{Weak}/(w_0)_{Weak}\}, \\ \{\mu(w_1)_{Strong}/(w_1)_{Strong}, \\ \mu(w_1)_{Medium}/(w_1)_{Medium}, \\ \mu(w_1)_{Weak}/(w_1)_{Weak}\}, \ldots\} \end{aligned} \tag{6.35}$$

It is worth noting that the ranges were selected based on the assumption that the membership functions for the different fuzzy sets should be roughly evenly spaced, with some overlap between adjacent sets to ensure smooth transitions. The exact values of the ranges were then adjusted through experimentation to achieve the desired degree of overlap and smoothness of the transition.

6.3.3 Fuzzy Rules

Fuzzy rule learning approach was proposed due to the advantages of FL. Firstly, previous research that used TPB to model user intention focused on human feedback through surveys to measure each factor of the theory [55]. However, in social networks, that option would be difficult since users are not always responsive. Secondly, FL is well capable of dealing with linguistic uncertainty. In particular, it considers a classification problem to be a " degree of grey" rather than a "black and white" problem [77](an implicit intention mining problem).

ATB Fuzzy Rules

Representing ATB depends on three variables of valence, arousal, and dominance. The linguistic values for ATB are defined as follows:
$ATB = \{Negative, Neutral, Positive\}$

The fuzzy rule set that is used to build the ATB is as follows:

Rule 1: If (x is Negative) and (y is Low) and (w is Weak) then (ATB is Negative)

Rule 2: If (x is Negative) and (y is Low) and (w is Medium) then (ATB is Negative)

Rule 3: If (x is Negative) and (y is Low) and (w is Strong) then (ATB is Neutral)

Rule 4: If (x is Negative) and (y is Medium) and (w is Weak) then (ATB is Negative)

Rule 5: If (x is Negative) and (y is Medium) and (w is Medium) then (ATB is Negative)

Rule 6: If (x is Negative) and (y is Medium) and (w is Strong) then (ATB is Neutral)

Rule 7: If (x is Negative) and (y is High) and (w is Weak) then (ATB is Negative)

Rule 8: If (x is Negative) and (y is High) and (w is Medium) then (ATB is Negative)

Rule 9: If (x is Negative) and (y is High) and (w is Strong) then (ATB is Neutral)

Rule 10: If (x is Neutral) and (y is Low) and (w is Weak) then (ATB is Neutral)

Rule 11: If (x is Neutral) and (y is Low) and (w is Medium) then (ATB is Neutral)

Rule 12: If (x is Neutral) and (y is Low) and (w is Strong) then (ATB is Neutral)

Rule 13: If (x is Neutral) and (y is Medium) and (w is Weak) then (ATB is Neutral)

Rule 14: If (x is Neutral) and (y is Medium) and (w is Medium) then (ATB is Neutral)

Rule 15: If (x is Neutral) and (y is Medium) and (w is Strong) then (ATB is Neutral)

Rule 16: If (x is Neutral) and (y is High) and (w is Weak) then (ATB is Neutral)

Rule 17: If (x is Neutral) and (y is High) and (w is Medium) then (ATB is Neutral)

Rule 18: If (x is Neutral) and (y is High) and (w is Strong) then (ATB is Neutral)

Rule 19: If (x is Positive) and (y is Low) and (w is Weak) then (ATB is Neutral)

Rule 20: If (x is Positive) and (y is Low) and (w is Medium) then (ATB is Neutral)

Rule 21: If (x is Positive) and (y is Low) and (w is Strong) then (ATB is Positive)

Rule 22: If (x is Positive) and (y is Medium) and (w is Weak) then (ATB is Neutral)

Rule 23: If (x is Positive) and (y is Medium) and (w is Medium) then (ATB is Neutral)

Rule 24: If (x is Positive) and (y is Medium) and (w is Strong) then (ATB is Positive)

Rule 25: If (x is Positive) and (y is Low) and (w is Strong) then (ATB is Positive)

Rule 26: If (x is Positive) and (y is Medium) and (w is Strong) then (ATB is Positive)

Rule 27: If (x is Positive) and (y is High) and (w is Strong) then (ATB is Positive)

SN Fuzzy Rules

Now, in order to represent SN using the user's network factors, one should consider how other people within the community believe the individual should behave, also, how most people usually behave. The former is referred to as injunctive SN, and the latter is the descriptive SN. In other words, to represent injunctive (SN), the attitude of other people toward a specific behaviour needs to be studied. The calculation of SN is based on NB and MC as presented in section 6.2.2. The fuzzy rules for SN are presented as follows:

Rule 1: If NB is low or MC is low, then SN is low

Rule 2: If NB is low or MC is medium, then SN is medium

Rule 3: If *NB* is low or *MC* is high, then *SN* is high

Rule 4: If *NB* is medium or *MC* is low, then *SN* is medium

Rule 5: If *NB* is medium or *MC* is medium, then *SN* is medium

Rule 6: If *NB* is medium or *MC* is high, then *SN* is high

Rule 7: If *NB* is high or *MC* is low, then *SN* is high

Rule 8: If *NB* is high or *MC* is medium, then *SN* is high

Rule 9: If *NB* is high or *MC* is high, then *SN* is high

PBC Fuzzy Rules

In TPB, PBC refers to a person's perception of the ease or difficulty of performing the behaviour of interest[6]. Perceived behavioural control varies across situations and actions, which results in a person having varying perceptions of behavioural control depending on the situation[6, 15]. With the emotion factors, representing the PBC depends on the person's both valence (x) and arousal (y) toward the behaviour. The following are fuzzy rules based on the relationship between valence and arousal and their impact on PBC:

Rule 1: If x is Negative and y is Low, then *PBC* is Low.

Rule 2: If x is Neutral and y is Low, then *PBC* is Medium.

Rule 3: If x is Positive and y is Low, then *PBC* is Medium.

Rule 4: If x is Negative and y is Medium, then *PBC* is Low.

Rule 5: If x is Neutral and y is Medium, then *PBC* is Medium.

Rule 6: If x is Positive and y is Medium, then *PBC* is High.

Rule 7: If x is Negative and y is High, then *PBC* is Medium.

Rule 8: If x is Neutral and y is High, then *PBC* is High.

Rule 9: If x is Positive and y is High, then *PBC* is High.

Intention Rules

Next, present the fuzzy rules that capture the relationships between ATB, SN, and PBC in predicting intention. These rules are formulated as "IF-THEN" statements, specifying the conditions for each input variable and the corresponding output variable (intention). The rules are as follows:

Rule 1: If ATB is Negative and SN is Low and PBC is Low, then Intention is Low.

Rule 2: If ATB is Negative and SN is Low and PBC is Medium, then Intention is Low.

Rule 3: If ATB is Negative and SN is Low and PBC is High, then Intention is Low.

Rule 4: If ATB is Negative and SN is Medium and PBC is Low, then Intention is Low.

Rule 5: If ATB is Negative and SN is Medium and PBC is Medium, then Intention is Low.

Rule 6: If ATB is Negative and SN is Medium and PBC is High, then Intention is Medium.

Rule 7: If ATB is Negative and SN is High and PBC is Low, then Intention is Low.

Rule 8: If ATB is Negative and SN is High and PBC is Medium, then Intention is Medium.

Rule 9: If ATB is Negative and SN is High and PBC is High, then Intention is Medium.

Rule 10: If ATB is Neutral and SN is Low and PBC is Low, then Intention is Low.

Rule 11: If ATB is Neutral and SN is Low and PBC is Medium, then Intention is Low.

Rule 12: If ATB is Neutral and SN is Low and PBC is High, then Intention is Low.

Rule 13: If ATB is Neutral and SN is Medium and PBC is Low, then Intention is Low.

Rule 14: If ATB is Neutral and SN is Medium and PBC is Medium, then Intention is Medium.

Rule 15: If ATB is Neutral and SN is Medium and PBC is High, then Intention is Medium.

Rule 16: If ATB is Neutral and SN is High and PBC is Low, then Intention is Medium.

Rule 17: If ATB is Neutral and SN is High and PBC is Medium, then Intention is High.

Rule 18: If ATB is Neutral and SN is High and PBC is High, then Intention is High.

Rule 19: If ATB is Positive and SN is Low and PBC is Low, then Intention is Low.

Rule 20: If ATB is Positive and SN is Low and PBC is Medium, then Intention is Low.

Rule 21: If ATB is Positive and SN is Low and PBC is High, then Intention is Medium.

Rule 22: If ATB is Positive and SN is Medium and PBC is Low, then Intention is Medium.

Rule 23: If ATB is Positive and SN is Medium and PBC is Medium, then Intention is High.

Rule 24: If ATB is Positive and SN is Medium and PBC is High, then Intention is High.

Rule 25: If ATB is Positive and SN is High and PBC is Low, then Intention is Medium.

Rule 26: If ATB is Positive and SN is High and PBC is Medium, then Intention is High.

Rule 27: If ATB is Positive and SN is High and PBC is High, then Intention is High.

6.3.4 Fuzzy Inference System

This research used the Mamdani inference method to calculate the fuzzy output sets. The fuzzy rules were applied to the input values to determine the degree of membership for each fuzzy set. Using the Mamdani method, the degree of membership for each fuzzy output set was then calculated based on the degree of membership of the input fuzzy sets. This process allowed for the determination of the most appropriate output values based on the input values and the defined rules.

The fuzzy inference system (FIS) is a critical component in modelling the intention estimation process based on the (TPB). In this chapter, the linguistic values had been defined, FL sets had been created, and fuzzy rules had been established. This section will delve into the details of constructing the fuzzy inference system and the steps involved in its operation.

The FIS employed in this research is based on the Mamdani inference method, widely used for its intuitive rule-based approach and ability to handle imprecise input data. The Mamdani inference system consists of four principal stages: fuzzification, rule evaluation, rule aggregation, and defuzzification.

- **Fuzzification:** In the fuzzification step, the primary goal is to transform the crisp input values into fuzzy sets, representing the degree of membership of each input variable within the predefined linguistic terms. This transformation is essential as it enables the fuzzy inference system to capture the inherent uncertainty and imprecision associated with human intentions and decision-making processes.

  The fuzzification process begins with applying membership functions to the crisp input values for valence, arousal, and dominance. These membership functions that were previously defined in the context of the research are Trapezoidal. Each input value is evaluated against the membership functions of its respective linguistic terms, resulting in a degree of membership for each term. This step essentially translates the crisp input values into a fuzzy representation, where each input variable is associated with one or more linguistic terms to varying degrees.

  - **Example**: Suppose having a Twitter post related to a new healthy life style resolution for the new year. The aim si to predict the intention of the user to engage with this resolution using ATB, SN, and PBC. Consider the input values as ATB = 0.8, SN = 0.6 and PBC = 0.7. In the fuzzification process, these values would be assessed against the membership functions for the defined linguistic terms for each variable as ("positive," "neutral," and "negative") in ATB. The result would be a degree of membership for each term, indicating how strongly the input values are associated with each linguistic term. For instance:

    * ATB: Positive (Degree of Membership: 0.8)
    * SN: Medium (Degree of Membership: 0.6)
    * PBC: High (Degree of Membership: 0.7)

The fuzzification step is crucial to the fuzzy inference process, as it sets the stage for the subsequent evaluation of fuzzy rules and aggregating their results. By representing the input variables as fuzzy sets, the fuzzification process allows the fuzzy inference system to model better the uncertainty and vagueness that characterize human intentions, thereby enhancing the system's ability to generate meaningful and accurate intention estimations.

- **Rule Evaluation:** The rule evaluation step, a crucial component of the fuzzy inference process, focuses on assessing the applicability of the predefined fuzzy rules to the fuzzified input values. By doing so, the system can determine the relevance of each rule in capturing the relationships between the input variables, such as (ATB, SN, and PBC) and the output variable (Intention) within (TPB) framework.

  In this step, the previously fuzzified input values are used in conjunction with the established fuzzy rules to calculate the degree of membership for each output fuzzy set. The fuzzy rules expressed as a series of if-then statements describe the associations between the input and output linguistic terms. The fuzzy inference system evaluates each rule by applying the appropriate fuzzy operators, such as the minimum (min) operator for the AND conjunction and the maximum (max) operator for the OR disjunction, to the input values' degrees of membership.

  - For instance, revisiting the example presented in the fuzzification step. One of the fuzzy rules states: "*IF ATB is Positive AND SN is Medium AND PBC is High THEN Intention is High*". The fuzzy inference system assesses this rule by computing the minimum degree of membership among the three input linguistic terms (positive ATB, medium SN, and high PBC). According to this rule, the resulting value represents the degree of membership for the output linguistic term (High Intention).

  The rule evaluation process is performed for all fuzzy rules, resulting in a set of output fuzzy sets with corresponding degrees of membership. This step is essential in identifying the most relevant rules for the given input values and understanding how the input variables contribute to the overall intention estimation. By evaluating the fuzzy rules, the system can leverage the expert knowledge embedded in the rules to generate nuanced and context-specific insights into the relationships between the input and output variables, ultimately enhancing the accuracy and interpretability of the intention estimations.

- **Rule Aggregation:** The fuzzy output sets resulting from all the rules are combined during rule aggregation. This step generates an aggregated fuzzy output set for the output variable (Intention). The commonly used aggregation methods include Max, Min, and Sum, depending on the fuzzy rules' nature and the desired granularity level.
  - Continuing with the proposed example, using the rule "IF ATB is Positive AND SN is Medium AND PBC is High THEN Intention is High.", the intermediate result for the aggregated Intention is [0.7 0.9 1.0] (High).
- **Defuzzification:** After applying the fuzzy rules to these input sets, fuzzy output sets for each factor were obtained. To obtain a crisp output value for each factor, the fuzzy output sets were needed to be defuzzified.

  The centroid method can be used for each output fuzzy set to calculate the crisp output value. The centroid of a fuzzy set is the centre of gravity of the set, which can be calculated as the weighted average of the values in the set. In this work case, the values in the output fuzzy sets represent the degree of membership of the output in a particular set.

  The first step involves discretizing the set into a finite number of points to compute the centroid for each output fuzzy set. Subsequently, the weighted average of these points can be calculated, with the weights determined by the degree of membership of each point within the set. This weighted average yields the crisp output value for each factor.

$$Output = \frac{\sum_{i=1}^{n}(x_i \times \mu_A(x_i))}{\sum_{i=1}^{n}\mu_A(x_i)} \tag{6.36}$$

  Where:

  - x_i represents the variable output value.

- $\mu_A(x_i)$ denotes the degree of membership for the fuzzy output set A at the value x_i.
- n is the total number of variable output values.

By calculating the weighted average, the centroid method ensures that the crisp output value accurately reflects the contributions of each fuzzy output set, thereby preserving the context-sensitive insight derived from fuzzy rules. Once the crisp output values for ATB, SN, and PBC have been obtained, they can be combined obtain the overall intention estimate for the user. This combination can be done using a weighted sum approach, where the weights are determined based on the relative importance of each factor in determining the user's intention.

- In the example that used above the aggregate the intermediate fuzzy outputs using a weighted average method:

$$Defuzzification = \frac{(0.7 \times 0.8 + 0.9 \times 0.6 + 1.0 \times 0.7)}{(0.7 + 0.9 + 1.0)} = 0.82$$

  This value (0.82) indicates the degree of membership of the user's intention in the *"High"* linguistic term. In fuzzy logic, this value represents the extent to which the user's intention aligns with the category of *"High Intention."*. This suggests a relatively strong inclination or likelihood of the user to engage with activity in the given Twitter post. In the context of the example, a *"High"* intention implies that the user is highly inclined to participate, perhaps by liking, sharing, or commenting on the post related to the new healthy lifestyle resolution. The final intention Estimate of the value (0.82), serves as the definitive intention estimate for the user. It is a crisp, quantifiable value that summarizes the fuzzy inference process's outcome, providing a clear indication of the user's intention. This information can be valuable for decision-making and tailoring content strategies to target users with a strong intention to participate in the specified online activity.

6.4 Summary

In conclusion, this chapter presents a proposed FL model for intent in social networks that integrates sentiment analysis and the TPB model. The proposed model offers a comprehensive approach to understanding the complex relationship between user attitudes and intentions in social media, and provides a powerful tool for predicting and influencing user behaviour. Through the incorporation of sentiment analysis with TPB factors such as ATB, SN, and PBC, the proposed model allows for more accurate and reliable predictions of user intentions, particularly when dealing with imprecise and incomplete information. The chapter provides a detailed discussion of the methods used to construct and evaluate the fuzzy intention model, highlighting its potential applications in various domains, including businesses and organizations seeking to better understand their customers' attitudes and intentions towards their products or services.

Chapter 7

Intention Mining Experiments

7.1 Data Mining Techniques: Experimental Methodology

This study aims to explore intent mining using data mining techniques. Specifically, the KNIME tool was used to preprocess, analyse, and visualize a dataset of tweets related to users' intention. This work goal is to identify the tweets with intentions, and analyse their frequency and distribution.

The KNIME tool provides a user-friendly environment for data mining and machine learning, allowing for efficient data preparation and analysis. A range of data mining techniques will be utilized, including clustering and association rule mining, to identify patterns and relationships within the dataset.

The results of this experiment will provide insights into the intentions expressed by Twitter users on the selected topic, and can be used to inform business and marketing strategies. It will also serve as a foundation for further research into intent mining using data mining analysis techniques.

7.1.1 Experiment Setup

The experimental setup is a critical aspect of any research project as it lays the foundation for the subsequent analyses and results. This study aims to investigate the use of data mining techniques for intent mining on social media data. To achieve this, KNIME 4.7.1 64-bit platform was utilized, which is a powerful and widely used data mining tool.

The machine used for the experiments was a Windows 10 64-bit operating system with an Intel i7 2.2GHz processor and 12GB RAM. This system configuration provided sufficient processing power and memory to enable processing and analysing large datasets efficiently. The KNIME platform was selected as it provides a user-friendly interface and a comprehensive set of data mining and analytics tools that are essential for processing social media data.

The data for this study were collected from Twitter using the Twitter API connection node combined with Twitter search nodes. The Twitter API connection node enabled establishing a connection with the Twitter data stream and retrieve data in real-time. The Twitter search nodes were used to filter the data and retrieve only the tweets that matched specific constraints, such as keywords, hashtags, or specific user handles.

The collected tweets were then preprocessed to remove noise and irrelevant data. This was achieved through a series of data cleaning and preparation techniques, including text normalization, stop word removal, and stemming. The cleaned data were then transformed into a structured format that could be used for further analysis.

7.1.2 Pre-processing Data

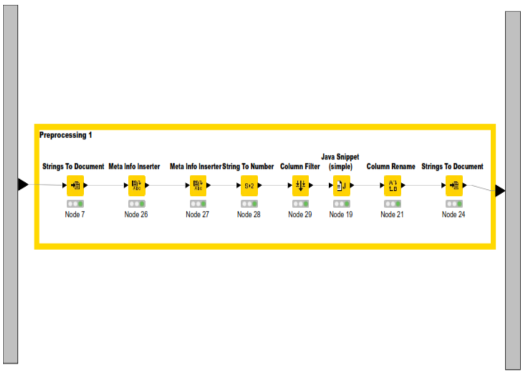
The collected dataset from Twitter contains a significant amount of noise due to the informal language and spelling errors present in the text. Therefore, several pre-processing steps must be taken to ensure the accuracy and validity of the data before applying the next experiments steps.

To achieve this, a series of pre-processing techniques had been employed using KNIME text mining nodes, presented in the form of Meta-nodes (a collection of nodes) as depicted in figure 7.1a. The first pre-processing step (Pre-processing-1) involves the conversion of all tweet texts into documents data type, as this is a requirement for KNIME text mining tools. Furthermore, all URLs present in the documents have been removed to avoid any potential conflicts when building word vectors in the subsequent steps.

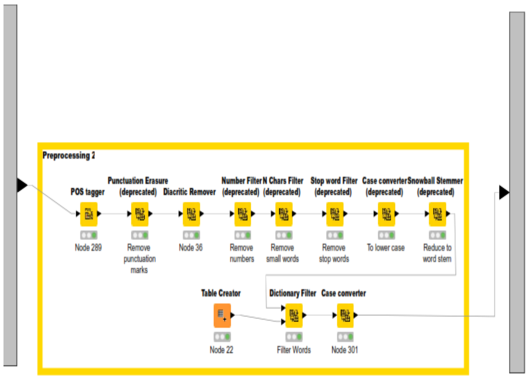
The second pre-processing step (Pre-processing-2) employs a range of different text filtering techniques, also using KNIME nodes as shown in figure 7.1b. These techniques include Part of speech (POS tagger), Punctuation Erasure, Diacritic Remover, Number Filter, N Chars Filter, Stopword Filter, Case Converter

Filter, Snowball Stemmer, Dictionary Filter, and Case Converter. The output of this step is a filtered document text, where each document record has been prepared to undergo the next step, as illustrated in figure 7.1c.

It is important to note that the pre-processing steps used in this experiment are essential for reducing the noise in the dataset and improving the accuracy of the subsequent analysis. By employing these techniques, transforming the raw data into a format that is more suitable for data mining techniques was achieved.



(a) Pre-processing 1: Convert to Documents and Removing URLs



(b) Pre-processing 2: Filtering the Document Text Format

Row ID	Orig Document	Preprocessed Document
Row25	""@lbdn123 The only time you would want to meet him would be to play p...	"lbdn123 time meet play poker easier read spot dog"
Row26	""two more names than I expected this early tbh, but I respect the move...	"name expect tbh respect move pursu tha..."
Row27	""RT @gucci1017: I charge 50 a verse but after #Droptopwop co...	"gucci1017 charg vers droptopwop quarter million"
Row28	""RT @seanhannity: Also I want to humbly thank all of you for the suppo...	"seanhann humbl thank support shown promis stop seek
Row29	""RT @SportQuotesBest: Some people want it to happen, some wish it w...	"sportquotesbest peopl happen wish happen happen -mi
Row30	""Takeoff drop a single and all of a sudden it's like when ya BD see u got ...	"takeoff drop singl sudden famili"
Row31	""@PerryGreco I want one""	"perrygreco"
Row32	""RT @busy_izzywhizzy: No, I don't want to see posts from March thank...	"busyizzywhizzi post march thank"
Row33	""RT @NaMo_Satya: Ask these people's that not to get in any kind of rel...	"namosatya peoplrelationship pakistaniindian muslim mar
Row34	""RT @MetroBoomin: And I charge 40 a beat but after #droptopwop I w...	"metroboomin charg beat droptopwop"
Row35	""RT @jurygroup: Did @realDonaldTrump want @EmmanuelMa...	"jurygroup realdonaldtrump emmanuelmacron win french
Row36	""RT @venusbrownies: I want someone to look at me the same way The ...	"venusbrowni look pope look obama"
Row37	""i want them GONE ""	"gone"
Row38	""RT @Skinny215: People still can't differentiate between a "need" and "...	"skinny215 peopl differenti look goofi skinni"
Row39	""i want kyungsoo to choke me im not kidding i would kill for kyungsoo to c...	"kyungsoo choke kid kill kyungsoo choke"
Row40	""RT @KEVINGETEM: I don't want nobody around me that I always gotta ...	"kevingetem gotta question"
Row41	""@loncbot come to titty typhoon in the next 20 mins if u want an ass kic...	"loncbot titti typhoon min ass kick"
Row42	""@RepAdamSchiff Au contraire - in politics and business, the Trump stra...	"repadamschiff contrair polit busi trump strategi ambigu"
Row43	""I want a lot of keychains. □""	"lot keychain"
Row44	""RT @DavidCamercon: @TRobinsonNewEra and the police r more concer...	"davidcamercon trobinsonnewera polic concern feel musl
Row45	""RT @Cyn_Santana: Keep your mediocre ass love. I want that extra shi...	"cynsantana mediocr ass love extra"

(c) The Output of Pre-processing Step

Fig. 7.1 The Workflows of Dataset Preprocessing Steps Using KNIME

7.1.3 Feature Selection: Phase-One

The aim of this approach is to identify the features in tweets that indicate intent, and this was achieved by utilizing IG Algorithm. Tweet vectors were used to select the feature set for classifying tweets into those with intent and those without.

To create tweet vectors, KNIME document vector node was used to represent each tweet's term space as a Boolean value. The dimensionality of the feature vectors were then reduced in KNIME using a set of nodes, as shown in figure 7.2a. This set of nodes included the Bag of Words Creator (BOW) and the Term Frequency (TF) nodes, which created a table of the terms existing in the document and calculated their frequencies in the documents set, respectively.

IG algorithm was applied to reduce the document vector from 2751 to 82 features. These 82 features were selected based on the dataset, i.e., the text of the tweets. Since the IG algorithm calculates the mutual information of the dataset, the selected features have the highest mutual information. In other words, all 82 terms have an information gain greater than zero.

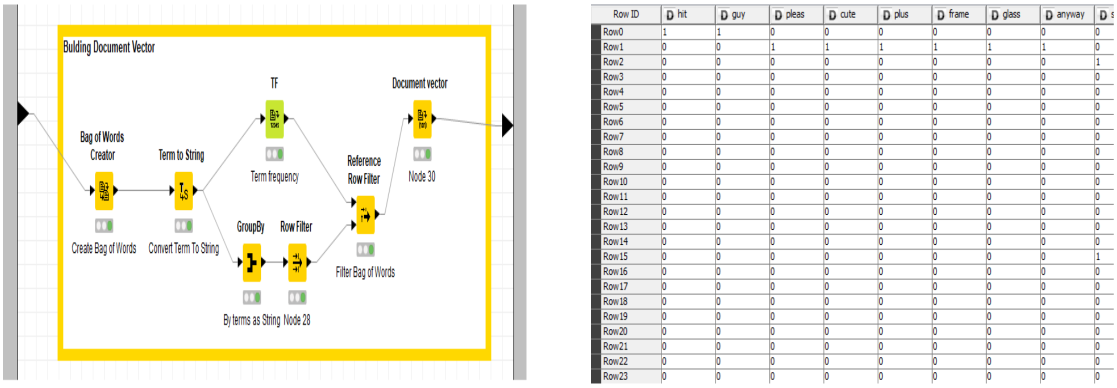
These 82 features were used to train four well-known classification models: DT, ANN, and NB. The dataset was partitioned using Random train-test split technique for prediction in the classification phase. The data is randomly divided into two sets, 70% training set and 30% test set.

It is worth noting that these features were not derived from the context of the posts themselves, but rather from the IG algorithm for documents. To sum up, IG was used to identify the features that users most frequently used in their posts. Then these textual features were employed in the sentiment analysis techniques to extract user intent.

Table 7.1 shows the overall performance of the various feature selection and classifiers using IG algorithm with 82 features, in terms of Recall, Precision, F-measure, and Accuracy. The results demonstrate that the Decision Tree Classifier achieved the highest Recall rate of 88.86%, while Naive Bayes had the highest Precision rate of 83.35%. The F-measure was highest for Decision Tree Classifier, at 86.14%. The accuracy was the highest for the Decision Tree Classifier at 83.32%. Figure 7.3 shows clearly that DT out perform the rest of the algorithms in accuracy when applying IG feature selection.

7.1.4 Feature Selection : Approach-Two (Hybrid)

This section presents the second approach for classifying tweets based on two feature selection algorithms. This approach entails setting the features threshold to eight features in the form of terms vector. The experiment is built to run in two phases, the first of which involves two levels of feature selection based on



(a) Building the Document Vector Using KNIME (b) The Output of Building The Document Vector

Fig. 7.2 Building the Document Vector

Table 7.1 Overall Performances of Various Feature Selection and Classifiers Using IG-Information Gain Algorithm, using 82 Features

Classification	Recall	Precision	F-measure	Accuracy
DT	**88.86%**	**83.58%**	**86.14%**	**83.32%**
ANN	88.36%	82.66%	85.41%	82.40%
SVM	84.30%	82.75%	83.52%	81.70%
NB	85.74%	83.35%	84.53%	80.61%

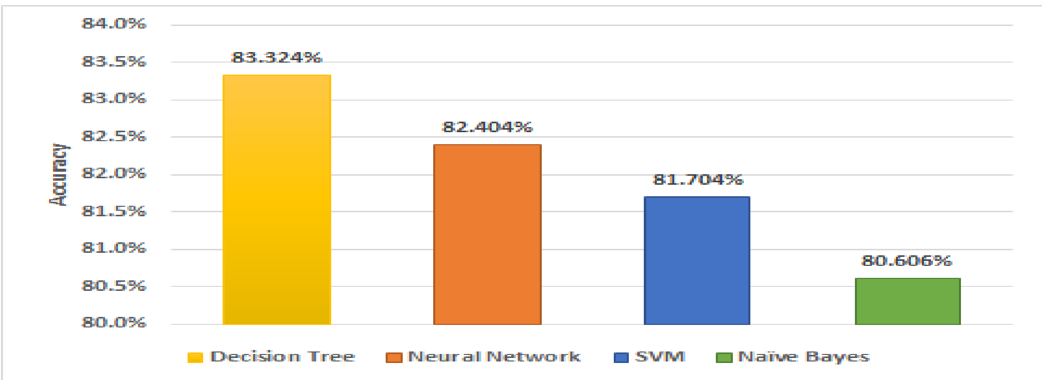


Fig. 7.3 Accuracy of Applying IG Features Selection

two different algorithms.

The first level is imported from the previous experiment by using IG Algorithm extracted 82 features. In the second level, a different feature selection algorithm is used, namely the Forward-Feature Selection Algorithm (FFS). FFS starts with an empty set of features and adds features to the set. The FFS algorithm is applied with four algorithms to select features: NB, SVM, ANN, and IG. The final step involves using classification algorithms, specifically DT, SVM, ANN, and NB, after setting the features. The data is partitioned using 10-fold cross-validation to ensure the robustness of the results.

The hybrid approach offers a unique solution to the feature selection problem by combining two distinct algorithms, which has the potential to increase the accuracy of tweet classification. Additionally, the use of multiple algorithms and two levels of feature selection enables the selection of the most relevant and informative features from the dataset.

The results indicated that while the hybrid feature selection approach led to a marginal reduction in accuracy for all classification algorithms, it concurrently demonstrated the advantageous trait of significantly reducing computational complexity. DT and ANN achieved the highest accuracy among the four algorithms. Furthermore, it is important to note that the effectiveness of this approach is dependent on the choice of algorithms and the threshold for the number of features selected. To ensure optimal performance, further experimentation may be required to determine the ideal combination of algorithms and feature threshold.

Decision Tree (DT)

DT setup, which is based on C4.5 [123], for the experimentation was as follows:

- The quality measure is calculated based on the Gini Index as splitting technique, with no pruning.
- The minimum number of nodes is 2. The split point value is calculated according to the mean value of the two attribute values that separate two partitions. Working on eight cores to speed performance.

The data was partitioned using 10-fold cross-validation to ensure the robustness of the results. Applying two phases to reduce feature gives a relatively close results, even though, the features are reduced to eight. By observing Table 7.2, almost the same accuracy values have been resulted for DT with very slight difference.

Since the decision trees learning method predicts the values of target variable by learning simple decision rules inferred from the data features, it resulted in a relative high outcome. It is robust to noisy data, and since it is a heuristic algorithm, that means a decision is obtained locally and does not guarantee to return the globally optimal solution.

Table 7.2 Experiment results of using Decision Tree classifier with eight features selected through the Hybrid feature selection

Feature Selection	Recall	Precision	F-measure	Accuracy
IG+NB (10-fold CV)	88.56% ± 0.45%	83.61% ± 0.36%	86.02% ± 0.24%	83.21% ± 0.57%
IG+SVM (10-fold CV)	88.56% ± 0.45%	83.61% ± 0.36%	86.02% ± 0.24%	83.21% ± 0.57%
IG+ANN (10-fold CV)	88.76% ± 0.38%	83.64% ± 0.31%	86.12% ± 0.16%	83.32% ± 0.43%
IG+DT (10-fold CV)	88.46% ± 0.52%	82.44% ± 0.55%	85.35% ± 0.40%	82.29% ± 0.79%

Naive Bayes (NB)

The basic NB classifier is used to decide the right class of the input data by referring to the highest probability values that calculated by the trainer classifier using the Bayes formula 5.1.

For applying the NB for feature selection in the second experiment, the probability of the word feature occurrence in a text document is independent of the word's position and the occurrence of other words in the text document.

The data was partitioned using 10-fold cross-validation to ensure the robustness of the results. Applying NB classifier with eighty two features produced the lowest accuracy in the IG feature selection from the first experiment with an accuracy of 80.61% as shown in Table 7.1. However, the accuracy increased when the feature set reduced to eight of eleven features in the second experiment as shown in Table 7.3.

Table 7.3 Experiment results of using Naive Bayes classifier with the features selected through the Hybrid feature selection

Feature Selection	Recall	Precision	F-measure	Accuracy
IG+NB (10-fold CV)	87.26% ± 0.47%	83.38% ± 0.38%	85.28% ± 0.25%	82.43% ± 0.58%
IG+SVM (10-fold CV)	87.26% ± 0.47%	83.38% ± 0.38%	85.28% ± 0.25%	82.43% ± 0.58%
IG+ANN (10-fold CV)	87.17% ± 0.39%	83.29% ± 0.32%	85.19% ± 0.17%	82.33% ± 0.45%
IG+DT (10-fold CV)	87.00% ± 0.55%	81.93% ± 0.58%	84.38% ± 0.41%	81.23% ± 0.81%

Artificial Neural Network (NN)

The NN algorithm is used based on FeedForward Learning with two inner layers with 100 output units each, and learning rate of 0.1. XAVIER initialization weight strategy [59] is used with ReLU Activation Function. The number of training iteration is one. The optimization Algorithm used is Stochastic Gradient Descent(SGD). The loss function that used is Mean Squared Error. he data was partitioned using 10-fold cross-validation to ensure the robustness of the results. Applying the hybrid feature selection shows an improvement in an accuracy as ANN improves the accuracy into 82.23%, as seen in table 7.4.

Table 7.4 Experiment results of using FeedForward Neural Network classifier with the features selected through the Hybrid feature selection in experiment 2

Feature Selection	Recall	Precision	F-measure	Accuracy
IG+NB (10-fold CV)	87.48% ± 0.42%	82.21% ± 0.34%	84.76% ± 0.22%	81.65% ± 0.54%
IG+SVM (10-fold CV)	88.17% ± 0.47%	82.55% ± 0.39%	85.26% ± 0.25%	82.23% ± 0.57%
IG+ANN (10-fold CV)	87.97% ± 0.40%	82.29% ± 0.33%	85.0% ± 0.18%	81.94% ± 0.48%
IG+DT (10-fold CV)	88.36% ± 0.51%	81.83% ± 0.56%	84.97% ± 0.38%	81.77% ± 0.77%

Support Vector Machine (SVM)

SVM which is based on LibSVM algorithm [27] has been used with overlapping penalty set to one, kernel used is Radial Basis Function (RBF) with Gamma equals to one. The data was partitioned using 10-fold cross-validation to ensure the robustness of the results. For SVM classification technique, the highest accuracy is reduced when applying hybrid comparing to IG as shown in tables 7.1 and 7.5.

Table 7.5 Experiment results of using Support Vector Machine classifier on the features selected through the Hybrid feature selection in experiment 2

Feature Selection	Recall	Precision	F-measure	Accuracy
IG+NB (10-fold CV)	86.10% ± 0.41%	82.55% ± 0.36%	84.28% ± 0.25%	81.28% ± 0.52%
IG+SVM (10-fold CV)	86.10% ± 0.41%	82.55% ± 0.36%	84.28% ± 0.25%	81.28% ± 0.52%
IG+ANN (10-fold CV)	86.28% ± 0.38%	82.72% ± 0.34%	84.46% ± 0.21%	81.49% ± 0.47%
IG+DT (10-fold CV)	85.83% ± 0.49%	81.45% ± 0.47%	83.58% ± 0.32%	80.34% ± 0.68%

7.1.5 Results Discussion

The results of the first experiment, using information gain (IG) feature selection, showed that decision tree (DT) achieved the highest accuracy with 83.34%. This indicates that the selected features contain enough information to make accurate predictions. However, applying two phases of feature selection to reduce the number of features did not show significant improvement in accuracy, although it significantly reduced data processing time. Figure 7.5, illustrates the accuracy for the learning classification models over the collected dataset for the two-phase approach. It is worth noting that while most of the classification techniques produced high accuracy results, they still do not fully capture the context of the tweets and may not represent the intention accurately. This is partly due to the labelling process, which was done by searching for certain strings in the tweet texts, meaning that tweets with similar phrases to the intention vector were labelled, regardless of their true intent. This approach may not be effective for all tweets, as certain phrases may have different meanings depending on context.

Furthermore, the limited search words used to retrieve data from Twitter may have also limited the accuracy of the results. Using more words and terms could potentially provide more accurate and diverse data, and lead to better results.

Overall, this experiment highlights the limitations of data mining approaches in accurately capturing the intent of social media users. It emphasizes the need for more nuanced and sophisticated approaches to intent mining that take into account context, user behaviour, and other factors that can influence the interpretation of social media data.

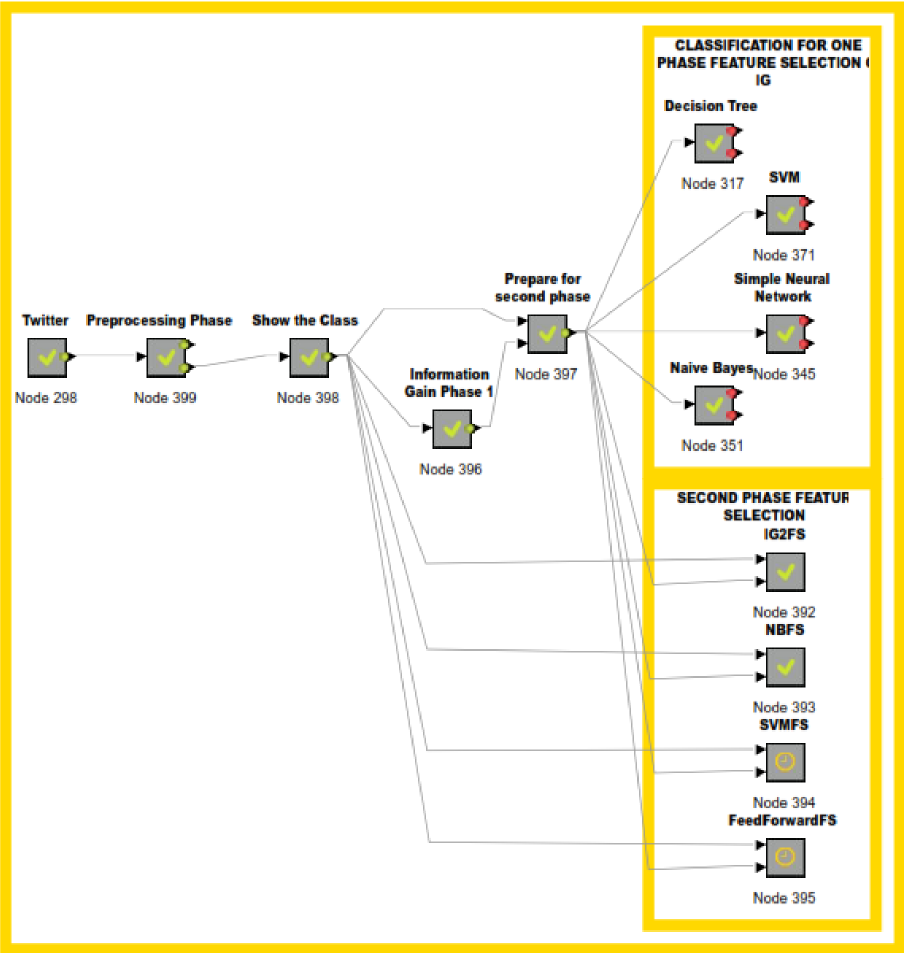
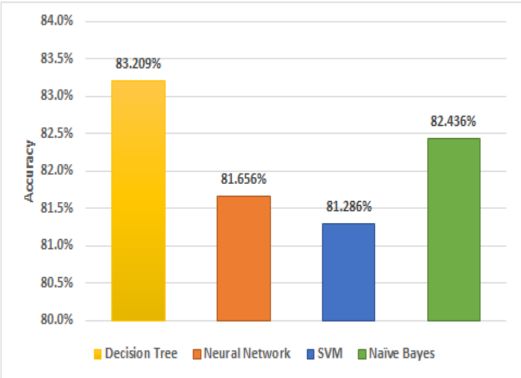


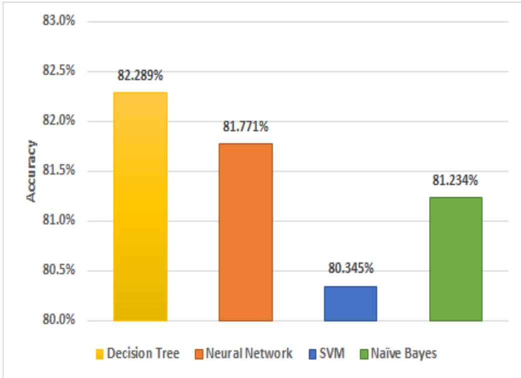
Fig. 7.4 Over all Workflow of The Experiments Using Knime

7.2 Sentiment Analysis: Experimental Methodology

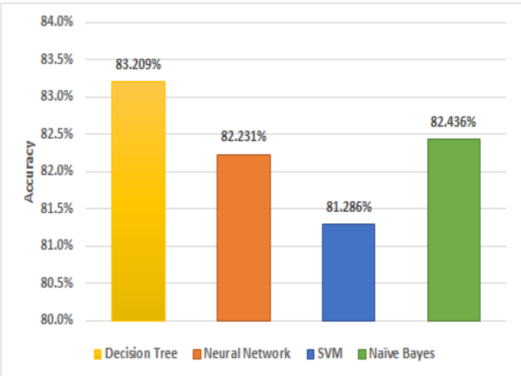
The second experiment conducted in this study aimed to explore the effectiveness of NRC-VAD lexicons in sentiment analysis for intent mining. The experiment utilized the same setup and classification techniques as the first experiment, but with a different focus on the feature selection process. By using NRC-VAD lexicons, the study aimed to investigate whether incorporating affective word features could improve the accuracy of intent mining, particularly in the context of sentiment analysis. This section will describe the experiment setup, pre-processing techniques, and feature selection methods used in the second experiment.



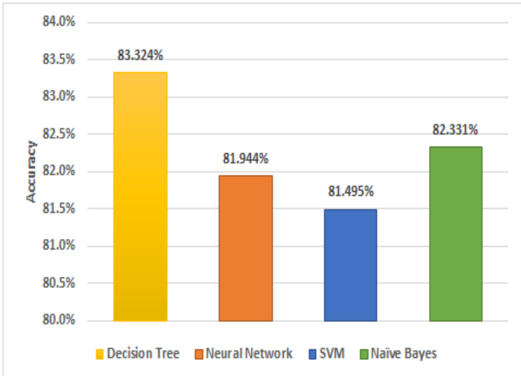
(a) Accuracy of Applying Classification after Two-Stage 8 Features Selection Based on IG and NB



(b) Accuracy of Applying Classification after Two-Stage 8 Features Selection Based on IG and IG



(c) Accuracy of Applying Classification after Two-Stage 8 Features Selection Based on IG and SVM



(d) Accuracy of Applying Classification after Two-Stage 8 Features Selection Based on IG and ANN

Fig. 7.5 Overall Accuracy of Various Feature Selection and Classifiers using the Information Gain Algorithm and applied algorithms NB, SVM, ANN, and IG, and the same set of words selected by the two-stage

7.2.1 Experiment Setup

The second experiment, which focuses on sentiment analysis for intent mining, was conducted using the same experimental methodology and classification techniques as the first experiment. The aim of this experiment is to investigate the impact of using NRC-VAD lexicons on the accuracy of intent mining.The NRC-VAD lexicon provides a rating of words on three dimensions: valence (positive/negative), arousal (activation/sleepiness), and dominance (weak/strong). The aim of using the NRC-VAD lexicon is to improve the accuracy of sentiment analysis and consequently improved the intent mining process. The experiment used KNIME to build the sentiment analysis pipeline and classify the intent of tweets.

7.2.2 Feature Selection

To select the most relevant features for sentiment analysis, two feature selection methods were employed in this experiment: Information Gain (IG) and NRC-VAD lexicons. The IG method was used to select the most important features from the text data, while the NRC-VAD lexicons were used as an alternative feature selection method to compare against the IG method. These lexicons provide a way to extract sentiment features from text data based on the sentiment scores of individual words.

For the IG method, the feature selection was performed on the training data using KNIME software, and the selected features were then used to train the classification models. The IG method works by calculating the information gain of each feature, which measures how much that feature contributes to the classification of the data. The higher the information gain, the more important the feature is for classification. The top 82 features with the highest information gain were selected, where 82 was determined by the experiment.

Table 7.6 The distribution of the intentions with top 10 keywords based on IG and NRC- VAD

Intention	Keywords	#Posts
I0	start, new, learn, watch, tri, need, buy, go, help, goal	380
I1	gym, work, god, week, time, day, gain, stop, chang, start	302
I2	year, like, goal, let, ve, make, go, hope, happi, think	339
I3	peopl, stop, make, ralli, think, start, time, good, care, hour	271
I4	f***, twitter, stop, get, seve, break, word, past, lose, thing	252
I5	day, make, go, stop, tweet, tri, play, time, focus, selfi	262
I6	stop, drink, smoke, right, cigarett, chrisbrown, club, na, plan, supper	288
I7	better, friend, make, time, continu, love, shit, awesom, use, spend	279
I8	eat, happi, amp, new, healthi, game, year, cariloha, need, food	308
I9	make, money, possible, thank, live, life, tapp, run, happy, like	284

For the NRC-VAD lexicon method, the feature selection was performed by matching the words in the text data to the NRC-VAD lexicons and using their sentiment scores as features. Specifically, the valence scores were used as positive and negative sentiment features, and the arousal scores were used as intensity features.

Both feature selection methods were evaluated using the same classification models as in the previous experiment: Naive Bayes, Decision Tree, Support Vector Machine, and Artificial Neural Network. The performance of each classification model was compared using the selected features from both feature selection methods, and the results were analysed to determine which method produced the best performance for sentiment analysis.

Table 7.7 Results for using NRC-VAD features to build four classification models SVM, DT, NB, and NN with multi-intentions for each post

Category	SVM		DT		NB		NN	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
Personal Growth	78%	48%	71%	0%	74%	12%	75%	38%
Humor	82%	52%	76%	34%	74%	8%	79%	46%
Health/Fitness	80%	36%	76%	5%	76%	5%	82%	46%
Recreation/Leisure	83%	55%	79%	36%	78%	5%	80%	51%
Career	77%	62%	70%	30%	75%	44%	78%	65%
Education/Training	90%	54%	87%	47%	85%	0%	85%	55%
Finance	81%	40%	81%	37%	78%	10%	82%	45%
Family/Friends/Relationships	77%	45%	72%	20%	73%	14%	78%	44%
Time Management/Organization	72%	76%	57%	70%	65%	74%	71%	75%
Philanthropic	81%	56%	77%	28%	75%	15%	76%	47%

It was noted that unlike the first experiment, where a wide range of features were used, including text-based features such as bag-of-words and n-grams, the focus of the second experiment was solely on the affective features derived from the NRC-VAD lexicon. From the results shown in the table 7.7, applying the SVM classifier on multi-classification in multi-intentions problem presents the best result with respect to accuracy when compared to DT, NB and NN. In addition, predicting intentions, such as Recreation & Leisure and Education & Training both get the highest accuracy in all the classifiers, see table 7.7, which indicates that features that describe the performance of those two categories are most often represented in the text.

7.3 Fuzzy Logic Inference System: Experimental Methodology

The proposed FL intention inference system was designed based on the discussion in Chapter 6 using the Mamdani fuzzy inference approach. The Trapezoidal fuzzy membership function was chosen to convert each input attribute's numerical value into a fuzzy linguistic term, as it is widely used in practice [34]. The following subsections detail the proposed FL implementation for the intention estimation problem.

7.3.1 Overview Fuzzy Inference Process

The fuzzy inference process is composed of four main steps: fuzzification, rule evaluation, aggregation, and defuzzification. This process enables the fuzzy inference system to combine the TPB fuzzy rules and produce a final crisp output.

- *Fuzzification:* In this step, the crisp input values are converted into fuzzy sets using the membership functions defined for each input variable (valence, arousal, and dominance). This provides a degree of membership for each input variable in the corresponding fuzzy sets (Negative, Neutral, and Positive for valence; Low, Medium, and High for arousal and dominance).

- *Rule Evaluation:* The fuzzy rules are evaluated based on the fuzzy input values. The degree of membership for each rule's antecedent (input) is determined using the fuzzy operators (AND, OR) defined earlier. The AND operator takes the minimum of the input memberships, while the OR operator takes the maximum. This results in a firing strength for each rule.

- *Aggregation:* The firing strengths of all rules are combined to form an aggregated output fuzzy set for each output variable (ATB, SN, PBC, and intention). The aggregation process usually involves taking the maximum of the firing strengths for each rule that contributes to a particular output fuzzy set.

- *Defuzzification:* The aggregated output fuzzy sets are then converted back into crisp values using a defuzzification method. In the context of your implementation, you could mention the specific defuzzification method used, such as the centroid or centre of gravity method. This final crisp output represents the estimated intention based on the TPB model variables.

7.3.2 TPB Fuzzy Rules in MATLAB

The TPB model variable rules are represented in MATLAB based on the discussion provided in Chapter 6, Section 6.3.3. The following summarizes the rules for ATB, SN, PBC, and intention:

- **ATB Fuzzy Rules:** ATB depends on the valence, arousal, and dominance variables of the NRC-VAD. The complete list of rules can be found in Section 6.3.3.
  $(ATB, \mu) = \mu_x and \mu_y and \mu_w$

  1) "$Valence == Negative \& Arousal == Low \& Dominance == Weak => ATB = Negative(1)$"
  .
  .
  .
  27) "$Valence == Positive \& Arousal == High \& Dominance == Strong => ATB = Positive(1)$"

- **SN Fuzzy Rules:** SN fuzzy rules are represented in MATLAB using normative beliefs (NB) and motivation to comply (MC):

  1) "$NB == Low || MC == Low => SN = Low(1)$"
  .
  .
  .
  9) "$NB == High || MC == High => SN = High(1)$"

- **PBC Fuzzy Rules:** PBC fuzzy rules are represented in MATLAB using valence and arousal:

  1) "$Valence == Negative \& Arousal == Low => PBC = Low(1)$"
  .
  .
  .
  9) "$Valence == Positive \& Arousal == High => PBC = High(1)$"

- **Intention Fuzzy Rules:** The intention fuzzy rules are represented in MATLAB using ATB, SN, and PBC:

1)"$ATB == Negative\&SN == Low\&PBC == Low => Intention = Low(1)$"

.

.

.

27)"$ATB == Positive\&SN == High\&PBC == High => Intention = High(1)$"

7.3.3 FL Model

The fuzzy model was implemented using MATLAB FL designer toolbox. The prototype uses a fuzzy inferencing system editor (FISE) based on the Mamdani inferencing system. This FIS utilised the minimum (And) operator, maximum (Or) operator and the centroid defuzzification method. The proposed TPB fuzzy intention model was designed based on the sentiment analysis list of five fuzzy discourse domains, valence, arousal, and dominance in addition to normative belief and motivation to comply . These domains were used as inputs to estimate the TPB intention model variables ATB,PBC, and SN.

For ATP factor, FIS was implemented with three inputs and one output, as given in Fig.7.6.

The three inputs, valence, arousal and dominance, were implemented as a fuzzy construct with three membership functions. These functions were represented as Trapezoidal functions with geometric vertices given in tables 7.8, 7.9 and 7.10, respectively.

Table 7.8 Input valence Membership Functions

Negative	0	0	0.3272	0.4181
Neutral	0.3195	0.4795	0.6305	0.7378
Positive	0.654	0.7278	1	1

Table 7.9 Input arousal Membership Functions

Low	0	0	0.2456	0.3748
Medium	0.2585	0.4103	0.5678	0.6947
High	0.566	0.727	1	1

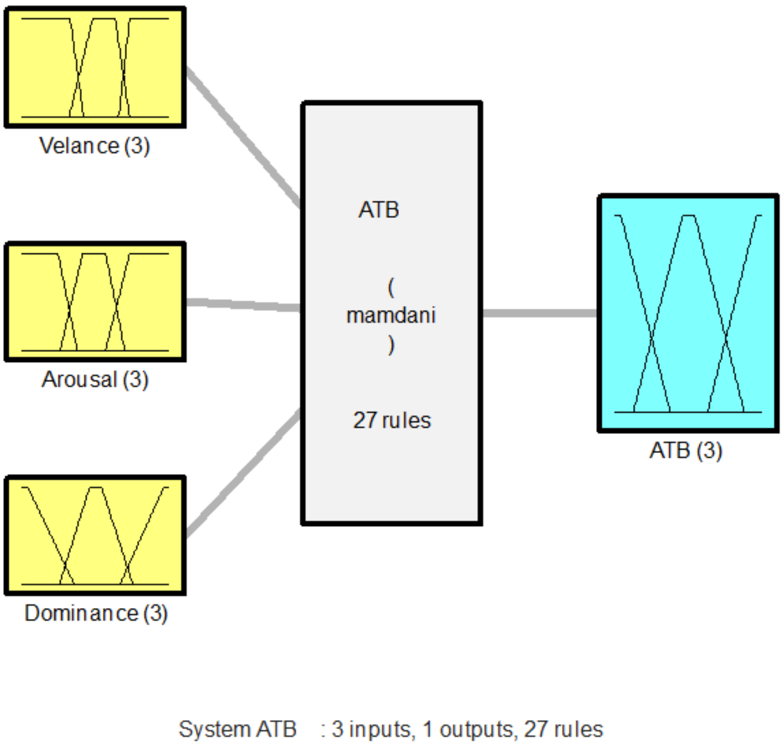


Fig. 7.6 ATB FIS structure

Table 7.10 Input dominance Membership Functions

Weak	-0.3749	-0.0418	0.0418	0.3548
Medium	0.2544	0.4582	0.5418	0.7508
Strong	0.6629	0.9582	1	1

The output ATB had three membership functions, as shown in Fig.7.7.

For SN factor, FIS was implemented with two inputs and one output as given in Fig.7.8. The two inputs were NB which is implemented by having three membership functions and MC with three membership functions. NB was represented as geometric vertices given in tables 7.11 and 7.12.

The output (SN) had three membership functions as shown in Fig.7.9 .

Fort PBC factor, FIS was implemented with two inputs and one output as given in Fig.7.10. The two inputs were valence which was implemented by having three membership functions and arousal with three membership functions. both valence and arousal were represented in tables 7.8 and 7.9 respectively.

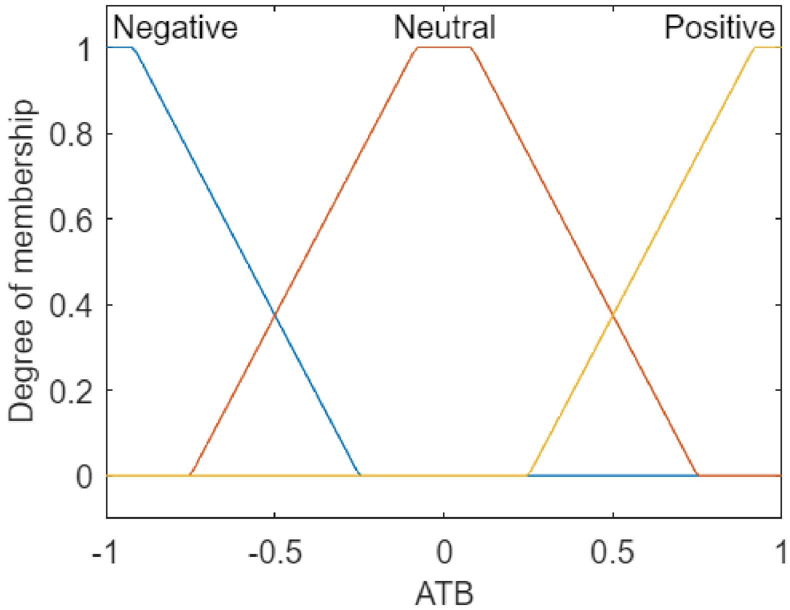


Fig. 7.7 ATB Output Membership Functions

Table 7.11 Input NB MFS

Low	-0.375	-0.04167	0.04167	0.375
Medium	0.125	0.4583	0.5417	0.875
High	0.625	0.9583	1	1

Table 7.12 Input MC MFS

Low	-0.375	-0.04167	0.04167	0.375
Medium	0.125	0.4583	0.5417	0.875
High	0.625	0.9583	1	1

The output (PBC) had three membership functions as shown in Fig.7.11.

For Intention, FIS was implemented with three inputs and one output as given in Fig.7.12.

The three inputs (ATB, SN, and PBC) were implemented by having three membership functions. For ATB, the membership functions are represented as Trapezoidal functions with geometric vertices given in table 7.13.

For SN, the membership functions were represented as Trapezoidal functions

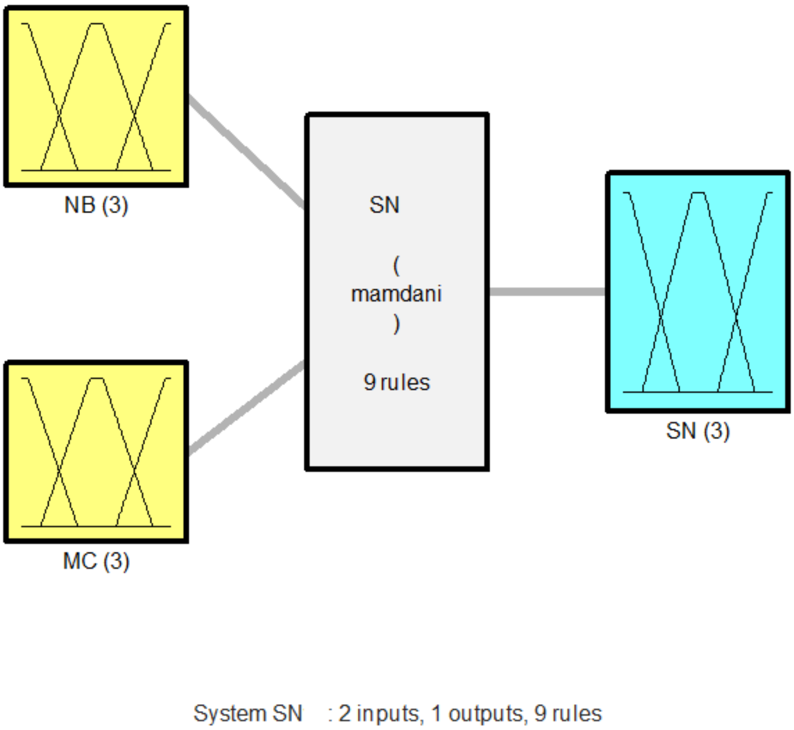


Fig. 7.8 SN FIS structure

Table 7.13 Input ATB MFS

Negative	-1.9	-1.1	-0.9	-0.1
Neutral	-0.9	-0.1	0.1	0.9
Positive	0.1	0.9	1.1	1.9

with geometric vertices given in table 7.14.

Table 7.14 Input SN MFS

Low	-1.9	-1.1	-0.9	-0.1
Medium	-0.9	-0.1	0.1	0.9
High	0.1	0.9	1.1	1.9

For PBC, the membership functions were represented as Trapezoidal functions

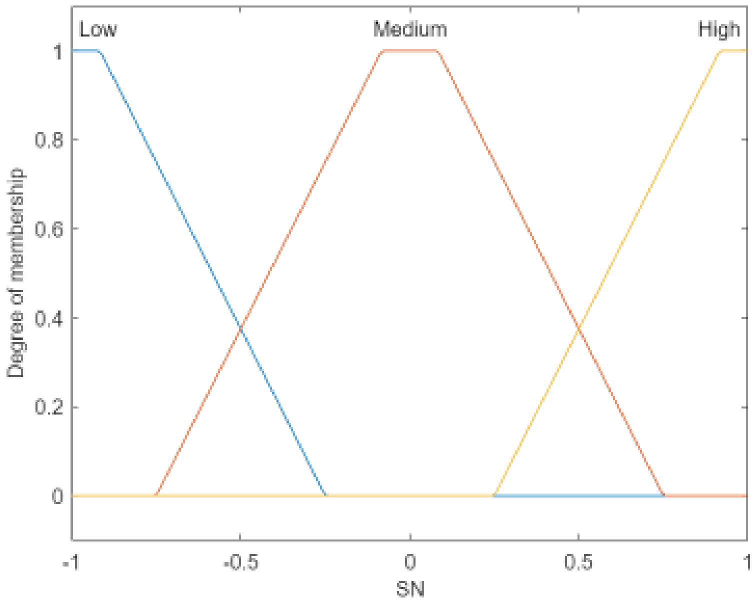


Fig. 7.9 SN Output Membership Functions

with geometric vertices given in table 7.15.

Table 7.15 Input PBC MFS

Low	-1.9	-1.1	-0.9	-0.1
Medium	-0.9	-0.1	0.1	0.9
High	0.1	0.9	1.1	1.9

The intention output had three membership functions each, as shown in Fig.7.13. The Intention output membership functions shown in Figure 7.13 represent the fuzzy sets that correspond to different levels of intention, such as Low, Medium, and High. These membership functions were used to aggregate the firing strengths of all the rules in the Intention FIS and form an aggregated output fuzzy set for the Intention variable. The defuzzification process then converted this aggregated output fuzzy set into a single crisp value, which represents the final estimated intention. This crisp value is the final output of the overall fuzzy inference system, allowing the analysis and prediction of an individual's behaviour based on the TPB model variables.

After presenting the membership functions for the Intention FIS, that illus-

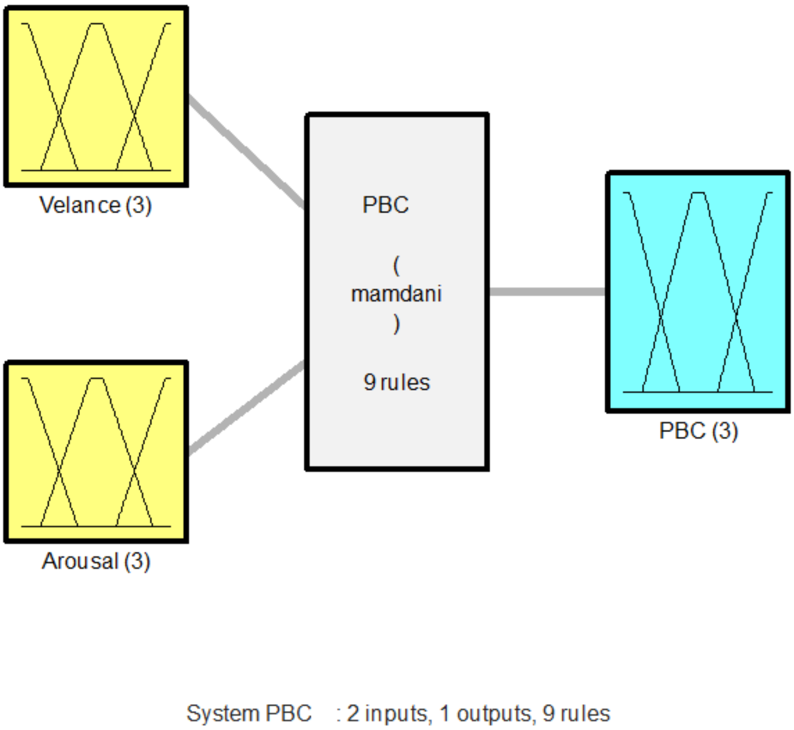


Fig. 7.10 PBC FIS structure

trating how the input variables (ATB, SN, and PBC) were transformed into fuzzy sets using trapezoidal functions. Now, the focus will be on the defuzzification process, a crucial step in converting the aggregated output fuzzy set into a single crisp value representing the final estimated Intention based on the TPB model variables.

The defuzzification process is visually represented in Figure 7.14. The x-axis shows the Intention ranging from -1 (Low) to 1 (High), and the y-axis represents the Output Membership values, spanning from -1 to 1. The Aggregated output fuzzy set curve in the plot illustrates the combined firing strengths of all the rules in the Intention FIS, forming an aggregated output fuzzy set for the Intention variable. This curve provides insight into how the different rules contribute to the overall output fuzzy set based on the input variable memberships. The Defuzzified output line (x = 0.4556) signifies the final estimated Intention, obtained by applying the defuzzification process, such as the centroid method, to the

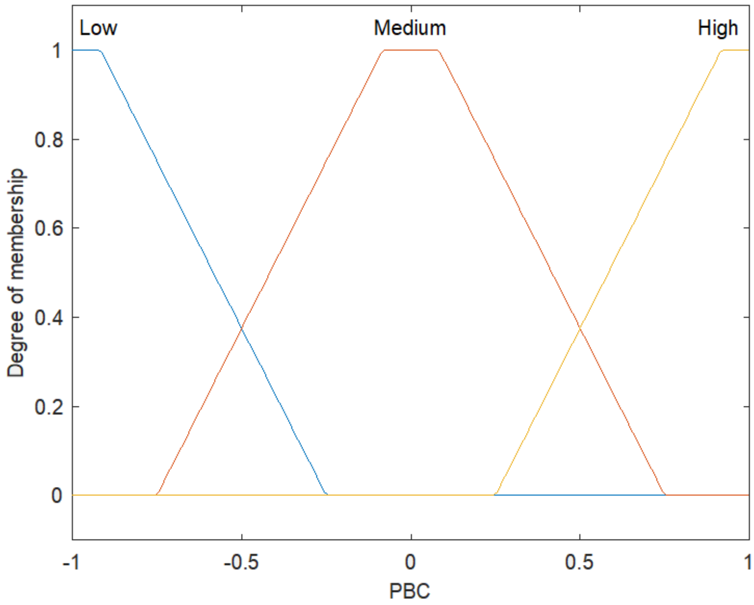


Fig. 7.11 PBC Output Membership Functions

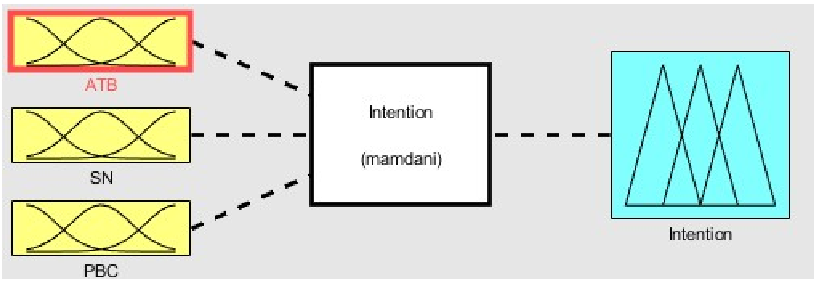


Fig. 7.12 Intention FIS structure

aggregated output fuzzy set. The intersection point of the Defuzzified output line and the Aggregated output fuzzy set curve (x = 0.4556, y = 0.254) represents the centroid of the aggregated output fuzzy set, serving as the crisp output of the Intention FIS.

To better understand the relationships between the input variables (ATB, SN, and PBC) and their combined effect on the Intention output, three 3D plots are presented, illustrating the interactions between these variables. The shape of these surfaces indicates how the combinations of input variables contribute to the Intention estimation.

- **ATB and SN interaction:** In Figure 7.15a, the surface representing the interaction between ATB and SN forms a distinct pattern. This pattern

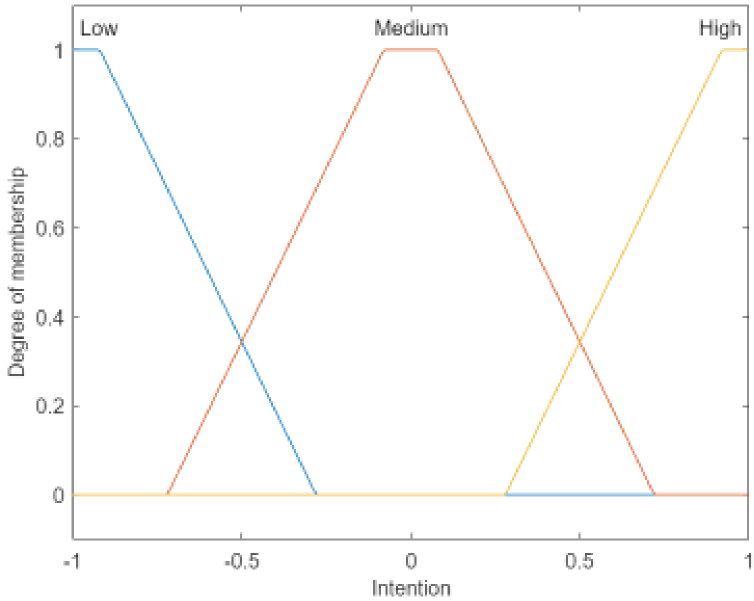


Fig. 7.13 Intention Output Membership Functions

reflects how the combinations of high and low values of ATB and SN can lead to different degrees of Intention. The highest and lowest points on the surface correspond to specific combinations that result in the most significant increase or decrease in Intention, respectively. Figure 7.15a reveals an interesting pattern, where the influence of SN in social networks seems to be more significant than that of ATB, especially when it comes to driving the Intention output. As observed in the plot, even with negative or positive ATB, the Intention output remains relatively low compared to when the SN increases. However, when SN increases, the Intention output increases with the ATB.

This observation implies that social norms within social networks are critical in shaping individuals' intentions to perform a specific behaviour. In this context, the social pressure exerted by peers and the desire to conform to social expectations substantially impact intention more than the individual's attitudes towards the behaviour.

This finding highlights the importance of considering the role of social networks when designing interventions or strategies aimed at promoting specific behaviours. By leveraging the power of social norms and the

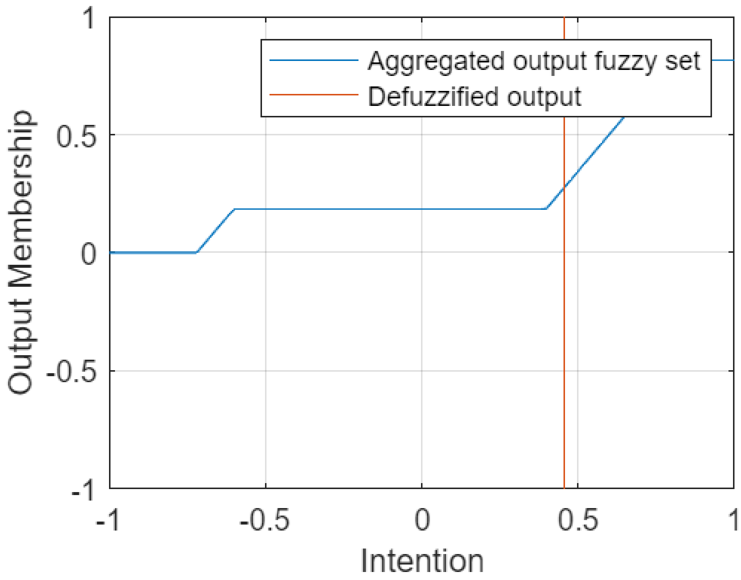


Fig. 7.14 Defuzzification of the Intention FIS Output

influence of peers within social networks.

This insight can also inspire further research to understand better the underlying mechanisms through which social norms in social networks impact individuals' intentions and to explore the factors that contribute to the formation and change of such norms.

- **ATB and PBC interaction:** Figure 7.15b illustrates the interaction between ATB and PBC in determining the Intention output. While both ATB and PBC appear to have a similar influence on the output, it is noteworthy that PBC demonstrates a slightly more substantial impact than ATB. This finding suggests that individuals' perceived behavioural control may play a more significant role in shaping their intentions than their attitudes towards the behaviour.

  This observation aligns with the theory of planned behaviour, which posits that perceived behavioural control is a critical determinant of Intention. In social networks, the users' perception of their ability to perform a specific action or behaviour might influence their Intention more than their attitudes.

  This insight emphasizes the importance of considering ATB and PBC when modelling social network users' intentions, as both factors contribute to shaping the outcome. It also supports the relevance of research question 1, which aims to explore the use of the theory of planned behaviour in modelling users' intentions.

Moreover, this finding encourages further research to investigate the factors contributing to users' perceived behavioural control and how these factors interact with attitudes to shape Intention. Such research can contribute to a deeper understanding of the interplay between ATB and PBC and help develop more effective strategies to influence users' intentions on social networks.

- **SN and PBC interaction:** Figure 7.15c demonstrates a similar pattern, where the influence of SN in social networks appears to be more significant than that of PBC in determining the Intention output, in line with the research question on modelling social network users' Intention to perform specific actions using the theory of planned behaviour. As observed in the plot, despite varying levels of PBC, the Intention output remains relatively low compared to when SN increases. However, the Intention output increases when SN increases, regardless of the PBC levels.

  This observation suggests that, similar to the ATB-SN relationship, social norms within social networks play a crucial role in shaping individuals' intentions, even when considering their perceived behavioural control. In this context, the social pressure and the desire to conform to social expectations have a more profound impact on Intention than the individual's perception of their ability to perform the behaviour.

  The finding further emphasizes the importance of applying fuzzy logic to handle the uncertainty and vagueness in users' decision-making processes when modelling social network users' intentions. By harnessing the power of social norms and the influence of peers within social networks, practitioners can create more effective campaigns and initiatives that lead to higher intention levels and, ultimately, a greater likelihood of individuals engaging in the desired behaviour.

  In addition, using sentiment analysis to evaluate users' opinions and emotions towards different products, services, and topics provides valuable insights for modelling users' intentions on social networks. This approach can also stimulate additional research to comprehend better the underlying mechanisms through which social norms in social networks affect individuals' intentions in the presence of varying levels of perceived behavioural control.

This line of inquiry can explore the factors contributing to the formation and change of such norms and their interplay with individuals' perceptions of control over their behaviour. The effectiveness of the proposed approach in modelling social network users' intentions compared to existing methods can further be assessed and refined, paving the way for more accurate and reliable predictions of user behaviour on social media platforms.

Considering the interactions between ATB, PBC, and SN, and their influence on the Intention output, a comprehensive understanding of the relationships among these variables emerges. The 3D plots in Figures 7.15b, 7.15c, and 7.15a demonstrate the varying degrees of impact each input variable has on shaping individuals' intentions in the context of social networks.

Overall, it is evident that social norms (SN) have a more significant influence on Intention output compared to both attitudes towards the behaviour (ATB) and perceived behavioural control (PBC). However, PBC shows a slightly more substantial impact on Intention than ATB, highlighting the importance of individuals' perception of their ability to perform the desired behaviour.

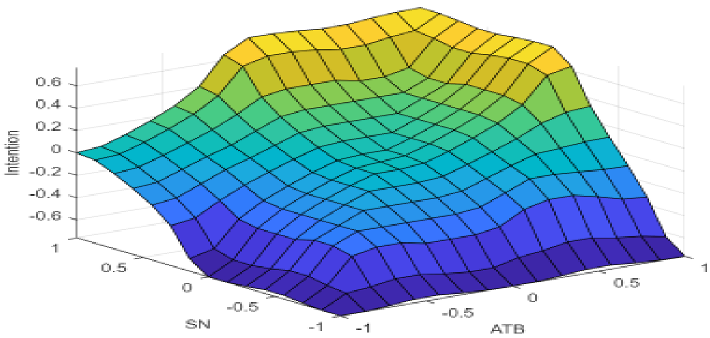
When modelling social network users' intentions, these observations emphasize considering all three input variables (ATB, PBC, and SN). Each factor contributes differently to shaping the outcome. This understanding is consistent with the research questions, which aim to explore the use of the theory of planned behaviour, fuzzy logic, and sentiment analysis in the context of modelling social network users' intentions.

Moreover, the insights gained from these 3D plots can inform the development of more effective interventions or strategies to promote specific behaviours on social networks. By considering the relative influence of ATB, PBC, and SN, practitioners can tailor their campaigns and initiatives to target the most impactful factors, leading to higher intention levels and a greater likelihood of individuals engaging in the desired behaviour.

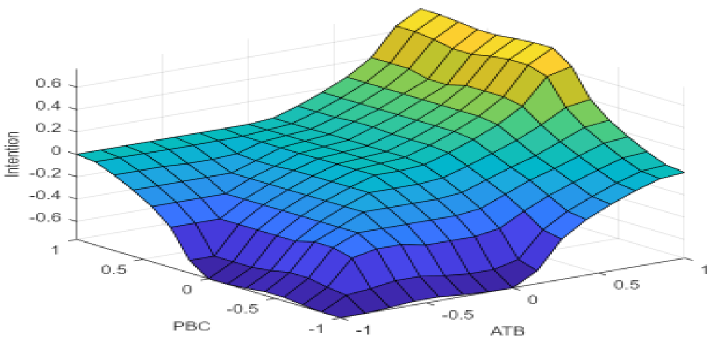
Furthermore, this comprehensive understanding of the relationships among input variables can inspire additional research to explore how these factors interact and shape individuals' intentions. Such research can contribute to advancing knowledge in the field and improve the effectiveness of strategies for influencing social network users' behaviours.

7.4 Summary

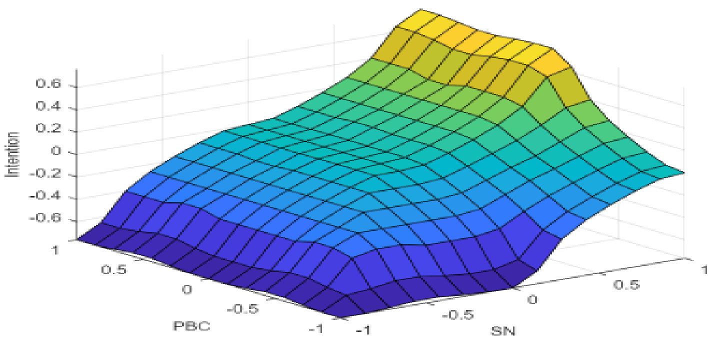
In order to provide a minimal number of features that estimate the user's intention, a set of feature selection techniques were used. Various classifiers were used as well, such as Decision Tree, Naive Based, Support Vector Machine, to verify the predicate features whether the user's data is expected. Feature Selection is the method of choosing the right minimum feature set from the total features available. Embedded feature-based filtering models use a part of the learning process of the supervised learning algorithm for the selection of features. A hybrid solution was proposed using IG as a filter feature selection model since the problem is a supervised learning problem. The proposed hybrid feature selection approach is based on the threshold of the features set to determine the number of features that exceed the maximum score in the form of the terms vector. The first part is a hybrid feature selection based on IG feature selection. Followed by the Forward-Feature Selection Algorithm (FFS).



(a) 3D plot of Intention as a function of ATB and SN interaction



(b) 3D plot of Intention as a function of ATB and PBC interaction



(c) 3D plot of Intention as a function of SN and PBC interaction

Fig. 7.15 3D plots of the relationships between input variables (ATB, SN, and PBC) and Intention output.

Chapter 8

Conclusion and Future Research

This thesis aimed to develop a novel approach to model and predict social network users' intentions based on the Theory of Planned Behaviour (TPB) using fuzzy logic (FL) in the context of sentiment analysis. Throughout this research, critical research questions have been addressed, a comprehensive understanding of the factors influencing social network users' intentions have been developed, and a novel approach for modelling and predicting user intentions using FL and sentiment analysis have been proposed.

8.1 Summary of Findings

The key findings of this research can be summarized as follows:

- The meticulously hydrated and curated corpus dataset of "NewyearResolution2015" has been designed to address the study's research questions. It encompasses tweet textual features, user features, and network feature sets, aiming to estimate user intentions on Twitter based on their behaviour, interests, and connections within their social network. The database structure is adeptly organised, representing each user profile as a row and the columns containing the extracted features of textual, user, and network attributes.

- The Theory of Planned Behaviour (TPB) has proven to be a pragmatic framework for modelling social network users' intentions to perform specific actions. By scrutinising users' attitudes, subjective norms, and perceived behavioural control, insights into the factors influencing their social network behaviour were obtained.

- Fuzzy Logic (FL) has emerged as a beneficial approach for managing the uncertainty and vagueness inherent in users' decision-making processes when modelling social network users' intentions. Incorporating FL in the proposed methodology allowed us to account for the inherent ambiguity and complexity of social media data, ultimately enhancing the prediction accuracy of user intentions.

- Sentiment analysis, employing the NRC Valence, Arousal, and Dominance (VAD) Lexicon, assessed users' opinions and emotions towards various topics, products, and services concerning modelling users' intentions on social networks. This facilitated a more comprehensive examination of user sentiment and intention on social networks, ultimately guiding the development of more effective and efficient recommendation systems and other technologies.

- The proposed approach, which combines TPB and FL, exhibited heightened efficacy in modelling social network users' intentions compared to existing methods. This study represents the first of its kind to apply FL in the context of sentiment analysis for predicting social network users' intentions based on the TPB, thereby underscoring the novelty and potential of the proposed methodology.

8.2 Implications

The findings of this research have significant implications for various stakeholders involved in the design, development, and management of social network platforms, as well as for researchers and practitioners working in sentiment analysis and social network user intention prediction. By understanding the factors that influence users' intentions and decision-making processes, this research can inform the development of more effective recommendation systems, marketing strategies, public awareness campaigns, and educational initiatives to promote responsible and informed use of social media platforms.

8.3 Limitations and Future Research

Despite the promising findings of this research, several limitations should be acknowledged. First, this study focused on Twitter data, and the findings may need to be more generalizable to other social media platforms. Future research should explore the applicability of the proposed approach to different platforms, such as Facebook, Instagram, and LinkedIn.

Second, the study utilized a limited number of datasets for the analysis, which may not fully capture the diversity of social network users' intentions. Future research could expand on this work using additional datasets, such as MovieTweetings, Amazon dataset, AffectiveTweets, or other user-generated content.

Third, this research focused on using the NRC-VAD Lexicon for sentiment analysis. Future studies could investigate the performance of other sentiment analysis lexicons and their suitability for modelling user intentions.

Lastly, this research employed a Mamdani trapezoidal fuzzy logic system. Future work could explore the use of other types of fuzzy logic systems, such as fuzzy neural systems or type-2 fuzzy systems, to further improve the accuracy and robustness of the proposed approach.

8.4 Final Remarks

In conclusion, this thesis has significantly contributed to sentiment analysis and social network user intention prediction by developing a novel approach combining TPB and FL. The results of this research will be of interest to researchers, practitioners, and businesses alike, as it has the potential to inform the development of new technologies and services that support the effective use of social networks.

Through the proposed approach, the effectiveness of incorporating FL and sentiment analysis in modelling and predicting social network users' intentions based on the TPB have been demonstrated. This research has expanded the current understanding of the factors influencing user behaviour on social networks, paving the way for future studies in this field.

As social media platforms grow and become increasingly integrated into our daily lives, understanding and predicting user intentions will be crucial in developing more effective and efficient technologies, services, and interventions. The findings of this thesis serve as a foundation for future research and innovation in the field of social network user intention prediction and sentiment analysis, ultimately benefiting users and stakeholders across various domains.

In the coming years, it is anticipated that the insights and contributions derived from this research will play a significant role in influencing the design, development, and management of social media platforms. This influence is expected to result in the enhancement of recommendation systems, the refinement of targeted marketing strategies, and the facilitation of informed decision-making for both users and stakeholders.

References

[1] Adedoyin-Olowe, M., Gaber, M. M., and Stahl, F. (2013). A Survey of Data Mining Techniques for Social Media Analysis. *Journal of Data Mining & Digital Humanities*, 2014:25.

[2] Agarwal, S. and Sureka, A. (2016). But i did not mean it!—intent classification of racist posts on tumblr. In *2016 European Intelligence and Security Informatics Conference (EISIC)*, pages 124–127.

[3] Agarwal, S. and Sureka, A. (2017). Analysis of Linguistic Features for Classification of Racist/Radicalized Posts on Tumblr.

[4] Aggarwal, C. C. (2011). *Social Network Data Analytics*, volume 70. Springer US.

[5] Agrawal, S. and Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, 60(1):708–713.

[6] Ajzen, I. (1991). The theory of planned behavior. *Orgnizational Behavior and Human Decision Processes*, 50:179–211.

[7] Ajzen, I. (2006). Theory of planned behavior diagram. Available at: https://people.umass.edu/aizen/tpb.diag.html. Accessed: 2023-03-07.

[8] Antwarg, L., Rokach, L., and Shapira, B. (2012). Attribute-Driven Hidden Markov Model Trees for Intention Prediction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1103–1119.

[9] Aziz, Y. A. and Chok, N. V. (2013). The role of halal awareness, halal certification, and marketing components in determining halal purchase intention among non-muslims in malaysia: A structural equation modeling approach. *Journal of International Food & Agribusiness Marketing*, 25(1):1–23.

[10] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC'10)*, volume 25, pages 2200–2204, Valletta, MT.

[11] Baccianella, S., Esuli, A., and Sebastiani, F. (2013). Using micro-documents for feature selection: The case of ordinal text classification. *Expert Systems with Applications*, 40(11):4687–4696.

[12] Bakker, I., van der Voordt, T., Vink, P., and de Boon, J. (2014). Pleasure, arousal, dominance: Mehrabian and russell revisited. *Current Psychology*, 33(3):405–421.

[13] Benamara, F., Taboada, M., and Mathieu, Y. (2016). Evaluative Language Beyond Bags of Words: Linguistic Insights and Computational Applications. *Computational Linguistics*, (October 2015):1–64.

[14] Berthold, K. T. M. (2012). Technical Report The KNIME Text Processing Feature: An Introduction. Technical report.

[15] Bhatt, V. and Nagvadia, J. (2021). *FACTORS INFLUENCING CONSUMER'S ONLINE BUYING BEHAVIOR: AN EMPIRICAL STUDY‖ Doctor of Philosophy in Management*. PhD thesis.

[16] Birjali, M., Kasri, M., and Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

[17] Blei, D. M., Griffiths, T. L., Gordon, M. I., and Tanenbaum, J. B. (2003). Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *NIPS'03*.

[18] Bourlai, E. and Herring, S. C. (2014). Multimodal communication on tumblr: "i have so many feels!". In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, page 171–175, New York, NY, USA. Association for Computing Machinery.

[19] Bradley, M. M. and Lang, P. P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. *University of Florida: The Center for Research in Psychophysiology*, Technical:0.

[20] Bravo-marquez, F., Frank, E., Mohammad, S. M., and Pfahringer, B. (2016). Determining Word – Emotion Associations from Tweets by Multi-Label Classification.

[21] Buckley, L., Kaye, S.-A., and Pradhan, A. K. (2018). Psychosocial factors associated with intended use of automated vehicles: A simulated driving study. *Accident Analysis & Prevention*, 115:202–208.

[22] Budalakoti, S. and Barber, K. S. (2010). Authority vs Affinity: Modeling User Intent in Expert Finding. In *2010 IEEE Second International Conference on Social Computing*, pages 371–378. IEEE.

[23] Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating Gender on Twitter. *Association for Computational Linguistics*, 146:1301–1309.

[24] Cambria, E., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

[25] Celikyilmaz, A., Hakkani-Tür, D., and Feng, J. F. J. (2010). Probabilistic model-based sentiment analysis of twitter messages.

[26] Chan, S.-Y. Y., Hui, P., and Xu, K. (2009). Community detection of time-varying mobile social networks. *Complex Sciences*, 4:1154–1159.

[27] Chang, K.-W., Yih, W.-t., and Meek, C. (2013). Multi-Relational Latent Semantic Analysis. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1612, Seattle, Washington, USA. Association for Computational Linguistics.

[28] Chang, S. E., Shen, W.-C., and Yeh, C.-H. (2016). A comparative study of user intention to recommend content on mobile social networks. *Multimedia Tools and Applications*, pages 1–19.

[29] Chaouali, W. (2016). Once a user, always a user: Enablers and inhibitors of continuance intention of mobile social networking sites. *Telematics and Informatics*, 33(4):1022–1033.

[30] Chaturvedi, I., Cambria, E., Welsch, R. E., and Herrera, F. (2018). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77.

[31] Chau, P. Y. and Hu, P. J.-H. (2002). Investigating healthcare professionals' decisions to accept telemedicine technology: an empirical test of competing theories. *Information & Management*, 39(4):297–311.

[32] Chen, H., Branavan, S. R. K., Barzilay, R., and Karger, D. R. (2009). Content Modeling Using Latent Permutations. *Journal of Artificial Intelligence Research*, pages 129–163.

[33] Chen, K.-H., Chen, L.-F., and Su, C.-T. (2013). A new particle swarm feature selection method for classification. *Journal of Intelligent Information Systems*, (510):1–24.

[34] Chen, S.-M. (1996). A fuzzy reasoning approach for rule-based systems based on fuzzy logics. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 26(5):769–778.

[35] Chen, Z., Lin, F., Liu, H., Liu, Y., Ma, W.-Y., and Wenyin, L. (2002). User IntentionModeling in Web Applications Using Data Mining. *World Wide Web*, 5(3):181–191.

[36] Choi, K.-S., Cho, W.-H., Lee, S., Lee, H., and Kim, C. (2004). The relationships among quality, value, satisfaction and behavioral intention in health care provider choice. *Journal of Business Research*, 57(8):913–921.

[37] Conner, M. and Norman, P. (2005). *Predicting health behaviour: Research and practice with social cognition models*. Open University Press., 2nd edition.

[38] Crowdflower (2016). 2015 new year's resolutions.

[39] Dai, H. K., Zhao, L., Nie, Z., Wen, J.-R., Wang, L., and Li, Y. (2006). Detecting online commercial intention (OCI). *Proceedings of the 15th international conference on World Wide Web*, pages 829–837.

[40] Daly, E. and Haahr, M. (2009). Social Network Analysis for Information Flow in Disconnected Delay-Tolerant MANETs. *IEEE Transactions on Mobile Computing*, 8(5):606–621.

[41] Dang, N. C., Moreno-García, M. N., and De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3):483.

[42] Das, A., Gollapudi, S., Kıcıman, E., and Varol, O. (2016). Information dissemination in heterogeneous-intent networks. In *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*, pages 259–268, New York, New York, USA. ACM Press.

[43] Devi, D. V. N., Kanth, T. V. R., Mounika, K., and Swathi, N. S. (2018). Assay: Hybrid approach for sentiment analysis. In *Information and Communication Technology for Intelligent Systems*, pages 309–318. Springer Singapore.

[44] Dickerson, J. P., Kagan, V., and Subrahmanian, V. S. (2014). Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 620–627.

[45] Ding, X., Liu, T., Duan, J., and Nie, J.-Y. (2015). Mining User Consumption Intention from Social Media Using Domain Adaptive Convolutional Neural Network. *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*, pages 2389–2395.

[46] Dooms, S., De Pessemier, T., and Martens, L. (2013). MovieTweetings: a movie rating dataset collected from twitter. *Workshop on Crowdsourcing and Human Computation for Recommender Systems, held in conjunction with the 7th ACM Conference on Recommender Systems*, (November 2015):2.

[47] Effrosynidis, D., Karasakalidis, A. I., Sylaios, G., and Arampatzis, A. (2022). The climate change twitter dataset. *Expert Systems with Applications*, 204:117541.

[48] Eight, F. (2019). Twitter us airline sentiment.

[49] Ellison, N. B., Steinfield, C., and Lampe, C. (2007). The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168.

[50] Elzinga, R., Reike, D., Negro, S. O., and Boon, W. P. (2020). Consumer acceptance of circular business models. *Journal of Cleaner Production*, 254:119988.

[51] Eshragh, F. and Mamdani, E. (1979). A general approach to linguistic approximation. *International Journal of Man-Machine Studies*, 11(4):501–519.

[52] Estivill-Castro, V., Lombardi, M., and Marani, A. (2018). Improving binary classification of web pages using an ensemble of feature selection algorithms. In *Proceedings of the Australasian Computer Science Week Multiconference*. ACM.

[53] Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *LREC 2006*, pages 417–422.

[54] Filho, J. A. B. L., Pasti, R., and de Castro, L. N. (2016). Gender classification of twitter data based on textual meta-attributes extraction. In *New advances in information systems and technologies*, pages 1025–1034. Springer.

[55] Fishbein, M. and Ajzen, I. (2011). *Predicting and changing behavior: The reasoned action approach*. Taylor & Francis.

[56] Friedkin, N. E. (2010). The attitude-behavior linkage in behavioral cascades. *Social Psychology Quarterly*, 73(2):196–213.

[57] Gautam, S. S., Abhishekh, and Singh, S. R. (2018). An improved-based TOPSIS method in interval-valued intuitionistic fuzzy environment. *Life Cycle Reliability and Safety Engineering*, 7(2):81–88.

[58] Gibert, K., Sànchez-Marrè, M., and Sevilla, B. (2012). Tools for Environmental Data Mining and Intelligent Decision Support. *Iemss.Org*.

[59] Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proc. AISTATS*, pages 249–256.

[60] Go, A. (2010). Exploiting the unique characteristics of tweets for sentiment analysis. *echnical report, Technical Report*.

[61] Go, A., Bhayani, R., and Huang, L. (2009a). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

[62] Go, A., Bhayani, R., and Huang, L. (2009b). Twitter Sentiment Classification using Machine Learning.

[63] Go, A., Bhayani, R., and Huang, L. (2010). Twitter Sentiment - A Sentiment Analysis Tool.

[64] Go, A., Huang, L., Bhayani, R., Sentiment, D., and Work, R. (2009c). Sentiment Analysis of Twitter Data. *Entropy*, 2009(June):30–38.

[65] Goh, E., Ritchie, B., and Wang, J. (2017). Non-compliance in national parks: An extension of the theory of planned behaviour model with pro-environmental values. *Tourism Management*, 59:123–127.

[66] Granitto, P. M., Furlanello, C., Biasioli, F., and Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90.

[67] Granovetter, M. (1973). The Strength of Weak Ties.

[68] Habib, M. W. and Sultani, Z. N. (2021). Twitter sentiment analysis using different machine learning and feature extraction techniques. *Al-Nahrain Journal of Science*, 24(3):50–54.

[69] Hájek, P. (1998). Metamathematics of fuzzy logic, vol. 4. *Trends in Logic*.

[70] Harper, F. M. and Konstan, J. A. (2015). The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1—-19:19.

[71] He, C., Parra, D., and Verbert, K. (2016). Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 56:9–27.

[72] He, R. and McAuley, J. (2016). Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. *Www*, pages 507–517.

[73] Horvitz, E., Breese, J., Heckerman, D., Hovel, D., and Rommelse, K. (1998). The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. *Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 256–265.

[74] Huang, Z., Jing, Z., Bai, Y., and Fang, Z. (2022). Does public environmental education and advocacy reinforce conservation behavior value in rural southwest china? *Sustainability*, 14(9):5505.

[75] Hutto, C. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

[76] Ibrahiem, S. S., Ismail, S. S., Bahnasy, K. A., and Aref, M. M. (2019). Multi-emotion classification evaluation via twitter. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE.

[77] Jefferson, C., Liu, H., and Cocea, M. (2017). Fuzzy approach for sentiment analysis. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE.

[78] Kaur, R. and Singh, S. (2016). A survey of data mining and social network analysis based anomaly detection techniques. *Egyptian Informatics Journal*, 17(2):199–216.

[79] Khan, F., Borah, S., and Pradhan, A. (2016). Mining Consumption Intent from Social Data : A Survey Mining Consumption Intent from Social Data : A Survey. *International Conference on Computing & Communication (ICCC-2016)*, (March).

[80] Khodabandelou, G., Hug, C., and Salinesi, C. (2015). *Mining Users' Intents from Logs*, volume 6.

[81] Kim, Z. M., Jeong, Y.-s., Hyeon, J., Oh, H., and Choi, H.-j. (2016). Classifying Travel-related Intents in Textual Data. *International Journal of Computing, Communication and Instrumentation Engineering*, 3(1):96–101.

[82] Kontopoulos, E., Berberidis, C., Dergiades, T., and Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, 40(10):4065–4074.

[83] Kumar, S., De, K., and Roy, P. P. (2020). Movie recommendation system using sentiment analysis from microblogging data. *IEEE Transactions on Computational Social Systems*, 7(4):915–923.

[84] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

[85] Lamsal, R. (2020). Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*, 51(5):2790–2804.

[86] Lee, H., Ferguson, P., O'Hare, N., Gurren, C., and Smeaton, A. F. (2010). Integrating interactivity into visualising sentiment analysis of blogs. *City*.

[87] Li, M. (2022). Application of sentence-level text analysis: The role of emotion in an experimental learning intervention. *Journal of Experimental Social Psychology*, 99:104278.

[88] Li, Q., Li, P., Ren, Z., Ren, P., and Chen, Z. (2022). Knowledge bridging for empathetic dialogue generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10993–11001.

[89] Lin, D. and Pantel, P. (2001). Discovery of Inference Rules for Question Answering. *Natural Language Engineering*, 7(4):343–360.

[90] Liu, B. (2015). *Sentiment Analysis*. Cambridge University Press.

[91] Longo, J. (2013). *Towards policy analysis 2.0*. phdthesis, University of Victoria.

[92] Lopez, C. E. and Gallemore, C. (2021). An augmented multilingual twitter dataset for studying the COVID-19 infodemic. *Social Network Analysis and Mining*, 11(1).

[93] Mamdani, E. and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*, 7(1):1 – 13.

[94] Mamdani, E. H. (1977). Application of fuzzy logic to approximate reasoning using linguistic synthesis. *IEEE transactions on computers*, (12):1182–1191.

[95] Martins, J., Costa, C., Oliveira, T., Gonçalves, R., and Branco, F. (2019). How smartphone advertising influences consumers' purchase intention. *Journal of Business Research*, 94:378–387.

[96] McEachan, R., Conner, M., and Lawton, R. (2005). A meta-analysis of theory of planned behavior studies: the impact of behavior type. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1):1–11.

[97] Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

[98] Mishael, Q., Ayesh, A., and Yevseyeva, I. (2019). Users intention based on twitter features using text analytics. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 121–128. Springer.

[99] Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

[100] Mohammad, S. M. and Bravo-Marquez, F. (2017). Emotion Intensities in Tweets. In *In Proceedings of the Joint Conference on Lexical and Computational Semantics (*Sem)*, pages 65–77, Vancouver, Canada.

[101] Mohammad, S. M., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. In *NAACL HLT 2018 - International Workshop on Semantic Evaluation, SemEval 2018 - Proceedings of the 12th Workshop*, pages 1–17, New Orleans, LA, USA.

[102] Mohammad, S. M. and 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*, number June, pages 26–34, California.

[103] Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., and Ureña-López, L. A. (2014). Ranked WordNet graph for Sentiment Polarity Classification in Twitter. *Computer Speech and Language*, 28(1):93–107.

[104] Montoro, A., Olivas, J. A., Peralta, A., Romero, F. P., and Serrano-Guerrero, J. (2018). An ANEW based fuzzy sentiment analysis model. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE.

[105] Nettleton, D. F. (2013). Data mining of social networks represented as graphs. *Computer Science Review*, 7(1):1–34.

[106] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM REVIEW*, 45:167–256.

[107] Nguyen, T.-T.-N., Nguyen, T. A., Do, S. T., and Nguyen, V. T. (2022). Assessing stakeholder behavioural intentions of BIM uses in vietnam's construction projects. *International Journal of Construction Management*, 23(13):2279–2287.

[108] Nielsen, A. and Nielsen, F. Å. (2011). A new ANEW : Evaluation of a word list for sentiment analysis in microblogs. In Rowe, M., Stankovic, M., Dadzie, A.-S., and Hardey, M., editors, *1st Workshop on Making Sense of Microposts*, volume 718, page 6. University of Denmark.

[109] Nugraha, E., Wibawa, A., Hakim, M., Kholifah, U., Dini, R., and Irwanto, M. (2019). Implementation of fuzzy tsukamoto method in decision support system of journal acceptance. In *Journal of Physics: Conference Series*, volume 1280, page 022031. IOP Publishing.

[110] Oh, K. J., Kim, Z., Oh, H., Lim, C. G., and Gweon, G. (2016). Travel intention-based attraction network for recommending travel destinations. *2016 International Conference on Big Data and Smart Computing, BigComp 2016*, pages 277–280.

[111] Palomino, M. A., Varma, A. P., Bedala, G. K., and Connelly, A. (2020). Investigating the lack of consensus among sentiment analysis tools. In Vetulani, Z., Paroubek, P., and Kubis, M., editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 58–72, Cham. Springer International Publishing.

[112] Park, D. H., Fang, Y., Liu, M., and Zhai, C. (2016). Mobile App Retrieval for Social Media Users via Inference of Implicit Intent in Social Media Text. pages 959–968.

[113] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic Inquiry and Word Count: LIWC 2001.

[114] Rane, A. and Kumar, A. (2018). Sentiment classification system of twitter data for us airline service analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 769–773. IEEE.

[115] Rangra, K. and Bansal, K. L. (2014). Comparative Study of Data Mining Tools. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(6):2277–128.

[116] Rashid, A., Shoaib, U., Shahzadsarfraz, M., and Technology, I. (2016). Knowledge discovery in database using intention mining. 28(6):5145–5151.

[117] Riza, L. S., Bergmeir, C. N., Herrera Triguero, F., Benítez Sánchez, J. M., et al. (2015). frbs: Fuzzy rule-based systems for classification and regression in r. American Statistical Association.

[118] Ross, T. J. (2009). *Fuzzy Logic with Engineering Applications*. John Wiley & Sons.

[119] Rout, J. K., Choo, K. K. R., Dash, A. K., Bakshi, S., Jena, S. K., and Williams, K. L. (2018). A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research*, 18(1):181–199.

[120] Saif, H., He, Y., Fernandez, M., and Alani, H. (2016). Contextual semantics for sentiment analysis of twitter. *Information Processing & Management*, 52(1):5–19.

[121] Salaheldeen, H. M. (2015). *Detecting , Modeling , and Predicting User Temporal Intention in Social Media*. PhD thesis.

[122] Sapountzi, A. (2016). *Insights from social networks: a big data analytics approach.* Msc, University of Macedonia.

[123] Shafer, J., Agrawal, R., and Mehta, M. (1996). Sprint: A scalable parallel classi er for data mining. In *Proc. 1996 Int. Conf. Very Large Data Bases*, pages 544–555.

[124] Subrahmanian, V., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., and Menczer, F. (2016). The darpa twitter bot challenge. *Computer*, 49(6):38–46.

[125] Tafreshi, S. (2021). *Cross-Genre, Cross-Lingual, and Low-Resource Emotion Classification*. PhD thesis, The George Washington University.

[126] Takagi, T. and Sugeno, M. (1983). Derivation of fuzzy control rules from human operator's control actions. *IFAC Proceedings Volumes*, 16(13):55–60.

[127] Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE transactions on systems, man, and cybernetics*, (1):116–132.

[128] Thelwall, M., Buckley, K., and Paltoglou, G. (2011). Sentiment Strength Detection for the Social Web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

[129] Thelwall, M., Buckley, K., Paltoglou, G., and Cai, D. (2010). Sentiment in Short Strength Detection Informal Text. *Journal of the American Society for Information Science*, 61(12):2544–2558.

[130] Tiwari, P., Bhat, A. K., and Tikoria, J. (2017). Predictors of social entrepreneurial intention: an empirical study. *South Asian Journal of Business Studies*, 6(1):53–79.

[131] Tusar, M. T. H. K. and Islam, M. T. (2021). A comparative study of sentiment analysis using nlp and different machine learning techniques on us airline twitter data. In *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, pages 1–4.

[132] Tyshchuk, Y. (2015). *Modeling human behavior in the context of social media during extreme events caused by natural hazards.* Rensselaer Polytechnic Institute.

[133] Tyshchuk, Y. and Wallace, W. A. (2018). Modeling human behavior on social media in response to significant events. *IEEE Transactions on Computational Social Systems*, 5(2):444–457.

[134] Vashishtha, S. and Susan, S. (2019). Fuzzy rule based unsupervised sentiment analysis from social media posts. *Expert Systems with Applications*, 138:112834.

[135] Vashisth, P. and Meehan, K. (2020). Gender classification using twitter text data. In *2020 31st Irish Signals and Systems Conference (ISSC)*, pages 1–6.

[136] Vastardis, N. and Yang, K. (2013). Mobile social networks: Architectures, social properties, and key research challenges. *IEEE Communications Surveys and Tutorials*, 15(3):1355–1371.

[137] Vineet, G., Devesh, V., Harsh, J., Deepam, K., and Shweta, K. (2014). Identifying purchase intent from social posts. *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014)*, pages 180–186.

[138] Wang, J., Cong, G., Zhao, X., and Li, X. (2015a). Mining User Intents in Twitter: A Semi-Supervised Approach to Inferring Intent Categories for Tweets. *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15)*, pages 339–345.

[139] Wang, Y., Yin, G., Cai, Z., Dong, Y., and Dong, H. (2015b). A trust-based probabilistic recommendation model for social networks. *Journal of Network and Computer Applications*, 55:59–67.

[140] Warriner, A. B., Kuperman, V., and Brysbaret, M. (2013). Norms of valence, arousal, and dominance for 13,915 Enlish lemmas. *Behavior Research Methods*, 45(4):1191–1207.

[141] Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*. Association for Computational Linguistics.

[142] Wu, S., Zhang, Q., Chen, W., Liu, J., and Liiu, L. (2019). Research on trend prediction of internet user intention understanding and public intelligence mining based on fractional differential method. *Chaos, Solitons & Fractals*, 128:331–338.

[143] Yadollahi, A., Shahraki, A. G., and Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Comput. Surv.*, 50(2).

[144] Yang, Y. and Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, pages 412–420.

[145] Yin, J., Liu, C., Wang, W., Sun, J., and Hoi, S. C. (2020). Learning transferrable parameters for long-tailed sequential user behavior modeling. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '20, page 359–367, New York, NY, USA. Association for Computing Machinery.

[146] Žabkar, V., Brenčič, M. M., and Dmitrović, T. (2010). Modelling perceived quality, visitor satisfaction and behavioural intentions at the destination level. *Tourism Management*, 31(4):537–546.

[147] Zadeh, L. A. (2015). Fuzzy logic—a personal perspective. *Fuzzy Sets and Systems*, 281:4 – 20. Special Issue Celebrating the 50th Anniversary of Fuzzy Sets.

[148] Zhang, C., Fan, W., Du, N., and Yu, P. S. (2016). Mining User Intentions from Medical Queries : A Neural Network Based Heterogeneous Jointly Modeling Approach. *Proceedings of the 25th International Conference on World Wide Web*, pages 1373–1383.

[149] Zhao, X. W., Guo, Y., He, Y., Jiang, H., Wu, Y., and Li, X. (2014). We Know What You Want to Buy: A Demographic-based System for Product Recommendation on Microblogs. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1935–1944.

Appendix A

Appendix A

A.1 Online Social Network

Online social networks have become an integral part of modern society, providing a platform for individuals to connect and share information. This section explores the various aspects of online social networks in relation to intention mining. The section begins by examining the different types of online social network platforms, such as Facebook, Twitter and LinkedIn, and how they are used by individuals to connect and share information. This will be followed by an in-depth examination of social network analysis, which is the process of using statistical and computational methods to understand the structure and dynamics of social networks. This will include a discussion of the various metrics and techniques used to analyse social networks, such as centrality measures, community detection, and network visualization. This section aims to provide a comprehensive overview of the current state-of-the-art in online social network research and its relevance to the task of intention mining.

A.1.1 Online Social Network Platforms

Online social networks platforms, such as Facebook, Twitter and LinkedIn, have become a popular means for individuals to connect and share information. These platforms often have user-friendly interfaces that allow for easy access via different technologies, including mobile devices. This section will examine the different types of online social network platforms and how they can be used as a source of input for intention mining algorithms.

Social Networking Communities Platforms

A social network community platform is an online platform that enables individuals and organizations to connect and interact with one another. These platforms allow users to create a personal or professional profile, share information and content, and establish connections with other users. Social network community platforms have become an integral part of our daily lives, and their popularity has grown rapidly in recent years.

One of the main features of social network community platforms is the ability for users to create a personal or professional profile. These profiles allow users to share information about themselves, including their interests, education, and employment history. This information can be used by other users to connect with and learn more about the person. Additionally, users can share content such as photos, videos, and links on their profile, which can be viewed by their connections.

Another key feature of social network community platforms is the ability for users to establish connections with other users. This is often referred to as "friending" or "following" on different platforms. These connections allow users to interact with one another by sending messages, commenting on posts, and sharing content. Furthermore, social network community platforms allow users to join groups, which are communities of users with similar interests. These groups can be used to share information, ask questions, and connect with like-minded individuals.

Additionally, the social network community platforms provide a set of functionalities that are tailored to the type of the platform, for instance, LinkedIn is a platform specifically designed for professionals and it helps individuals and organizations to stay in touch, while Facebook is one of the most popular social networking services that support various usage patterns and technological capabilities that support both online and offline connections.

- **Facebook[1]:** Facebook is a social networking community platform that allows users to connect and interact with each other online. It is one of the most widely used social networking services and is considered a rich source of information for social network research due to its vast range of functions and services [49]. It supports a wide range of usage patterns, from online

[1] https://www.facebook.com/

and offline connections, which makes it a valuable tool for studying human behaviour and intentions.

Facebook website and mobile applications provide users with a wide range of services and functionalities, such as open platform with APIs that allows third-party providers to create applications that enhance the original Facebook services. For instance, location-aware services can be accessed through mobile Facebook applications, enabling users to update their geographic status, browse the current locations of their Facebook friends, and sort friends by distance from the users' current location. This makes Facebook a valuable source of data for intention mining and an important platform to be studied in the field of online social networks.

- **LinkedIn [2]:** LinkedIn is a social networking community platform that is specifically designed for professionals. It allows individuals and organizations to connect and interact with one another online, fostering professional relationships and networking opportunities. LinkedIn is also a powerful tool for job seekers and hiring managers, as it facilitates the process of finding and listing job vacancies. For job seekers, LinkedIn allows them to create a professional profile and connect with potential employers and recruiters. This gives job seekers access to a wide range of job opportunities that may not be available on other job search platforms. For organizations, LinkedIn is an effective recruitment tool, as it allows them to list job vacancies and connect with potential candidates. Furthermore, LinkedIn has a feature of recommendation system, where current and former colleagues, supervisors and partners can endorse the candidate's skills and qualifications, which can make the recruitment process more efficient and effective. This makes LinkedIn a valuable source of data for intention mining and an important platform to be studied in the field of online professional networks.

In conclusion, social network community platforms have become an essential aspect of our digital lives. They provide a wide range of functionalities and services, which makes them valuable sources of data for intention mining. They allow individuals and organizations to connect and interact with one another, fostering professional and personal relationships and networking opportunities. Furthermore, the different platforms have different features that make them

[2] https://www.linkedin.com

suitable for different purposes, this makes it important to study the different platforms in the field of online social networks.

A.1.2 Microblogging Platforms

Microblogging platforms are a type of social media that allows users to share short-form text and multimedia content, such as text messages, images, and videos. These platforms are characterized by their limited character count, which typically ranges from 280 characters (Twitter) to 500 characters (Tumblr). This feature makes them particularly well-suited for sharing quick updates and real-time information.

One of the key features of microblogging platforms is the ability for users to share real-time information. This can include personal updates, news, and current events. Additionally, users can share multimedia content such as photos and videos, which can be used to provide context and visual interest to their posts. This feature makes these platforms ideal for sharing information about events, such as concerts, conferences, and other gatherings.

Another key feature of microblogging platforms is the ability for users to follow other users and see their updates in a single feed. This feature allows users to keep up-to-date with the latest information from their friends, family, and other people they follow. Additionally, users can discover new content by searching for specific keywords or hashtags.

In addition to the above, microblogging platforms have different types, for example, Twitter is one of the most popular microblogging platforms in the world, it allows users to post 280 characters, while Sina Weibo is a Chinese microblogging platform that allows users to post up to 500 characters, and Tumbler is a microblogging platform that allows users to post text, photos, videos and audio to a short-form blog, also it allows users to follow other blogs.

- **Twitter[3]:** Twitter is a microblogging social platform that allows users to share and interact with short messages known as tweets. Since its launch in 2006, Twitter has grown to become one of the largest and most widely-used social media platforms in the world, with over 330 million monthly active users. Twitter is widely used for a range of purposes, from personal communication and news sharing to business marketing

[3] https://www.twitter.com/

and political campaigning. Twitter users can follow other users in Twitter or can be followed with no need for any reciprocation. Twitter users get information about "what are you doing or thinking" as tweets of their Twitter friends in real time [84]. One of the differences between Facebook and Twitter is that, Facebook used to help users to interact and communicate with their friends and family in the real world, while Twitter helps users to communicate with any person that share the same interest. As in Facebook, Twitter has both website and a related mobile application, and associated APIs that support applications' programmers and enable them to develop new functions and services for Twitter. Moreover, Twitter enables mobile device users to send new tweets to the Twitter web-site not only through the mobile application, but also by short messaging service (SMS). Twitter's short message format, combined with the platform's ability to reach a large and diverse audience, has made it a valuable tool for studying human behaviour and intentions. The tweets shared on Twitter can contain a wealth of information about users' thoughts, opinions, and behaviours, which can be analysed using a range of text mining and natural language processing techniques. This makes Twitter a valuable source of data for intention mining, and an important platform to be studied in the field of online social networks.

- **Tumblr[4]:** Tumblr is a microblogging social platform that allows users to share and interact with a wide range of multimedia content, including text, photos, videos, and GIFs. The platform was launched in 2007 and has since grown to become one of the largest and most popular microblogging platforms in the world, with over 500 million active blogs. Tumblr is particularly popular among younger users and has been used for a range of purposes, from personal blogging and creative expression to business marketing and political activism. The platform's multimedia-rich content and its ability to reach a large and diverse audience make it a valuable tool for studying human behavior and intentions. The microblogs shared on Tumblr can contain a wealth of information about users' thoughts, opinions, and behaviors, which can be analyzed using a range of text mining, natural language processing and multimedia analysis techniques. This makes

[4] https://www.tumblr.com/

Tumblr a valuable source of data for intention mining, and an important platform to be studied in the field of online social networks. [3].

- **Sina Weibo[5]:** Weibo stands for "microblog" in Chinese words[6]. Sina Wibo is widely used in China and launched in 2009. It is one of the largest microblogging services in China with over 500 million monthly active users[149]. It implements many features from Twitter. The users posts are limited to 140 charterers. It also supports double hash-tagging "#HashName# " method. Sina Weibo is widely used for a range of purposes, from personal communication, news sharing, to business marketing and political campaigning. The platform's short message format, combined with its ability to reach a large and diverse audience in China, has made it a valuable tool for studying human behavior and intentions within China. The microblogs shared on Sina Weibo can contain a wealth of information about users' thoughts, opinions, and behaviors, which can be analyzed using a range of text mining and natural language processing techniques. This makes Sina Weibo a valuable source of data for intention mining within China, and an important platform to be studied in the field of online social networks within the country.

In conclusion, microblogging platforms are a popular form of social media that allows users to share short-form text and multimedia content. They are particularly well-suited for sharing real-time information and updates, and the ability to follow other users' updates in a single feed makes them ideal for keeping up-to-date with the latest information. Additionally, microblogging platforms come in different types each of them has its own features, and the ability to share multimedia content provides a rich source of data for intention mining.

A.1.3 Multimedia Social Platform

In recent years, multimedia social platforms such as TikTok, Flickr and Instagram have grown in popularity and have become an integral part of modern society. These platforms allow users to share and interact with a wide range of multimedia content, including photos and videos. As a picture is worth a thousand words, an increasing number of social media tools are using photos to share experience

[5] https://www.weibo.com/
[6] https://en.wikipedia.org/wiki/Microblogging_in_China/

or to tell stories. Audio is used in social networks, as it consider easier to understand than text. Videos can give more information about body language that is used during communication. This section explores the various aspects of multimedia social platforms in relation to intention mining. The section examines the different types of multimedia social platforms, such as TikTok, Flickr and Instagram, and how they are used by individuals to share multimedia content and connect with others. It also covers the unique features and functionalities of these platforms that make them valuable sources of data for intention mining.

- **Instagram [7]:** Instagram is a multimedia social platform that allows users to share photos and videos, as well as to view and interact with content shared by other users. Since its launch in 2010, Instagram has grown to become one of the largest and most popular social media platforms in the world, with over a billion monthly active users. The platform is particularly popular among younger users and has been used for a range of purposes, from personal photo sharing to business marketing. Instagram provides a range of features that allow users to edit and enhance their photos and videos, as well as to add captions, hashtags, and location information. The platform also allows users to interact with other users through features such as direct messaging, commenting, and liking. This makes Instagram a valuable source of data for intention mining, as the photos and videos shared on the platform can contain a wealth of information about the intentions and behaviors of users.

- **Flickr [8]:** Flickr is a multimedia social platform that allows users to share and organize their photos and videos. The platform was launched in 2004 and has since grown to become one of the largest online photo management and sharing platforms in the world. Flickr provides users with a range of tools for organizing and sharing their photos, including the ability to add tags and descriptions, create albums, and set privacy settings. In addition to personal photos and videos, Flickr is also used by professional photographers, businesses, and organizations to share and showcase their work. The platform has a large and active user community, which has led to the development of a wide range of third-party tools and applications

[7] https://instagram.com/
[8] https://www.flickr.com/

for accessing and analysing the data on the platform. This makes Flickr a valuable source of data for intention mining, as the photos and videos shared on the platform can contain a wealth of information about the intentions and behaviours of users.

- **TikTok:**[9] TikTok is a relatively new multimedia social platform that has gained significant popularity in recent years. It allows users to create and share short videos, often set to music, with a wide variety of creative filters and editing tools. The platform is particularly popular among younger users and has been used for a range of purposes, from comedic skits and lip-syncing to more serious content such as news and politics. TikTok has also been used as a means for businesses and influencers to promote their products and services. As a multimedia platform, TikTok provides a unique source of data for intention mining, as the videos shared on the platform can contain a wealth of information about the intentions and behaviours of users.

In conclusion, Multimedia social platforms are a type of social media that allow users to share and consume a wide range of multimedia content, including photos, videos, and audio. They are characterized by their ability to support a wide range of multimedia formats, and often include features such as image and video editing tools, and the ability to share content across multiple platforms. Additionally, multimedia social platforms come in different types each of them has its own features, and the ability to share multimedia content provides a rich source of data for intention mining.

A.1.4 Social Network Analysis

Social Network Analysis (SNA) is the procedures of studying, describing, and understanding the social network structure to make decisions accordingly. Structured Data in a Social network is represented by a graph as $G = (V, E)$, where V is a set of nodes or entities ; people, organizations, systems, and products; while E is a set of edges or relationships that connects the nodes through patterns of interactions. SNA may belong to one of two approaches, either a static or a dynamic analysis. The static assumes that changes on a social network is limited over time and analysis on the entire network can be done in once[4].

[9] https://www.tiktok.com/

The dynamic SNA is presented in two approaches the centralized and the distributed approaches. The two approaches differ mainly in how researchers see the network.

In the centralized approach, the social graph is considered as a whole and the assumption is made that the total amount of the social information is available to all nodes. While in the distributed approach, a more detailed implementation is used. Each node is considered to have a limited access to the social information, thus being able to see only a subset [122].

Many researchers[136, 105, 1, 5, 45, 78] have used graph theory in their researches to analyse social networks. Graph theory provided many tools for network analysing and manipulation to solve the related problems.

The process of SNA begins by building a graph for a social network. The graph is built from the generated graph concepts such as social ties, neighbours and communities need to be address. After that, network analysis can be performed on metrics such as centrality, edge expansion and clustering coefficient.
Social network matrices can be summarised as follows:

- **Social Ties:** are the meaningful social relationships such as friendship. Social ties strength was introduced and quantified by [67]. The author defined as ″*a (probably linear)* which combines several factors such as (the time as amount, the intensity of emotion, the intimacy (mutual confiding) and the reciprocal services which characterize the tie″). These values are used to represent the edge weight between two nodes in a graph , which specify the tie strength.

- **Social Neighbours:** are the set of nodes with which a host is related through specific social ties, where these ties would be preferably strong. Social neighbours are categorized according to the strength of the social ties between them, Community Member, Familiar Stranger, Stranger, and Friend.

- **Communities:** A community is ″ *a clustering of entities that are closely linked to each other, either by direct linkage or by some easily accessible entities that can act as intermediates* ″[26]. The creation of communities would be case by different reasons such as social relationships, common interests, family bonds, or work relationships between individuals. Applying this property in SN leads to improve scalability and optimize routing [136].

- **Centrality:** According to the graph theory, centrality is the ″ centrality represents the value of the relative importance of a vertex within a graph″ [40]. It is useful tool for categorizing the individuals inside the examined social network. For example, the socially most active member of a group has the highest centrality, because he or she is the one who brings all the other friends together. The most popular ways to measure centrality are three, which called Freeman degree, closeness, and betweenness centrality metrics. Degree centrality is the number of direct ties involving a given node, while closeness centrality is the reciprocal of the mean geodesic distance, which is the shortest path between a node and all other reachable nodes with in network.

- **Betweenness Centrality:** It represents the relative importance of a node's location and its ties. It was used in community detection algorithms. The definition was extended by Newman et. al[106] to include both links and the computed link betweenness to find the most important links in a network. They applied removing the links one by one, which made the network to be parsed into distinctive communities[136].

- **Edge Expansion:** The worst-case exit capacity from a set of nodes, in proportion to its down-link update traffic. It captures the rate at which information is flooded through a network.

- **Clustering Coefficient:** This factor represents the adjacency between the vertices. In other words, it describes how well is the connectivity and how well identifiable the communities in a network are.

- **Selfishness:** It consider a negative metrics the represent all the actions that aim for the maximization of personal profit, against the common good. An example of a selfish behaviour could a node not accepting to act as a relay for a message, in order to maximize its battery life span.

A.2 Overview of Data Mining Tools

Data mining tools refer to a set of software applications and algorithms designed to extract and process data from large datasets. They are used in research to identify patterns, relationships, and trends in the data, which can be used to make predictions, make decisions, and drive business insights. In this research, data mining tools can be used to process and analyse the social network dataset in order to build the intention model. These tools apply different techniques and algorithms, such as clustering, association rule mining, decision trees, and artificial neural networks, to extract and pre-process the data, in order to uncover meaningful insights.

Different data mining tools such as KNIME, WEKA, Orange and RapidMiner are used for data pre-processing, data exploration, data visualization, modelling and evaluation. They provide various techniques and algorithms for extracting insights from large datasets and help researchers in various domains to perform various data mining tasks such as regression, classification, clustering, association rule mining, and more. These tools also provide a user-friendly interface and easy-to-use environment for data scientists, data analysts, and researchers to perform their data mining tasks efficiently. The follows a brief description of different tools that are used in data mining.

- **KNIME** [10] (Konstanz Information Miner) is an open source workflow data mining platform based on Java and Eclipse platform. It works under different operating systems Windows, Linux, OS X [115]. KNIME. In addition, it supports big data analysis with graphical user interface. Its visual interface gives the ability to access data, applies data transformation, supports powerful predictive analytics [58]. KNIME workflow consists of connected nodes or extensions[14]. Moreover, KNIME supports integration of different data analytic tools such as R, Python scripting, WEKA and other third party applications such as Google Analytics.

  - Pros:

    * Open-source and free: KNIME is an open-source data mining tool, which means it is free for use and distribution. This makes it an ideal choice for researchers who have limited budgets.
    * Versatile and comprehensive: KNIME supports a wide range of machine learning algorithms and data pre-processing techniques, including data cleaning, feature selection, and data normalization. This makes it an ideal tool for researchers who require a comprehensive suite of data mining tools.

[10]https://www.knime.org

    * User-friendly interface: KNIME has a graphical user interface that is intuitive and easy to use, making it ideal for researchers who are new to data mining.
    * Easy integration: KNIME provides a wide range of plugins that enable researchers to integrate other data mining tools, such as WEKA, Orange, and RapidMiner. This makes it an ideal choice for researchers who want to use multiple tools for their research.
  - Cons:
    * Steep learning curve: Despite its user-friendly interface, the learning curve for KNIME can be steep, especially for researchers who are new to data mining. This can make it difficult to get started quickly with the tool.
    * Limited scalability: KNIME is not designed to handle large-scale data processing and analysis, which makes it unsuitable for research projects that require high-performance computing resources.
    * Limited support: As an open-source tool, KNIME has limited support resources compared to commercial data mining tools. This can make it difficult for researchers to get help and support when they need it.

  Overall, KNIME is an excellent data mining tool that is well suited for researchers who require a comprehensive suite of data analysis and reporting functionalities in a single platform. Despite its limitations, the open-source and free nature of the tool, as well as its versatility and user-friendly interface, make it an ideal choice for researchers who are working on a limited budget.

- **WEKA** [11] stands for (Waikato Environment for Knowledge Analysis), it is free data mining tool based on Java, making it accessible to researchers without incurring any licensing costs. It works on different operating system platforms. It combines several tools of data pre-processing, machine learning, visualization, and feature selection.
  - Pros:

[11] https://www.cs.waikato.ac.nz/ml/weka/downloading.html

    * Wide Range of Algorithms: WEKA offers a variety of algorithms for data mining tasks including classification, regression, clustering, and association rule learning, among others.
    * User-friendly Interface: WEKA has a graphical user interface (GUI) that makes it easy to use, even for researchers who are not experts in programming.
    * Easy Integration with Other Tools: WEKA can be easily integrated with other data analysis tools, making it a useful addition to the researcher's toolkit.
    * Strong Community Support: The open source nature of WEKA has resulted in a strong community of developers and users, who continuously contribute to its development and provide support to each other.
  - Cons:
    * Limited Data Handling Capabilities: WEKA may struggle to handle large datasets and may require additional software or tools to manage and process the data.
    * Outdated Algorithms: WEKA may contain older algorithms that may not be as effective as newer, more advanced algorithms available in other data mining tools.
    * Limited Customization Options: WEKA is a pre-packaged tool and may not offer the level of customization that some researchers require to meet the specific needs of their research.
    * Slower Processing Speed: WEKA is written in Java, which can be slower than other programming languages and may result in slower processing times for large datasets.

- **Orange** [12] is an open-source data-mining tool that implemented in C++ and Python. This tool supports different operating systems. It provides a comprehensive set of tools for data preprocessing, visualization, and analysis. It has a user-friendly interface, making it suitable for both beginners and advanced users. Some of the key features of Orange include:
  - Pros:

[12] https://orange.biolab.si

    * User-friendly interface: Orange has a user-friendly interface that allows for easy exploration and manipulation of data.
    * Data visualization: The tool provides a range of visualization options, allowing for a better understanding of data patterns and relationships.
    * Integration with other tools: Orange is compatible with other data mining and machine learning tools, providing the ability to use them in combination with Orange.
    * Large community: Orange has a large and active community of users who can provide support and share knowledge.
  - Cons:
    * Limited scalability: Orange is not designed to handle large datasets and can become slow or unresponsive when working with large data sets.
    * Limited model selection: The tool provides a limited range of machine learning algorithms and may not be suitable for more advanced or specialized research.
    * Lack of customization: Orange may not be suitable for users who require custom algorithms or workflows.

  In summary, Orange is a user-friendly and accessible data mining tool that provides a range of tools for data exploration and analysis. However, its limited scalability and lack of customization may make it less suitable for more advanced or specialized research.

- **RapidMiner:** [13] is a stand-alone application with user friendly interface that support variety operating systems. Furthermore, it offers a comprehensive range of data preprocessing, visualization, modelling, evaluation, and deployment tools. It is widely used in various industries such as finance, healthcare, marketing, and more for tasks such as predictive modelling, customer segmentation, market basket analysis, and more.
  - Pros:

[13] https://rapidminer.com

    * User-friendly interface: RapidMiner has a drag-and-drop interface that is easy to use, even for people with little to no programming experience.
    * Comprehensive features: RapidMiner offers a wide range of features, including data preprocessing, feature selection, modelling, evaluation, and deployment, making it a one-stop-shop for data mining tasks.
    * Integration with various data sources: RapidMiner can be integrated with various data sources, including databases, spreadsheets, and web services, making it easier to import data into the platform.
    * Large community: RapidMiner has a large and active user community, which means that users can find help and support easily, and also contribute to the platform through add-ons and plugins.
  - Cons:
    * Steep learning curve: Although the interface is user-friendly, the platform is complex and may take some time to get used to for first-time users.
    * Resource-intensive: RapidMiner can be resource-intensive and may require a powerful computer to run smoothly, especially when working with large datasets.
    * Cost: RapidMiner is a commercial software, and its price may be a barrier for some users, especially for those with limited budgets.
    * Limited scalability: RapidMiner may struggle with scalability when working with extremely large datasets, which can lead to slow performance and decreased efficiency.

  In conclusion, RapidMiner is a powerful and comprehensive data mining tool that offers a wide range of features for data preprocessing, visualization, modelling, evaluation, and deployment. However, its cost and resource requirements may limit its accessibility for some users, especially for those with limited budgets or working with extremely large datasets.

A critical comparison of KNIME, WEKA, Orange, and RapidMiner can be made based on several factors such as functionality, user-friendliness, scalability, integration capabilities, and cost. KNIME, known as a graphical user interface (GUI) based platform, is popular for its ease of use and visualization capabilities. It provides a wide range of data analysis and data mining tools, including machine learning algorithms, data preprocessing, and reporting functionalities. However, it has limited scalability and can slow down when working with large datasets.

WEKA, on the other hand, is a well-established data mining toolkit that provides a comprehensive set of machine learning algorithms. It has a simple and intuitive user interface that makes it easy for beginners to get started with data mining. However, it may not be as flexible and customizable as other tools.

Orange, another GUI-based tool, is known for its visual programming approach, which allows for easy customization and integration with other tools. It provides a wide range of machine learning algorithms and data preprocessing functionalities. However, it may not be as scalable as other data mining tools.

RapidMiner is a powerful and scalable data mining platform that provides an extensive set of machine learning algorithms and data analysis tools. It is widely used in industry and academia due to its robustness and scalability. However, its complex user interface may make it challenging for beginners to get started with data mining.

Furthermore, each tool has different strengths and weaknesses, therefore, selecting a tool for a particular research project will depend on the specific requirements and goals of the project. The table A.1 provides a high-level overview of some of the key advantages and disadvantages of each tool, and researchers should evaluate each tool carefully based on their individual needs. KNIME is a comprehensive and versatile data mining tool that provides a user-friendly interface and a wide range of data analysis functions. Its open-source nature makes it easily accessible for researchers, and its wide range of plugins and extensions allows for seamless integration with other tools. KNIME also offers an extensive library of nodes and workflows that makes it possible to automate complex data analysis processes and simplify the implementation of various data mining techniques. Furthermore, its visual programming approach allows researchers to easily understand and interpret the results of their data analysis. Given these strengths, KNIME is an excellent choice for researchers who are looking for a comprehensive and flexible data mining tool.

Tool	Operating System	Supported Languages	Features	Related papers
KNIME	Cross Platform	Java, R, Python	Easy to use graphical interface, ability to integrate multiple data sources, large community support	[70]
WEKA	Cross Platform	Java	Wide range of machine learning algorithms, simple and easy to use interface, large community support	[25]
Orange	Cross Platform	Python	Interactive data visualization, integration with scikit-learn and numpy libraries, ability to handle big data	[46]
RapidMiner	Cross Platform	Java	Advanced analytics and data visualization, ability to handle big data, seamless integration with big data tools like Hadoop and Spark	[71]

Table A.1 Data Mining Tools

A.3 Tweet Object Description

The tweet object is a fundamental unit of information on the Twitter platform. It encapsulates a single tweet, consisting of various fields that describe different aspects of the tweet. The Data dictionary: Standard v1.1 provides a comprehensive guide to the fields that are included in a tweet object, as well as their possible values and data types. Let's take a closer look at what this entails.

The created_at field indicates the date and time when the tweet was created. The id field provides a unique identifier for the tweet as an integer. The id_str field provides a unique identifier for the tweet as a string. The full_text field contains the complete text of the tweet. The truncated field is a Boolean value indicating whether the tweet has been truncated to fit within Twitter's character limit. The display_text_range field provides the start and end indices of the displayed text in the full text of the tweet. The entities field contains information about any entities (e.g., hashtags, URLs, user mentions) present in the tweet. The source field indicates the source or platform used to create the tweet. The in_reply_to_status_id field indicates the ID of the tweet that this tweet is in reply to, if any. The in_reply_to_user_id field indicates the ID of the user that this tweet is in reply to, if any. The in_reply_to_screen_name field indicates the screen name of the user that this tweet is in reply to, if any. The user field contains information about the user who created the tweet, including their ID, name, screen name, location, description, follower count, friend count, and more. The geo field indicates the geographic location of the tweet, if available. The coordinates field provides the coordinates of the tweet, if available. The place field indicates the place where the tweet was sent, if available. The is_quote_status field is a Boolean value indicating whether the tweet is a quote tweet. The retweet_count field indicates the number of times the tweet has been retweeted. The favorite_count field indicates the number of times the tweet has been favorited. The favorited field is a Boolean value indicating whether the tweet has been favorited by the authenticated user. The retweeted field is a Boolean value indicating whether the tweet has been retweeted by the authenticated user. The lang field indicates the language of the tweet. The Data dictionary: Premium v1.1 includes many of the same fields as the Standard v1.1, but it also provides additional fields that are specific to the premium search API, which allows developers to search for historical tweets using a variety of filters and criteria. Some of the additional fields included in the Premium v1.1 data dictionary include:

The matching_rules field, which provides an array of rules that matched the query used to search for the tweet. The withheld_in_countries field, which provides an array of country codes where the tweet is withheld due to local laws or regulations. In summary, the tweet object is a complex data structure that encapsulates various fields that describe different aspects of a tweet. The Standard v1.1 data dictionary provides a comprehensive guide to the fields that are included in a tweet object, while the Premium v1.1 data dictionary includes additional fields that are specific to the premium search API. Understanding these data dictionaries is essential for developers who are building applications that work with tweet data, whether in real-time or historically.