1.1 Sentiment Analysis

Sentiment analysis is defined as the computational study of identifying opinions, emotions and sentiments of users about specific entities. The entities may be an event, product, topic, services, individuals, attitudes, organizations etc. Sentiments are the key influencers of human behaviour. Human beliefs and perceptions of reality relays on how others see the world. Sentiment analysis is otherwise known as Opinion mining. It is the process to evaluate the text messages to determine whether the expression is positive, negative or neutral and to what extent. It plays a key role in the decision-making process of users and stakeholders. This is widely applied to healthcare materials as well as to the voice of the customer community such as social media, reviews, survey responses etc for the applications ranging from clinical medicine to marketing. Sentiment analysis is widely used to arrive at a decision as to determine users likes or dislikes about the product or if somebody is for or against something.

The use of sentiment analysis is frequently applied in social media to assist the marketing and customer service units to identify the sentiments of consumers. In marketing, sentiment analysis can be of great help to understand whether the consumers are satisfied or dissatisfied with the product or services. Similarly, industries use sentiment analysis to evaluate the impact of a new service or product.
advertisement campaign, user’s opinion towards recent promotional news of a company on social media. Sentiment analysis is also useful for the customer service agents to automatically sort user emails into urgent or spam categories based on the sentiment of the email.

The sentiment analysis task is critical since it allows the users to see the opinion of customers such as their likes and dislikes about the particular brand. The customer feedback, reviews from social media includes a treasure trove of practical business information. It is not always easy to identify the relevant website and to decipher the information so as to summarize the opinions in them. Sentiment analysis is, therefore, a vital tool to uncover those sentiments.

The main challenge of sentiment analysis is to identify the most relevant social media site to extract the user opinion. The extracted raw data may be noisy and it requires carrying out appropriate filtering tasks to prepare out the data for further analysis. Another key challenge is to identify the most suitable sentiment analysis task for the accurate prediction of sentiments.

![Fig 1.1: Steps in Sentiment Analysis Tasks](image)
The structure of sentiment analysis process is depicted in Fig 1.1 above. The raw data can be extracted from various social media sources such as blogs, discussion forums, micro blogs, Twitter, reviews and forums etc. The data thus gathered may be huge and unstructured therefore it is required to clean those data using various preprocessing tasks. The data preprocessing is nothing but a data cleaning process to structure the data before analysis. This involves stemming, punctuation removal, stop words removal etc. This is common to all text analysis problems. Furthermore, an efficient sentiment analysis technique is required to analyze and predict the sentiments. Sentiment analysis prediction is the measurement of positive, negative and neutral opinions. The main goal of sentiment analysis is to convert the unstructured data into meaningful information.

1.2 Sentiment Analysis Techniques

The sentiment analysis techniques are broadly classified into two categories; lexicon-based techniques and machine learning techniques. Lexicon-based techniques are based on rules and lexica whereas machine learning techniques use supervised and unsupervised learning. A detailed description of the two categories of sentiment analysis techniques are given below.

1.2.1 Lexicon-based Techniques

These techniques employ dictionaries of words and the sentiments are based on the analysis of individual phrases or words. This is then used to calculate the sentiment of the document. There are two common methods to create sentiment lexicons, they are corpus-based approach and dictionary-based approach. Both dictionary-based and corpus-based approaches use either semantic or statistical methods to determine sentiment polarity.

1.2.2 Machine Learning Techniques

In machine learning process, data plays an essential role and a suitable learning method called the machine learning technique is necessary to uncover hidden insights from the data. Due to the huge availability of data and the intrinsic ability to uncover the information from data, the machine learning techniques pulled out much more attention in the areas of data mining, information retrieval and pattern recognition.
While performing machine learning task, two important phases are to be considered:

- **Training phase**: In this phase, machine learning model will be trained based on the data provided, by that way pairing input with the expected result.

- **Test phase**: This phase estimates how well the machine learning model has been trained. It also estimates the model properties such as mean error, classification errors, recall, precision etc.

There are generally two commonly used techniques in machine learning based on the nature of dataset and the ongoing problem. They are called supervised learning techniques and unsupervised learning techniques.

### 1.2.2.1 Supervised Learning:

Supervised learning is a machine learning task of inferring a relation from the labelled training dataset. Training dataset is a set of labelled training examples. In this learning, each training example is a pair of input and the desired output value. These learning techniques analyze the training data and develop an inferred relation to map new examples. The information extracted from supervised learning technique is frequently being utilized for both prediction and recognition. Supervised learning phase is again categorized into classification and regression methods.

- **Classification**

Classification is one of the most commonly used techniques in machine learning with a wide range of applications including medical diagnosis, sentiment analysis etc. The classification models attempt to draw conclusions from observed values. In this task, the inputs are divided into multiple classes and the learner must build a model that assigns unseen input values to the classes. The most basic form of classification is a binary classifier. It has only a single output with two labels 0 and 1 respectively. The commonly used classification models include SVM, ID3, and Naïve Bayes etc.
Regression

Regression is a supervised learning technique used to predict the dependent variable, given a set of independent variables. It is parametric in nature because it makes certain assumptions based on the data set. While classification gives an outcome of “what kind”, regression gives the outcome of “how many”. The most popular regression algorithms are namely Linear Regression, Multivariate Adaptive Regression Splines, and Stepwise Regression etc.

1.2.2.2 Unsupervised Learning:

Unsupervised learning technique is tasked with finding relationships within data. There are no training examples used in this learning process. Instead, the model is given a set of data and tasked with detecting the patterns and correlations therein. The algorithms used to perform this are very different from those used for supervised learning. The popular examples of unsupervised learning methods are K-Means, Apriori algorithm etc. Unsupervised learning techniques are further grouped into clustering and association methods.

- Clustering: - Clustering is a problem where the inherent grouping of the data needs to be discovered. It is the task of dividing the data points into a different number of groups such that the groups are segregated with similar traits and are assigned them into different clusters. Clustering techniques are of different types such as partition-based, hierarchical-based and density-based clustering techniques etc.

- Association: - Association is a rule learning problem where the rules are discovered that describes large portions of the given data. Association rules are the antecedent and consequent statements that help unveil the connection between apparently unrelated data in the information repository or database. The antecedent is the item found in the data and consequent is the item found in combination with the antecedent.
1.3 Sentiment Analysis Applications

The sentiments are the key influencers of human behaviour. Sentiment analysis is essential to make timely decisions with automated insights based on the opinion of users collected from social media. The applications of sentiment analysis are plenty and overwhelming. This section demonstrates some of the commonly used applications of sentiment analysis.

1.3.1 Business

Businesses worldwide strive to earn delight and patronage of its customer base. Sentiment analysis in business is a major breakthrough for the entire brand revitalization. It helps industries to identify the potential risks, gives an insight into the future endeavor as well as the marketing strategies. The user experience regarding a particular entity may be positive, negative or neutral. These experiences are expressed in the form of posts or feedback through social media. Sentiment analysis helps to categorize the user’s experience to opinions so as to identify the opportunities for up-selling, design new marketing strategies and to improve the existing features. Many of the popular online shopping websites such as eBay, Flip Kart, Amazon etc allows users to exchange their opinions and analyses the sentiment of users. This helps to understand what is being properly applied with regards to services and products. Sentiment analysis helps such websites to win over the dissatisfaction of customers into active support by analyzing the massive volume of their opinionated data.

Sentiment analysis is often applied in business intelligence to identify the subjective reasons as for why the consumers are not buying a product, whether the consumer service support meet their customer expectations.

1.3.2 Brand Reputation Management

Brand Reputation simply states how a particular product brand is viewed by users. A favourable online reputation means customers are more likely to trust the brand and purchase products and services from that particular brand. The brand reputation management services require a large number of human interactions to differentiate positive, negative and neutral opinions. Sentiment analysis is applied to brand reputation management process as a means of discerning the direction of public
sentiments. Sentiment analysis also enables retrieving real-time data containing the context of a discussion of the specific brand to get to know the overall assessment of the brand. With this real-time data, companies and organizations get the opportunity to observe, analyze and respond to the issues regarding the product or services before they actually spin out of control.

1.3.3 Government

Sentiment analysis can be applied to discover public opinion and trends, especially for political campaigns. Twitter sentiment analysis has become a powerful online predictor of political sentiments and is proven to be a beneficial platform for the prediction of election results. Tweets describing political sentiment demonstrate the close correspondence to political parties and their positions, thereby indicating that tweets plausibly reflect offline political landscape. The sentiment analysis methods analyze the sentiment polarities to allow the government to draw a qualitative opinion of active users on social media based on the topics of interest to the public administration.

Sentiment analysis can also be applied in the field of political science, sociology, and psychology to examine the trends, opinions, ideological bias and gauge the reactions.

1.4 Rationale of the Study

Sentiment analysis has become a very active research field with a wide range of applications in nearly all domains. With the exponential growth of social media sites, it has become a formidable task to extract relevant information from blogs as well as forum postings. The micro blogging site, Twitter is increasingly being used by organizations to gather public opinion. It offers the fastest way to build a targeted audience even if they do not follow us back. Unlike other social media sites such as Facebook, the Twitter seems to be a little more diverse in what they are willing to share. This is one of the primary reasons why Twitter is more suitable for this particular context. Recent reports say that this marketing tool has received almost eight million unique visitors per month and around fifty million tweet posts appearing daily. Since the proliferation of commercial applications and Twitter being a fantastic
tool in giving a product brand a ‘voice’ and a ‘personality’, it has been chosen as the best programming platform to conduct sentiment analysis tasks in our work.

Although the Twitter-based information extracting techniques enable gathering direct responses from the target audience, not much by the way of research has been carried out to predict, model and forecast user behaviour using the already existing and often abundant supply of personal data housed by the Social Network. This ready and continuous stream of social media information could be analyzed with the use of an efficient machine learning algorithm, to predict social behaviour patterns. Thus predictive algorithm could prove useful in understanding and further analyzing social dimensions and to forecast trends in social behaviour.

Using sentiment analysis, the information extracted from the chosen Twitter sample population shall be analyzed using machine learning algorithms to derive at meaningful patterns. The analysis of primary data in this regard, employing automated data mining techniques is vital for gathering social behavioural patterns. The results would also help to demonstrate the inhered possibilities, advantages and applicability of the most commonly used machine learning algorithms.

Many of the companies and organizations surrounding sentiment analysis had already built their own sentiment analysis systems. These pragmatic applications and commercial interests urged a strong motivation for conducting research in the domain of sentiment analysis. Sentiment analysis offers many research problems that had never been studied before. This motivated us for further research.

1.5 Contributions

The major contributions of the work presented in this thesis are described as follows.

- Two real-world datasets of 16000 tweet samples based on popular product brands Samsung Galaxy S6 and Motorola MX were extracted from Twitter API over a consecutive time period of three years.
- A detailed study of recent literature works in sentiment analysis was presented.
Chapter 1

Introduction

Mining Social Media Information to Predict Social Behavioural Patterns

- Made a comparative analysis of the existing machine learning techniques namely supervised and unsupervised learning methods with the real-world datasets and the publicly available data obtained from GitHub.

- Designed and formulated a fuzzy-based partition clustering model (FPC).

- Designed and developed a fuzzy hybrid hierarchical model (FHHC) to deal with the problems of fuzzy-based partition clustering model and extended the FHHC model to EFHHC in order to achieve more reliable and promising results.

- Designed and developed a novel cluster-based approach (MFTCT) to overcome the limitations of the above-mentioned models.

- Devised and formulated an innovative supervised learning approach (MC4.5) for accurate and faster sentiment analysis results.

1.6 Objectives

The objectives of the research work are described as follows:

- To conduct a comprehensive study of recent literature works related to the existing sentiment analysis approaches and its applications.

- To gather real-world tweet samples from the Twitter Application Programming Interface (API).

- To execute sentiment analysis on tweets over the existing unsupervised approaches to learn the merits and demerits of the state-of-the-art methods.

- To perform a detailed study of supervised learning techniques and to carry out a comparative experimental analysis with both the supervised and unsupervised learning techniques to identify the most efficient machine learning algorithm in the domain of sentiment analysis.

- To propose a faster and effective machine learning model to cope with all the short comings of existing methods when the sentiment analysis is executed.
1.7 Thesis Outline

The structure of the thesis is organized as follows:

Chapter 2 covers the literature survey of most relevant studies and related concepts which includes different machine learning techniques namely supervised and unsupervised machine learning relevant to sentiment analysis.

Chapter 3 proposes a novel fuzzy-based partition clustering model for the sentiment analysis of real-world tweets regarding a popular product brand.

Chapter 4 discusses the limitations of partition clustering methods and proposes a new fuzzy-based hierarchical clustering model to overcome the shortcomings of partition clustering algorithms.

Chapter 5 proposes a new efficient model to deal with the problems of both the partition and hierarchical clustering algorithms.

Chapter 6 proposes a highly efficient and fast supervised learning model to predict the sentiment analysis of a trendy product brand.

Chapter 7 deals with the detailed experimental analysis of proposed models with other state-of-the-art methods and highlights the advantages of the proposed model.

Chapter 8 finally concludes the overall work and experimental results along with suggestions for the future scope of further research related to sentiment analysis.