UNIVERSITY OF SUSSEX

Dynamic Shifts In Political Attitudes

A Cluster Based Analysis Into The Impacts of Transformative Events on Political Attitudes

Alisha Alex
Candidate Number : 260776
Supervisor : Maxine Sherman

Submitted for the Degree of Masters of Science
University of Sussex
September 2023

September 2022

Statement Of Originality

This report is submitted as a requirement for the Degree of MSc Data Science at the University of Sussex. This is to certify that to the best of my knowledge, this thesis and the research presented is a product of my own labour as well as it is been developed by me as a resultant of my own research except where indicated in the text.

Alisha Alex

Acknowledgement

First and foremost, I would like to thank Almighty for the greater support and strength for pushing me forward.
I would like to sincerely express my gratitude to my Supervisor, Maxine Sherman for supervising this project and providing me with knowledge about clustering methodologies and politics. Her valuable guidance, support and dedication throughout every stage of my project have been greatly appreciated.

Finally, I want to express my heartfelt appreciation towards my family and friends, for their immense support and encouragement.

Abstract

Political attitudes are critical components of an individual's engagement with social and political landscapes. The catalysts for shifts in political attitudes encompasses personal experiences as well as social events. This thesis delves into the changes in political attitudes due the impacts of transformative events including the most popular Brexit as well as the pandemic virus outbreak Covid -19. In this research, we work on a dataset created by extracting relevant features from a survey dataset. The datasets are then divided into two groups marking the two periods that are not impacted by the respective events as well as well impacted by the events. For the given research, we consider the years before 2020 as an era unimpacted whereas 2020,2021 as a time period impacted by the events of Brexit and Covid-19. It is then followed by the implementation of an Unsupervised Clustering Algorithm called K-Prototype Clustering. It is then followed by analysing the results obtained. The Cluster evaluation is carried out by a Classification Model.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

In an ever-changing global landscape, political attitudes and beliefs undergo continuous evolution shaped by the impacts of momentous events [1]. The societies find themselves navigating through paradigm-shifting occurrences that which creates profound transformations in the political systems. From revolutions and economic crises to environmental disasters and spontaneous viral outbreaks, these events have the power to both challenge as well as reinforce long standing political ideologies and systems [2]. In this modern era, an understanding of how such transformative events impact public opinions and reshape political landscape is critical for the comprehension of complexities of governance and democracy. The consequences of changes in political attitude extend beyond the existing mere theoretical contemplation.

Transformative events are defined as events that possess the potential to change the collective consciousness as well as impact the ordinary lives of people and societies. Among such occurrences, two events have left remarkable imprints on the UK political stage, Brexit, and Covid 19 Pandemic [3]. This study conducts an exploratory voyage into the intricate dynamics of how Brexit and COVID-19 have impacted and altered the political attitudes of the UK Communities. These events, though vastly different in nature shares a common thread as stimulants of changes, causing intensive uncertainties, profound emotions as well as ideological divisions among the citizens.

1.1. Motivation

Political attitudes act as integral components for human demeanor and creates the foundation upon which the governments and societies are built. Societies and governments as whole are strongly impacted by the aggregates of political attitudes of individuals. These attitudes influence the type of political leader that comes to power, policies, and law implementations as well as overall stability of political systems. The formation of political attitudes can be influenced by multitude of factors, which includes various factors such as family upbringings, education, culture, media, peer influences as well as personal experiences. Over time, these attitudes become deeply imprinted in individual's identities and guide them in their political choices, which includes voting behavior, policy supports, and political activity participations. This study sets out to explore the fluid nature of political attitudes as well as the multifarious influences that molds them. The uproarious journey of the United Kingdom's departure from European Union and the unparalleled global health crisis have undoubtedly influenced political attitudes in interesting and diverse ways.

Brexit, short for "British exit," is defined as the process by which the United Kingdom decided to leave the European Union (EU). Originating from a national referendum held on June 23, 2016, in which 51.9% of voters chose to leave the EU while 48.1% voted to remain, Brexit became a widely popular acronym [29] .In 2016, the Brexit referendum created a crucial moment in the UK's political stage triggering intensive debates on issues such as national identity, sovereignty, and economic integration. The divisive nature of campaigns and the eventual decision to leave the European Union created a fundamental shift in political attitudes [4]. As separations deepened, communities and individuals found themselves reexamine their political views and values which then lead to significant alterations in political ideologies and alliances. As per 2016, June, 51.9% people chose to leave the EU whereas 48.1 % wanted to remain. However, as per a poll conducted in 2022, 57% of people consider Brexit as a wrong choice whereas 43% consider it as a right decision [5]. The idea of Brexit created strong concerns among the people and impacted their social, economic, and political views [6]. It is intriguing to research how this policy impacted the attitudes of people across the country. Did the EU referendum impact the party affiliations? Did it change people's views towards the traditional values and social inequalities? It is interesting to determine these changes across the two time periods.

Meanwhile, In early 2020, the emergence of the pandemic Covid-19 created an unrivalled global crisis, demanding for immediate responses from both government and citizens alike. The pandemic brought

public health and socio-economic concerns to the public forefront which led to the creation of adaptive refined policy decisions. Furthermore, the impacts of this viral outbreak extended far beyond health sectors, penetrating every aspect of society including governance, economy, and international relations. As individuals dealt with fear and the consequences of pandemic, their political attitudes underwent reflective transformations, including shifts of governmental efficacy perspectives, public responsibility as well as the role of science in policy making. It is interesting to research whether the events of Covid played a role in reshaping the political attitudes. Did Covid play a major role in the time-period of 2020,2021?

This study embarks to unravel these events and their far-reaching impacts on political attitudes. Through a comprehensive examination of opinion data based on the survey organized by the Natcen Social Research, we seek to understand the intricate interplay between these transformative events as well as the fluctuations in the political attitudes. Therefore, as we delve into these intersections of political attitudes and transformative events, we obtain a deeper understanding of the nature of political attitudes. The case of Brexit and COVID-19 pandemic serves as an illuminating canvas on which we paint an interesting picture of how the transformative events have influenced the political attitudes of people.
This research is set out to endeavor the complex and fluid nature of political attitudes and the multifarious impacts that molds them.

1.2. Dissertation Outline

1.2.1. Selection of Survey Questions from British Social Attitude Survey conducted by NatCen Social research that which captures political attitudes and construct a dataset consisting of features using the selected questions for multiple years.
1.2.2. Conduct a Cluster Analysis on the created dataset using K-Prototype Clustering
1.2.3. Analyze the patterns formed as well as conduct a comparative analysis between Pre- Brexit and Brexit discussion Time period as well as an analysis into the impact of COVID-19 Outbreak.

1.3. Structure Of Dissertation

An overview of the dissertation is presented below:
In **Chapter 2,** a comprehensive summary of prior research of the topic is presented, encompassing various techniques, datasets and algorithms used.
In **Chapter 3,** we explore several key facets of the topic, encompassing diverse methodologies and techniques employed in the study, such as understanding the scenarios and means of finding optimal number of clusters as well as K-prototype Clustering.
In **Chapter 4**, a detailed description of the Experiment Set up designed to undertake the study is presented encompassing the process of selection of Questions and their justifications to the process of K-prototype Clustering.
In **Chapter 5**, a comprehensive and detailed analysis of the conducted experiments are presented with obtained results through visualization and analyses.
In **Chapter 6**, an overview of the study is presented, including the conclusions drawn from the conducted research as well as suggestions for future work to be undertaken for further investigations.

Chapter 2
Background Research

In this section, we investigate previous research conducted on the shifts in political attitude and opinions as well as their transformations in response to various events. Multiple methods were entrusted to analyze and comprehend these alterations in political opinions. The primary focus of this literature review centers around cluster-based approaches that which have demonstrated remarkable success in understanding such political changes.

2.1. Determination of Political Attitudes

In context of this research, the careful identification of relevant questions plays the vital role. Selection of questions with political attitudes in this research project stands as an important determinant for the successful completion of this study. Attitudes are defined as an enduring orientation towards objects including ideas, people, etc. [7]. Hence, Political attitudes are relatively persisting orientations towards political objects [7].

In 1993, Professor Reiner Riemann, a renowned faculty member of Bielefeld University's psychology department, undertook an investigation into political attitudes. He explored the interplay between attitudes and politics by associating with political issues [8]. His study mainly included 162 political issues of relevance and categorized the attitudes of people based on their responses by developing a 5-dimension models including conservativism, authoritarianism, liberalism etc. His concept of identifying and categorizing the political attitudes by analyzing various sets of issues gives an interesting approach for determining the relevant features that capture the political essence by looking through a perspective consisting of relevant elements of social issues. Similarly, in 1996, Jeoffrey Evans contributed to the field by conducting a study that focused on the evolution of political attitudes. He established a clear understanding of these attitudes using the left-right scale, meticulously organizing, and describing various political stances [9]. His method of analyzing various perspectives of social norms by incorporating them into a system of scales gives insights to relevant attributes that defines a political attitude. These two studies comprehensively clarify the norms, contextual conditions, and the intricate relationship between attitudes and politics. In alignment with these scholarly works, I have chosen to include insights from these research papers in guiding for the selection of appropriate questions from the NatCen Survey that capture the essence of political attitudes.

2.2. Cluster-Based Analysis

Cluster-based analysis has long been an eminent approach for examining complex datasets, facilitating the comprehension of underlying patterns, structures as well as predictive insights. Being an unsupervised machine learning technique, cluster analysis has been widely used in multiple fields, which also includes politics which employs the investigation of survey data. Notably, multiple studies have utilized the concepts of cluster-based methodologies to explore various relationships based on the survey data.

In 1986, John A. Fleishman et al. conducted an interesting study that which primarily focus on the organization of political attitudes through the implications of cluster analysis. His research mainly aimed to classify the individuals into subgroups with relatively similar profiles of opinions on 12 political issues [10]. The data utilized in this study was obtained from the survey of 1980 American Election Study, conducted by the Survey Research Centre at the University of Michigan. The sample size consisted of 483 participants along with a set of 12 political issues were carefully selected which ranges from civil rights to inflation as well as unemployment.

For the achievement of this classification of individuals into distinct clusters, the study created a two-step clustering procedure. Initially, Ward's hierarchical clustering method was utilized, combining cases into clusters to minimize the error sum of squares. Subsequently, the clusters obtained from the Ward's method served as the starting point for an iterative reallocation clustering process. This iterative procedure involved moving cases between clusters until a goodness-of-fit criterion was satisfied,

ceasing when no further reallocation could improve the within-cluster sum of squares. The dissimilarity between individuals was computed using the Euclidean distance metric.
Finally, because of the cluster analysis, six distinct clusters were obtained, representing various political groups such as liberals, quasi-liberals, conservatives, among others. Interestingly, the study was able to demonstrate that the formed clusters exhibited systematic differences in political party affiliation, self-identified ideology, as well as voting patterns for the 1980 Presidential Election. Crucially, the research concluded that political attitudes were not organized along a single dimension of ideological liberalism or conservativism, emphasizing the subtle and varied nature of political attitudes among the surveyed population. This intriguing study conducted by Fleishman et al. provided significant contributions to the understanding of complex organizations of political attitudes as well as gave valuable insights into the difference in opinions on various political issues. The usage of cluster analysis in this research served as a viable approach for the exploration and comprehension of the underlying structure of political attitudes within the provided population.

Similarly, another interesting study was conducted by HyeHyun et al. in 2012 focused on analyzing the trust levels in government among the public in the United States and 19 European countries [11]. The researchers performed cluster analysis on two datasets including the surveys: Citizenship, Involvement, Democracy (CID) survey and the European Social Survey (ESS) by employing demographic variables, media usage, as well as social participation to classify the public into varied groups. The two-step cluster analysis technique was utilized, wherein the log likelihood distance measure was used to compute the distances between clusters. The optimal number of clusters was determined with the application of Schwarz's Bayesian Criterion (BIC) clustering criterion, and continuous variables were standardized for maintaining consistency. Consequently, three homogeneous public segments were identified for each of the 20 countries. Mainly, these segments exhibited variations in trust levels towards the governmental institutions, serving as a crucial indicator for the quality of government-public relationships. This study highlighted the effectiveness of cluster-based analysis in discerning and understanding the key aspects of government-public interactions in varying contexts.

K prototype clustering, a category of K-means clustering designed for mixed data types, has emerged as an influential approach for grouping the data points based on similarity for data containing mixed data types. The K-means method, extensively known as one of the most popular and widely used clustering methodologies, serves as its foundation mark.
In 2022, Hakan Stattin et al. conducted an interesting study that which aims to explore the transformation of political interest into diverse political values, attitudes, and behavior by investigating basic values [12]. This research included the application of cluster analysis on a dataset developed through a questionnaire-based investigation involving adolescents. The study hypothesized that the political interests play a critical role as a predictor of political engagement. To build the dataset, a comprehensive assessment was conducted in 57 classes across three high schools in Sweden by utilizing self-report questionnaires.
In this study, an approach of two-step cluster analysis was employed. Initially, a hierarchical cluster analysis using Ward's method was employed to identify the optimal number of clusters. Furthermore, armed with this knowledge of the number of clusters, a non-hierarchical cluster analysis such as K-means clustering was employed to obtain the final cluster solution. The initial hierarchical cluster analysis helped to reveal the existence of five clusters, followed by the subsequent application of K-Means clustering, which ultimately identified three distinct clusters characterized by lower or the average values on political interest as well as perceived importance of welfare of others.
The utilization of K-means clustering in this research demonstrates its importance as a powerful methodology for effectively handling the data and extracting meaningful insights from complex datasets. By the application of K-means clustering, the study successfully identified unique patterns of political interest and its implications for political engagement among the examined adolescent population. This research demonstrated its potential for enhancing our understanding of the relationships between various political factors, leading to valuable contributions to the field of political science research.

2.3. Brexit & Covid-19

Numerous studies have vastly studied the impacts of Brexit on various dimensions of British life. Among such investigations, an interesting study conducted by Joseph et al. in 2020 investigated the political attitudes of UK citizens following the EU referendum [13]. The primary focus of this research is not the outcomes of Brexit rather it focuses on dividing lines on political attitudes in the aftermath of Brexit. For the accomplishment of this, the researchers adopted the data from the 2016-2017 round of the European Social Survey, that which provided valuable understanding into the changing political landscape based on the impacts of Brexit. For the assessment of these impacts of Brexit on political attitudes, the study incorporated two political scales, namely GAL-TAN [14] as well as the left-right scale [15]. These scales provided a methodology for describing the dynamics of the "new" and "old" politics concerning Brexit. Furthermore, along with the usage of these scales, a statistical approach such as Multi Correspondence Analysis (MCA) is used to unravel the underlying patterns as well as structures within the data. The conclusions of this MCA provided a set of observations, allowing for a comprehensive analysis of the relationships between different variables. From the MCA analysis, it becomes evident that the divisions among politically interested and politically uninterested individuals, as well as difference between varied age groups and genders, are more noticeable in comparison to the differences between individuals in varying social class positions. This research mainly concentrated the analysis of political views based on the two scales by the utilization of one scale to analyze the post Brexit times whereas another for pre-Brexit times. These findings highlight the interesting interplay of political attitudes in the post-Brexit era, wherein factors such as engagement level, age, and gender play a more significant role in the formation of political perspectives than just the social class distinctions. The conclusions of this research increased our understanding of the post-Brexit political landscape and emphasized on the importance of considering multiple factors in comprehending citizens' attitudes towards this transformative event.

Likewise, In the year 2020, Willem et al. published a significant study exploring the impacts of Covid-19 on socio-political attitudes [16]. The researchers organized a survey targeting Covid-19 patients, enabling them to discern the great impacts of the pandemic on various social and political dimensions. Their study provided valuable conclusions into the potential emergence of a “new fault line” within the political landscape, arising from the increased demands for the implementation of straightforward policy solutions with respect to the challenges of pandemic.
This study conducted by Willem et al. contributed substantially to the existing literature, providing an elaborative understanding of the far-reaching effects of the Covid-19 pandemic on social and political attitudes. By investigating the perspectives of Covid-19 patients, the study revealed how the global health crisis may have reformed the political discourse as well as triggered calls for simple policy approaches. Such conclusions are crucial for policymakers and researchers alike, as they navigate through the complexities of governing in response to the pandemic and strive to address the evolving issues as well as the aspirations of the population.

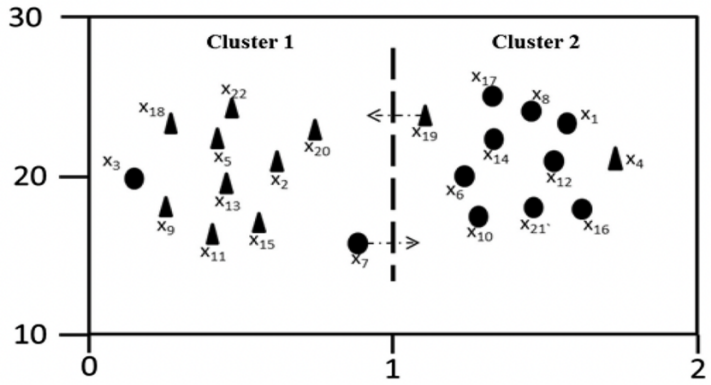
Chapter 3
Preliminaries

Partitioning set of objects into homogeneous groups or clusters is a fundamental operation in data mining [17]. Cluster analysis entails the application of clustering algorithms to discover patterns or groupings within a dataset, serving as a commonly employed methodology for Exploratory Data Analysis. Within the context of cluster analysis, these algorithms establish clusters in a manner where the similarity measures between elements within a cluster exceed those of elements in any other cluster. Successful cluster analysis necessitates the establishment of a similarity measure for every possible pairwise combination of entities slated for clustering. Such similarity metrics encompass options like Euclidean, probabilistic, cosine distance, and correlation. The selection of a specific cluster analysis approach essentially embodies the definition of a cluster itself [18]. The datasets to be mined often contain millions of objects described by tens and thousands of various types of attributes or variables [18]. Given this substantial complexity, it is imperative that the employed clustering algorithms exhibit scalability, effectively accommodating datasets of immense proportions and a diverse range of attribute types. The focus of this research is directed towards datasets comprising a mixture of numerical and categorical variables, which present unique challenges in terms of clustering analysis.

3.1. Clustering Algorithm

Among the established clustering techniques, the K-Means algorithm emerges as a widely adopted method, particularly valued for its efficacy in handling large datasets. Yet, its suitability falters when confronted with data containing categorical variables. To address this limitation, Huang introduced an innovative approach termed the K-prototype algorithm, engineered to proficiently manage clustering tasks involving a mix of data types encompassing both numerical and categorical variables. It is notable that in K-Means, the centroid is defined as its mean whereas in K-Modes it is the mode Value. Rooted in the principles of partitioning, the K-prototype algorithm represents an evolutionary step beyond the K-Means and K-mode clustering methodologies, specifically tailored to navigate the complexities of clustering mixed data types. This enhancement seeks to augment the existing clustering toolkit, facilitating more comprehensive data analysis and pattern discovery. The below figure shows a pictorial representation of categorical and noncategorical points in the respective clusters formed via k prototype clustering.

Fig 3.1: K- prototype clustering



Source: https://www.ias.ac.in/public

3.1.1. K-Means clustering

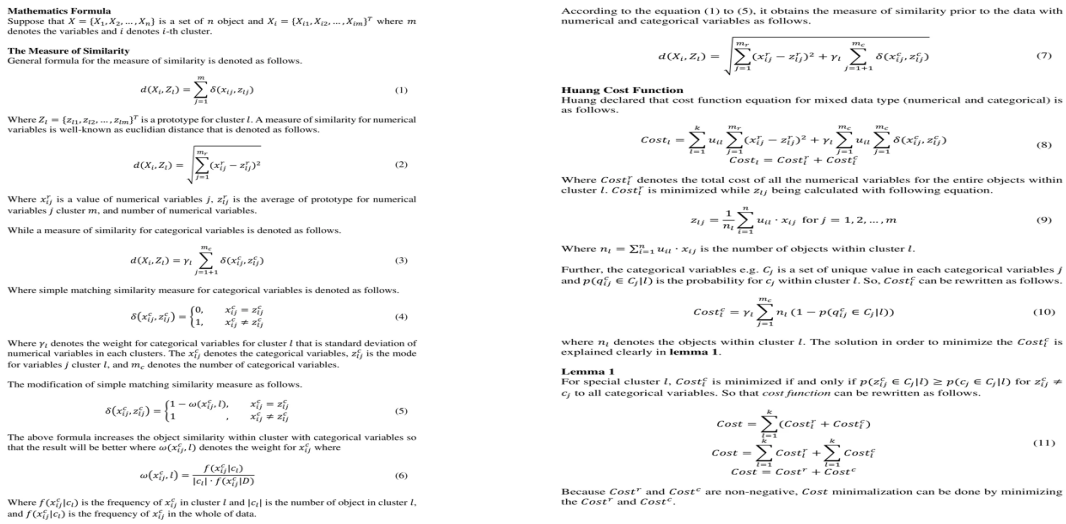
K-means algorithm is classified as a partitional or non -hierarchical clustering method.
Given a set of numeric objects **X** and an integer number **K** (<=n). The K-means algorithm searches for a partition of **X** into **K** clusters that minimizes the within groups of sum of squared errors [19]. This methodology is founded on the computation of distances between data points located in multidimensional space. The algorithm proceeds by iteratively assigning data points to the nearest

cluster centroid, followed by updating the centroids through the incorporation of data points residing within each cluster. This iterative process, particularly suitable for numerical data, draws upon distance metrics such as Euclidean distance for distance computation.
However, when confronted with categorical data, a notable challenge emerges due to the absence of a natural or meaningful approach for calculating distances. Unlike numerical variables where the magnitude between values carries interpretive value, categorical variables like colors or gender lack this inherent numeric interpretation, thereby rendering direct distance calculation infeasible. Further complexity arises from the algorithm's reliance on averaging data points within a cluster to facilitate centroid updates. This procedure, well-suited for numerical data, encounters hurdles in the realm of categorical data due to the absence of a straightforward method for defining a representative "average" value. This inherent trait poses a substantial obstacle to the successful application of the K-means algorithm to datasets characterized by categorical variables.

3.1.2. K-Modes clustering

K-Modes clustering has been primarily conceived to cater to the distinctive attributes of categorical data. Its methodology revolves around specialized dissimilarity metrics that meticulously account for the presence, absence, or discrepancies within categorical values. The fundamental premise of K-Modes entails the utilization of categorical-specific dissimilarity metrics. Notably, K-Modes employs a distance metric that systematically considers the disparities in features or the existence/non-existence of values within categorical attributes. This deliberate choice of distance measurement underscores the algorithm's proficiency in accommodating categorical data, effectively addressing the challenges that K-Means encounters when confronted with non-numeric attributes. K-Modes can provide solutions even when data have sparse marginal distribution [20]. This adaptive nature enhances its utility in scenarios characterized by unique data distributions and reinforces its position as a valuable tool for categorical data analysis within the realm of clustering techniques.
The K-prototype algorithm synergistically combines the principles of both K-means and K-modes clustering methodologies. Like K-Means, the K-Prototypes algorithm employs the 'Euclidean' distance to quantify the separation between numerical variables. Diverging from the K-Means approach, K-Prototypes assesses the separation between categorical variables by considering the count of shared categories. This cohesive integration of diverse distance measures underscores the algorithm's ability to effectively handle the complexities of mixed data types. The K-Prototype algorithm holds practical significance due to its suitability for real-world databases, where mixed data types are often prevalent and commonly encountered [19].

Fig 3.2: Mathematical derivation of K-Prototype



Mathematics Formula
Suppose that $X = \{X_1, X_2, \ldots, X_n\}$ is a set of n object and $X_i = \{X_{i1}, X_{i2}, \ldots, X_{im}\}^T$ where m denotes the variables and i denotes i-th cluster.

The Measure of Similarity
General formula for the measure of similarity is denoted as follows.

$$d(X_i, Z_l) = \sum_{j=1}^{m} \delta(x_{ij}, z_{lj}) \quad (1)$$

Where $Z_l = \{z_{l1}, z_{l2}, \ldots, z_{lm}\}^T$ is a prototype for cluster l. A measure of similarity for numerical variables is well-known as euclidian distance that is denoted as follows.

$$d(X_i, Z_l) = \sqrt{\sum_{j=1}^{m_r} (x_{ij}^r - z_{lj}^r)^2} \quad (2)$$

Where x_{ij}^r is a value of numerical variables j, z_{lj}^r is the average of prototype for numerical variables j cluster m, and number of numerical variables.

While a measure of similarity for categorical variables is denoted as follows.

$$d(X_i, Z_l) = \gamma_l \sum_{j=1+1}^{m_c} \delta(x_{ij}^c, z_{lj}^c) \quad (3)$$

Where simple matching similarity measure for categorical variables is denoted as follows.

$$\delta(x_{ij}^c, z_{lj}^c) = \begin{cases} 0, & x_{ij}^c = z_{lj}^c \\ 1, & x_{ij}^c \neq z_{lj}^c \end{cases} \quad (4)$$

Where γ_l denotes the weight for categorical variables for cluster l that is standard deviation of numerical variables in each clusters. The x_{ij}^c denotes the categorical variables, z_{lj}^c is the mode for variables j cluster l, and m_c denotes the number of categorical variables.

The modification of simple matching similarity measure as follows.

$$\delta(x_{ij}^c, z_{lj}^c) = \begin{cases} 1 - \omega(x_{ij}^c, l), & x_{ij}^c = z_{lj}^c \\ 1 \quad , & x_{ij}^c \neq z_{lj}^c \end{cases} \quad (5)$$

The above formula increases the object similarity within cluster with categorical variables so that the result will be better where $\omega(x_{ij}^c, l)$ denotes the weight for x_{ij}^c where

$$\omega(x_{ij}^c, l) = \frac{f(x_{ij}^c | c_l)}{|c_l| \cdot f(x_{ij}^c | D)} \quad (6)$$

Where $f(x_{ij}^c | c_l)$ is the frequency of x_{ij}^c in cluster l and $|c_l|$ is the number of object in cluster l, and $f(x_{ij}^c | c_l)$ is the frequency of x_{ij}^c in the whole of data.

According to the equation (1) to (5), it obtains the measure of similarity prior to the data with numerical and categorical variables as follows.

$$d(X_i, Z_l) = \sqrt{\sum_{j=1}^{m_r} (x_{ij}^r - z_{lj}^r)^2 + \gamma_l \sum_{j=1+1}^{m_c} \delta(x_{ij}^c, z_{ij}^c)} \quad (7)$$

Huang Cost Function
Huang declared that cost function equation for mixed data type (numerical and categorical) is as follows.

$$Cost_l = \sum_{i=1}^{k} u_{il} \sum_{j=1}^{m_r} (x_{ij}^r - z_{lj}^r)^2 + \gamma_l \sum_{j=1}^{m_c} u_{il} \sum_{j=1}^{m_c} \delta(x_{ij}^c, z_{lj}^c) \quad (8)$$
$$Cost_l = Cost_l^r + Cost_l^c$$

Where $Cost_l^r$ denotes the total cost of all the numerical variables for the entire objects within cluster l. $Cost_l^r$ is minimized while z_{lj} being calculated with following equation.

$$z_{lj} = \frac{1}{n_l} \sum_{i=1}^{n} u_{il} \cdot x_{ij} \quad \text{for } j = 1, 2, \ldots, m \quad (9)$$

Where $n_l = \sum_{i=1}^{n} u_{il} \cdot x_{ij}$ is the number of objects within cluster l.

Further, the categorical variables e.g. C_j is a set of unique value in each categorical variables j and $p(q_j^c \in C_j | l)$ is the probability for c_j within cluster l. So, $Cost_l^c$ can be rewritten as follows.

$$Cost_l^c = \gamma_l \sum_{j=1}^{m_c} n_l \left(1 - p(q_j^c \in C_j | l)\right) \quad (10)$$

where n_l denotes the objects within cluster l. The solution in order to minimize the $Cost_l^c$ is explained clearly in **lemma 1**.

Lemma 1
For special cluster l, $Cost_l^c$ is minimized if and only if $p(x_{ij}^c \in C_j | l) \geq p(c_j \in C_j | l)$ for $x_{ij}^c \neq c_j$ to all categorical variables. So that *cost function* can be rewritten as follows.

$$Cost = \sum_{l=1}^{k} (Cost_l^r + Cost_l^c)$$
$$Cost = \sum_{l=1}^{k} Cost_l^r + \sum_{l=1}^{k} Cost_l^c \quad (11)$$
$$Cost = Cost^r + Cost^c$$

Because $Cost^r$ and $Cost^c$ are non-negative, $Cost$ minimalization can be done by minimizing the $Cost^r$ and $Cost^c$.

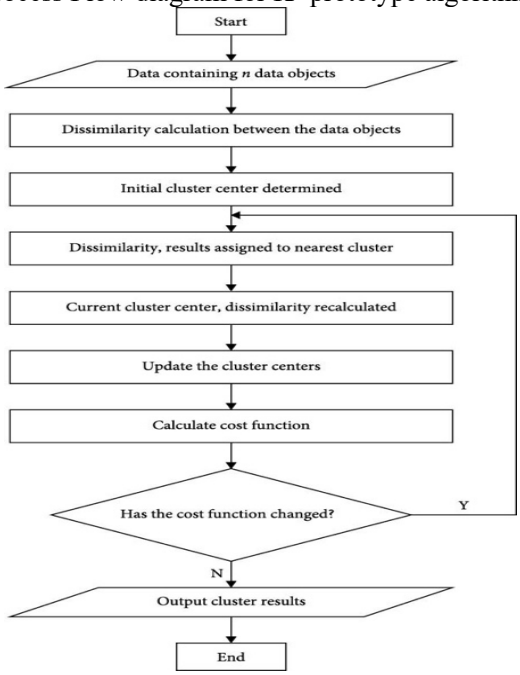
Source: https://towardsdatascience.com

3.1.3 The K-Prototype clustering algorithm

The K-Prototypes clustering algorithm closely resembles the K-Means clustering algorithm, but distinct variations emerge in the method of calculating data point distances and the approach to selecting fresh prototypes for clusters. The subsequent sections outline the sequential steps constituting the K-Prototypes clustering algorithm.

Step 1: Initially, we designate **K** data points from the input dataset to serve as the initial prototypes.
Step 2: Subsequently, we determine the distance between each data point and the existing prototypes. The calculation of these distances adheres to the methods outlined in the preceding sections.
Step 3: Upon establishing the distances between data points and prototypes, we proceed to allocate data points to clusters. In this process, every data point is assigned to the cluster whose prototype exhibits the closest proximity to the respective data point.
Step 4: Following the allocation of data points to clusters, we compute fresh prototypes for each cluster. The methodology for calculating these prototypes involves determining the mean for numerical attributes and the mode for categorical attributes.
Step 5: If the new prototypes are the same as the previous prototypes, we say that the algorithm has converged. Hence, the current clusters are finalized. Or else go back to Step 2 [21].

Fig 3.3: Process Flow diagram for K -prototype algorithm



Source: https://www.researchgate.net

3.2. Elbow Method

A pivotal stage within any unsupervised algorithm involves ascertaining the most suitable count of clusters for partitioning data. The Elbow method serves as a technique employed to ascertain the appropriate number of centroids (**K**) for utilization in a K-Prototype clustering algorithm. It is a visual method to test the consistency of the best number of clusters by comparing the difference of the sum of square error (SSE) of each cluster [22]. This technique encompasses the computation of the "cost" or "distortion" of clustering for diverse cluster quantities and subsequently identifying a juncture on a graph where the cost commences leveling off. This juncture exhibits an "elbow" shape, signifying the juncture beyond which the inclusion of more clusters yields negligible reduction in intra-cluster variability.

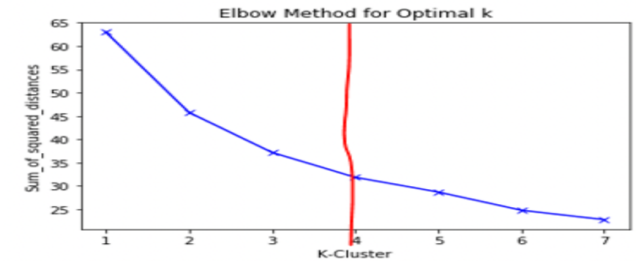
3.2.1. Plot Interpretation

The plot typically exhibits a descending pattern in cost as the cluster count rises. This trend occurs because smaller clusters tend to minimize the distances between data points and their respective cluster centers, culminating in reduced cost. Nevertheless, there reaches a certain point where the addition of more clusters scarcely impacts cost reduction. This juncture is recognized as the **Elbow Point**.

3.2.2. Elbow Point Identification

The Elbow point depicted on the plot denotes the optimal cluster count for the dataset. It corresponds to the point just prior to the cost curve's plateauing. Further incorporation of clusters after this point results in marginal reductions in cost.

Fig 3.4: Graph of Elbow method (example)



Source: https://www.researchgate.net

3.3. Silhouette Score Coefficient

Silhouette Score Coefficient is defined as a measure of cohesion among the data points in a cluster. It is a better choice of metric to calculate the quality of clusters formed by partitioning-based clustering algorithms [23]. It measures how similar is a datapoint within a cluster(cohesion) compared to other clusters(separation) [24]. The respective values range between -1 and 1. Higher the value of the silhouette score coefficient, optimal the cluster solution.

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

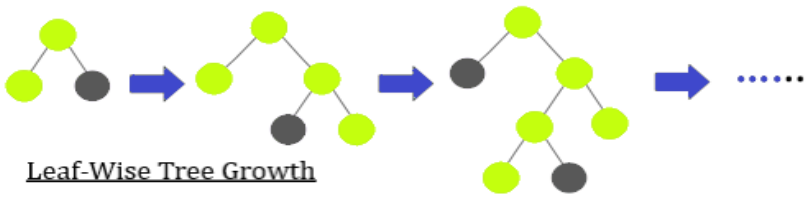
3.4. Cluster Evaluation methodology

The clusters are evaluated by developing a classification model on top and treating the clusters as labels. Higher the quality of clusters ,better the accuracy of the classification model that which predict them.

3.4.1 LightGBM (Light Gradient Boosting Machine)

LightGBM stands as a gradient-boosting framework that draws its foundation from decision trees, strategically harnessed to amplify the overall efficiency of the model. This framework operates on a learning algorithm grounded in trees [25]. Noteworthy is LightGBM's distinctive vertical growth pattern, progressing leaf by leaf. A remarkable strength of LightGBM lies in its organic adeptness at managing categorical features, obviating the requirement for laborious preprocessing procedures. This characteristic emerges as a substantial asset within the context of the dissertation's research objectives.

Fig 3.5: Architecture of LightGBM



Source: https://www.geeksforgeeks.org/

LightGBM is well known for its efficient handling of categorical features hence avoid the needs of preprocessing such as one-hot encoding. Furthermore, SHAP values can be calculated for LightGBM and for models such as XGBoost. However, LightGBM gives more straightforward and well-documented values compared to models like XGBoost.

3.4.2. Attributes for Evaluation

In our study, the obtained clusters are evaluated using the methodology of classification. Here, the clusters are evaluated based on the distinctiveness as well as informativeness of clusters.

3.4.2.1. Distinctiveness of Clusters

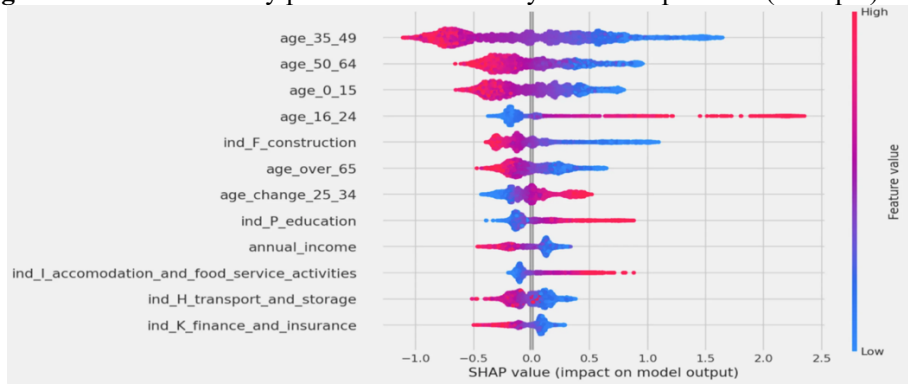
The research entails an evaluation of cluster differentiation effectiveness using a cross-validated F1 score. The F1 score, encompassing both precision and recall, serves as a metric for this assessment. By employing the F1 score, which adeptly combines precision and recall, the study aims to gauge the classification model's proficiency in distinguishing between clusters. This evaluative approach underscores the comprehensive analysis of clustering outcomes presented in the dissertation.

$$F1 = \frac{2(Prescion * Recall)}{Prescion + Recall}$$

3.4.2.2. Informativeness of Clusters

SHAP (SHapley Additive exPlanations) values find application in elucidating the outcomes of machine learning models by assigning contributions to each input feature that led to a prediction. Through the computation of SHAP values for the classification model, an assessment can be made regarding the predominant impact of individual features on the model's predictions within each cluster. This analytical process unveils the pivotal features that significantly contribute to the differentiation of clusters. SHAP feature importance provide a means to comprehend the primary drivers behind the prediction of each cluster. Consequently, this analysis aids in determining whether the clusters are shaped by a diverse array of features or if they adhere to an overly simplistic characterization [26].

Fig 3.6: A SHAP summary plot ranks variables by feature importance (example)



Source: https://medium.com/bricklane-tech

Chapter 4
Methodology & Procedures

In this section, we delve into the actions and techniques implemented to achieve the successful experimentation. We illustrate the suggested methodology and the procedure of gathering data. Here, we examine the process of creating the dataset by carefully selecting appropriate questions from the NatCen British Social Attitude Survey. Subsequently, we delve into the clustering methodology that was chosen for conducting the studies. Finally, we take deep dive into the evaluation of the generated clusters.

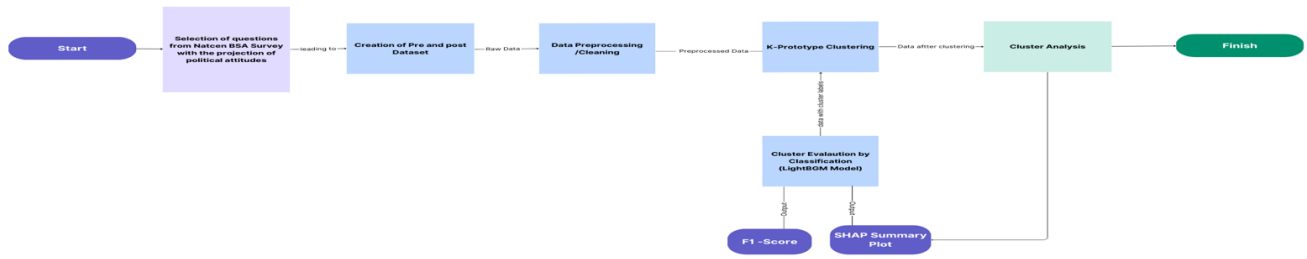
4.1. Programming Environment

The experimentation took place within the Google Colaboratory environment, utilizing Python 3.6, a highly utilized programming language known for its user-friendly nature and compatibility with a diverse range of integrated libraries. To execute the cluster analysis followed by evaluation, numerous Python libraries, such as Pandas, Matplotlib, UMAP, NumPy, Kmodes, shap, and lgbm were employed. The project's hardware setup involves a MacBook Air running macOS and powered by the innovative Apple M1 chip. This chip incorporates an 8-core CPU and an 8-core GPU, alongside a unified memory of 8 GB. This combination of hardware resources was leveraged to ensure efficient and powerful processing for the study.

4.2. Approach

The adopted approach unfolds in four distinct stages. In the initial stage, emphasis is placed on shaping the dataset. This involves careful selection of relevant questions from the British Social Attitude Survey, administered by NatCen, that effectively capture the nuances of political attitudes. The selection of survey questions is then well elaborated in the following section 4.3. This endeavor culminates in the creation of a feature set essential for subsequent analysis. Notably, the temporal framework considers the years 2018 and 2019 as the pre-Brexit era, a period unaffected by Brexit's impact. In contrast, the years 2020 and 2021 denote the post-Brexit epoch, marked by the confluence of both Brexit as well as the COVID-19 pandemic. The selected years marks the time-period when the Brexit was the peak point of discussions. The years were selected to ensure data compatibility as well as to ensure the need of having the same set of questions for all the years. The survey data corresponding to 2022 has not been released hence, we use the most recent available data after Brexit to create the Pre period which include 2020 & 2021. Importantly, it's crucial to recognize that these two events occurred simultaneously, leading to the intertwined impact of both on the post-Brexit period. While Brexit was anticipated, the advent of COVID-19 was unforeseen. The subsequent stage involves the application of K-Prototype Clustering to both datasets separately which are provided after the necessary data cleaning procedures. Following this, the third stage centers on the validation of the derived clusters through a Classification methodology. Finally, the fourth stage entails an in-depth analysis of the resultant clusters. This phase encompasses the computation of similarity scores to unveil the transformations within clusters between the pre and post periods, offering insights into the evolving landscape of political attitudes.

Fig 4.1: Block Diagram of the Experiment



4.3. Data Sets

Data plays a critical role in the process of cluster analysis. In this experiment, we use the survey data popularly known as British Social Attitudes conducted by NatCen social Research [27]. NatCen Social Research organizes a survey called British Social Attitude Survey annually incorporating survey questions including all relevant issues of the specific year. It is notable to know that the survey corresponding to particular year primarily discusses the relevant issues related to that time-period. However, some of the questions are repeated annually. The survey includes numerous questions that which covers various aspects of the attitudes of the British People. The Natcen opinion panel includes people from various parts of UK including England, Scotland, Wales, and Northern Ireland. These people share their experiences on variety of events and are managed by NatCen. The survey consists of question with multiple responses from which the responder can choose from. The responses include all possible answers along with a choice that which allows the responder to skip the question. The data corresponding to this survey is retrieved through the platform called UK Data Services [28].

4.3.1. Feature Selection

The Selection of questions have prime significance as it serves a major role in the commencement of this research endeavor. The study conducted by Professor Reimer Reinsman, which delves into the correlation between personalities, attitudes and politics gives an intriguing insight into how the political attitude can be defined as well as identified [8]. He mainly conducted this research by analyzing various events that occurred during the period of his study and developed attitudes based on the respective responses of people towards these events. Similarly, the study conducted by the Joeffrey Evans [9] on evaluating the political attitudes based left-right scale as well as libertarian-authoritarian scale is notable as his study incorporates multiple questions from the British social attitude survey. For instance, the research paper talks about statements such as “There is one law for rich and another for poor” [27] which mainly focus on inequality as well as exploitation. The study considers this as political attitude questions as it influences an individual’s long term political perspectives. Collectively, the two papers give insightful information regarding the selection of questions that which reflects the political attitudes. Hence, these two papers act as significant benchmarks in guiding the curation of questions.

The created dataset involves demographic variable which describes as well as categorize the individuals followed by set of event-based questions that which encompasses Brexit as well as covid events. These questions discuss the impact of these events. Finally, dataset encompasses political attitude questions which are prominently ordinal in nature. The selected questions for the analysis are listed below,

Table 4.1: Demographic features

Survey Questions	**Variables**
Sex of respondent	RSex
Respondent's religion	ReligSum20
What is your main source of income at present?	maininc5
Do you think of yourself as a little closer to one political party than to the others?	ClosePty
Generally speaking, do you think of yourself as a supporter of any one political party?	SupParty
political party identification	PartyIDN

Table 4.2: Event Specific questions (EU referendum & COVID-19)

Survey Questions	**Variables**
Do you think Britain's long-term policy should be towards EU	ECPolicy
Did you manage to vote in the 2016 referendum about the European Union?	EURefV2
If you were given the chance to vote again in the referendum, how would you vote ?	EURefb
Because of the Covid-19 pandemic, I would like to change jobs	CovWk1
Because of the Covid-19 pandemic, I would like to stop working	CovWk2
Because of the Covid-19 pandemic, I would like to reduce my working hours	CovWk3

Table 4.3: Political attitude questions

Survey Questions	**Variables**
How much interest do you have in politics?	Politics
If it had to choose, should govt reduce/increase/maintain levels of taxation and spending?	TAXSPEND
Opinions differ about the level of benefits for unemployed people. Which comes closest to your own view	dole
Most people on the dole are fiddling in one way or another	DoleFidl
Many people who get social security don't really deserve any help	SocHelp
The welfare state encourages people to stop helping each other	welfhelp
Cutting benefits would damage too many people's lives	damlives
The creation of the welfare state is one of Britain's proudest achievements	proudwlf
Government should redistribute income from the better-off to less well-off	redistrb
There is one law for the rich and one for the poor	RichLaw
Ordinary working people do not get their fair share of the nation's wealth	Wealth
Management will always try to get the better of employees if it gets the chance	Indust4
Big business benefits owners at the expense of workers	BigBusnn
For some crimes, the death penalty is the most appropriate sentence	DeathApp
The law should always be obeyed, even if a particular law is wrong	WrongLaw
Censorship of films and magazines is necessary to uphold moral standards	censor
Young people today don't have enough respect for traditional British values	tradvals
People who break the law should be given stiffer sentences	StifSent
How satisfied or dissatisfied are you with the way the National Health Service runs nowadays	NHSSat

The tables 4.1,4.2,4.3 gives an elaborative picture of all the selected questions/features and related variables. The dataset is then crafted by the integration of all the selected features. The dataset can be categorized into two distinct sets, namely Pre-Brexit data and post-Brexit data. The Pre Brexit data is formed by combining the survey data from 2018 and 2019 comprising of 4,535 survey responses whereas the post-Brexit data includes the data from 2020 and 2021 comprising of 3,882 survey responses. It is noteworthy to know that the surveyed group of individuals varies, thereby provides with a resourceful collection of diverse survey responses.

4.4. Clustering Algorithm

The newly created datasets are then preprocessed. The datasets then undergo dimensionality reduction using the UMAP dimensionality reduction technique. It is a dimensionality reduction technique useful for visualizing and analyzing large datasets with numerous features. It preserves the structure in the data and helps to identify patterns. The algorithm represents high dimensional data in low dimensional space while maintaining the data structure. Here, I have used UMAP to represent my dataset in a two-dimensional space. This algorithm projects the original data into a lower dimensional space such that each data point is represented by a pair of values in a two-dimensional space. Here we deal with a dataset created by selecting questions from a large survey data hence this provides a visual aid into the structures and patterns of the created datasets from the respective survey dataset. Scatter plots are used as visual representation of the datapoints in the 2-dimensional space allowing to observe any patterns, clusters, or grouping that might exist in the data.

The K-prototype clustering is a type of K means clustering Hence it is significant to determine the value of K. K stands for the number of clusters. The optimal value of k is then obtained for both datasets by employing Elbow method. This graphical representation will then provide an insight into the optimal

number of clusters. Here we plot a graph and obtain the point at which it represents a trade-off between minimizing the cost and prevent overfitting. It is then followed by the determination Silhouette Score Coefficient to check if the K values are accurate such that higher the coefficient value, optimal the cluster solution.
The datasets mainly comprise of categorical and ordinal data. Therefore, Once the K value is determined, the K-prototype algorithm is then applied to the two the datasets.

4.5. Performance Evaluation

Once the clusters are generated, it is important to interpret the resulting clusters. Therefore, in this study, we adopt the methodology of evaluation of the generated clusters through a classification model.
In the given study, the generated clusters are treated as labels followed by the development of a classification model on top. If the classification model predicts the cluster labels with high accuracy, then the clusters are of high quality. This approach hinges on premise that the cluster with high quality provides perfect predictive performances in the classification model.
In the given study, it is beneficial to adopt the LightGBM classification model as the classifier as it excels in handling both categorical and numerical features. It does not require to perform one-hot-encode on categorical features as it may lead to sparse data. LightGBM also provides built-in support for the calculation of SHAP values. Hence the shap values can be obtained for each instance in the datasets.

4.5.1. Cross-Validated F1 -Score

Cross-validation is typically applied to avoid overfitting as well as to ensure robustness. A higher value for the cross-validated F1-score suggests that that the clusters are distinct and well separated. That is, the employed classification model (LightGBM) is successful in recognizing the inherent differences between the clusters based on the provided features.

4.5.2. SHAP Feature Importance

Once the classification model is fit on the dataset with the cluster labels as target variables, a SHAP value analysis is then performed to understand the cluster characteristics.
The SHAP analysis then rank the features based on the feature importance and illustrate the relationship between the value of a feature and its impact on the cluster predictions. The contribution of a feature is computed by the difference that it brings to the final predicted value [].

Once the classification model is fit, the impact of the features on the prediction of each cluster is then visualized by a Summary plot. The generated graph includes the features of importance on the y -axis such that the feature with high contributions at the top and magnitude of the average impact on model on the x-axis. A feature is represented with colored bar where each color corresponds to a specific cluster showing the importance of the feature in predicting the respective clusters. Here the features are ordered as per their importance.

4.6. Cluster Analysis

After the development of the SHAP summary plot, it provides information regarding the important features as well as their contribution towards the clusters. It is significant to check if the generated cluster solutions are similar in nature. In this study, our primary goal is to determine whether the clustering solutions generated are similar in nature or not which in turn gives an insight into the changes in the dynamics of the political attitude landscape with respect to the pre and post time periods. The process of similarity check includes the computation of correlation matrices for each cluster separately for both the datasets. A detailed visual inspection of the generated correlation matrix is then adopted to check for the similarity. A similar cluster solution would imply that the solutions yield similar patterns of correlation among variables. If the identified clusters show high degree of similarity in the correlation structure, it indicates robustness.

Furthermore, the study also adopts a methodology to analyze the political landscape involving the clusters formed by the two datasets known as **Cluster Profile comparison**. The created datasets comprise of moderately similar amounts of data (4,535:3,882). Therefore, it is beneficial to employ cluster profiling. It includes the calculation of average values of features with political attitudes. The

computed average values are the compared across clusters in both datasets. This provides a sense of how the political attitudes differ between clusters within each datasets allowing to gain insights into how political attitudes differ between clusters within each dataset. The survey questions with political attitude mainly comprises of ordinal data. The political attitude feature set comprises of ordinal variables ranging from 1 to 5 such that 1 stands for “Strongly Agree”, 2 for “Agree”, 3 for “Neutral”, 4 for “Strongly Disagree” and 5 stands for “Disagree”. Due to the ordinal nature of this feature set, the study adopts the mode computation for each ordinal variable. This computation provides an insight into the most common political attitude value for each feature within the cluster. The resulting values can then be employed to understand the changes in the political attitudes.

For instance, consider the value of profile difference for a feature as 0, this indicates that the average values are the same for that cluster in both datasets. This can be interpreted as high level of similarity. Furthermore, if the value of the profile difference is a -1, it indicates that the average value of the specific feature in the post dataset is one unit lower than in the pre dataset for the specific cluster. This indicates a shift in political attitude of the respondents within a cluster over time, which leads to a more negative or lower stances on the specific issue represented by the variable. Finally, if the profile difference value is 1, it indicates that the average value of that variable in one cluster of the post dataset is one unit higher than the average value in the corresponding cluster in the pre dataset. That is there has been an increase in the average value from pre to post dataset for the clusters.

Chapter 5
Results and Discussions

This chapter covers and discuss the results acquired by the implementation of the proposed approaches and methodologies described in Chapter 3.

All the analysis were carried out in Google Collaboratory environment on the MacOS. The experiments were performed by creating two datasets by dividing the timeframe into Pre and Post Brexit time periods. It is then followed by the implementation of k prototype clustering on the two datasets respectively. The generated clusters are then evaluated with classification model followed by similarity analysis for the two datasets.

For the given datasets we perform the necessary preprocessing steps including the removal of null values followed by the removal of survey responses that include responses that do not provide any required responses. For instance, in case of ordinal variables -1 stands for "Not answered" this response may not be required for the analysis purposes. Hence, we remove such responses from the datasets.

In later stage we perform the UMAP Dimensionality reduction technique, this projects data in a two-dimensional space such that the two axis corresponds to the two dimensions of the embeddings generated by the UMAP algorithm.

Fig 5.1: Pre data set

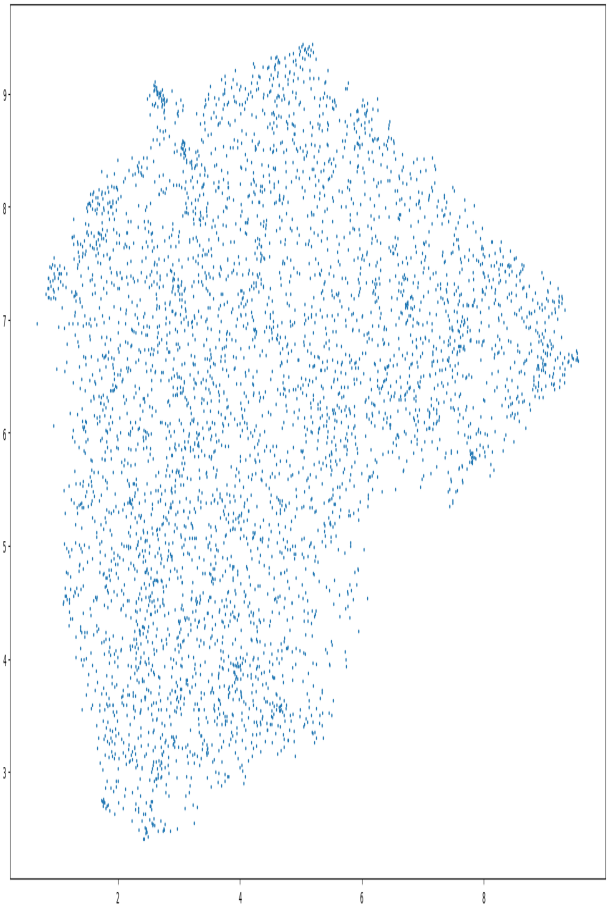
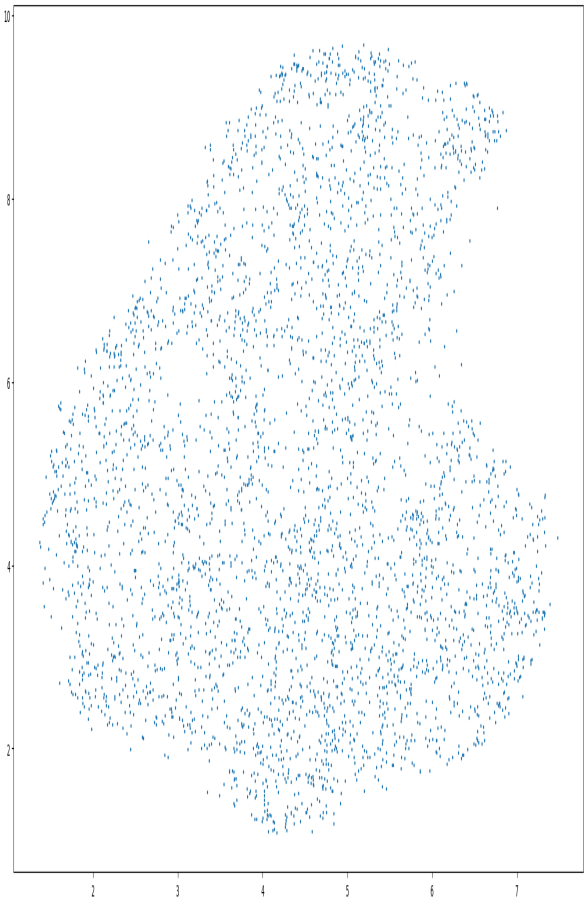


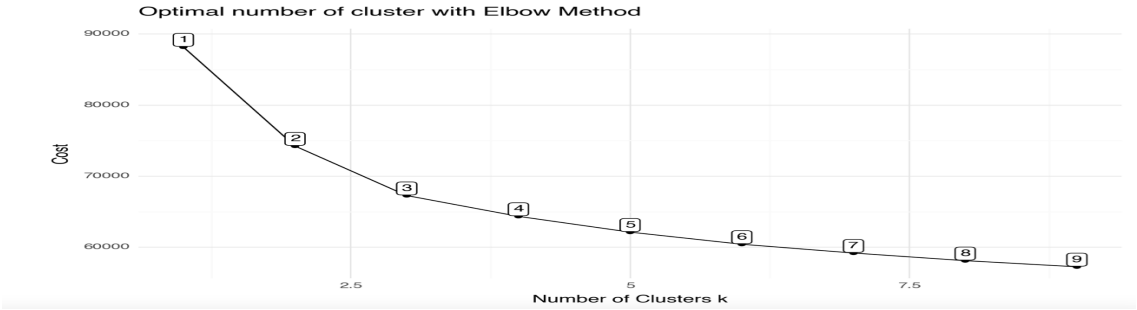
Fig 5.2: Post dataset



The above figures provide the scatter plot representations of datasets in a two-dimensional space implemented through UMAP algorithm. In the obtained plots, it is visible that datapoints are strongly close to one another. This may be because the data are naturally concentrated or may be because the data has inherent patterns or groupings that may be well defined and closer to one another.

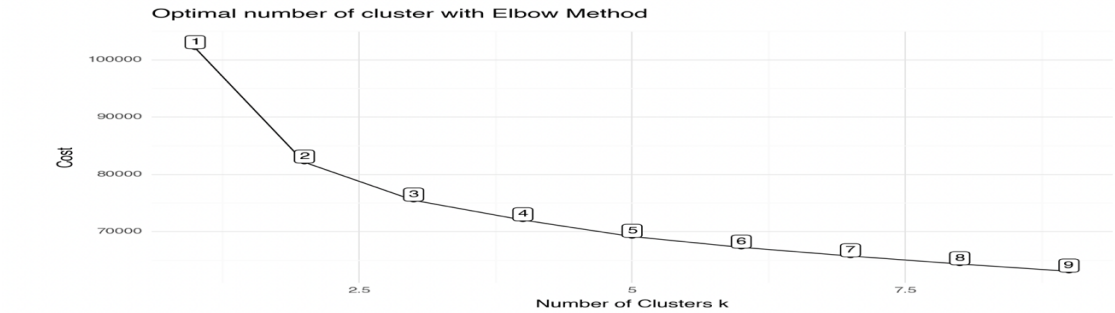
Later on, it is essential to determine the optimal value of **K** for the successful commencement of this analysis. We determine the optimal value of **K** by executing the Elbow method on the two datasets. The output provides a graphical representation as shown below,

Fig 5.3: Pre-Brexit Dataset: Elbow method graph



It is visually evident that the graph starts to level-off after the implementation of 3 as **K**. Hence **K** may have values either 2 or 3 as the optimal value for the **K** as per the Elbow method computation.
Similarly, the optimal value of **K** for the post Brexit dataset is implemented by the computation of elbow method in the similar way. The graphical representation of the optimal value of **K** for the post Brexit data is given below,

Fig 5.4: Post Brexit dataset: Elbow method graph



The graph above illustrates that the cost value level-off after the value of number of clusters is 3. Therefore, **K** can either be 2 or 3 as the optimal value of clusters for the post Brexit data.

Table 5.1: Optimal number of cluster computation by Elbow method

Pre Brexit dataset	Post Brexit dataset
3/2	3/2

However, **K** can either be 2 or 3. Selection of the optimal value for **K** is crucial for this analysis. Hence, we adopt the method of silhouette score to find the optimal value for **K** for the two datasets.
The obtained silhouette score for the Pre and Post datasets are given below,

Table 5.2: Silhouette Score coefficient values

Datasets	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9
Pre-dataset	0.111	0.085	0.060	0.050	0.046	0.042	0.039	0.033
Post-dataset	0.177	0.127	0.076	0.068	0.064	0.049	0.047	0.044

From the computed Silhouette scores for the pre and post datasets, it is visible that k=2 has the highest Silhouette score. Hence the optimal value for **K** is computed as **2**.

Once the K value is obtained, K-prototype algorithm is implemented on both datasets. This results in the formation of two clusters for both datasets. The resulting cluster solutions are listed below,

Fig 5.5: Cluster solution for pre dataset

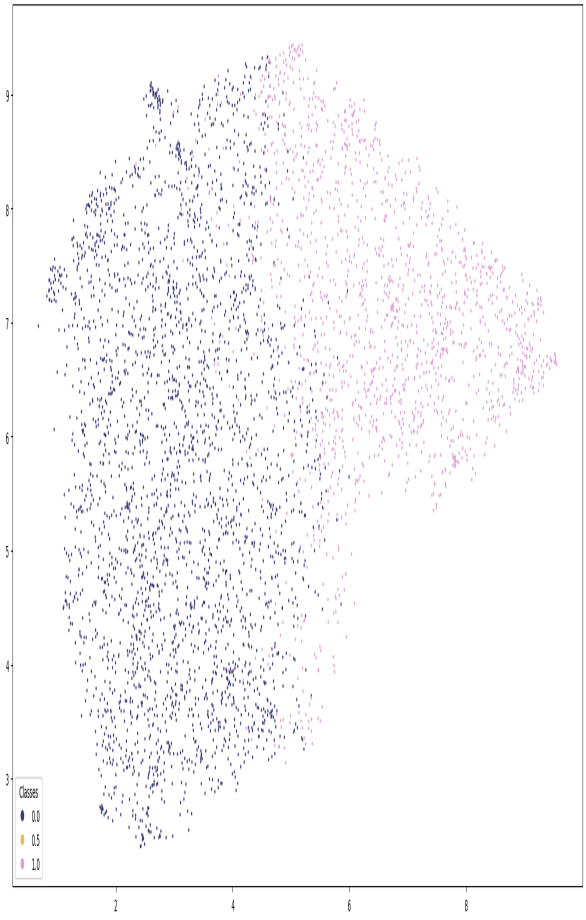
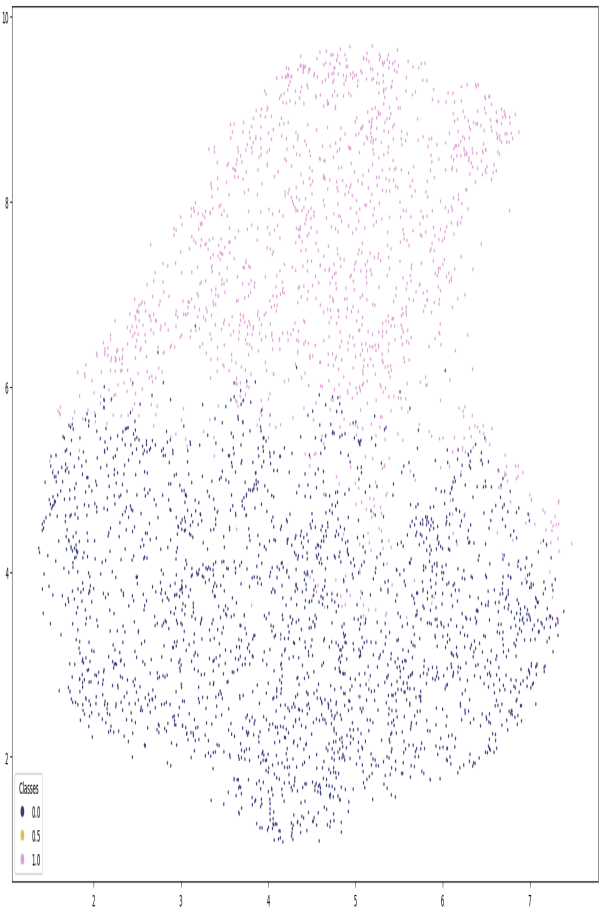


Fig 5.6: Cluster solution for post dataset



The above figures provide the scatter plot representation of the cluster solutions for the Pre and post datasets in a two-dimensional space. The predata consists of two clusters with each cluster containing a set 2886 and 1648 entries respectively. Based on the obtained cluster centroids, cluster 0 mainly consists of conservatives whereas the cluster 1 mainly consists of Labor. Likewise, in case of post Brexit dataset, the obtained two clusters contains a set of 2329 and 1551 entries respectively such that cluster 0 mainly consists of party identification as conservative whereas cluster 1 consists of Labour. Similarly, in both cluster solutions the people in cluster 0 believes that the tax must not be increased and spend on services whereas in cluster 1 believes that the tax must be increased and spend on health and education. This trend of similarity is evident across the cluster solution.

The second stage of the study involves the evaluation of the generated clusters. The evaluation of the generated clusters is performed with the application of a classification model (LightGBM) as explained in Chapter 3. Here, we evaluate the clusters by considering the clusters as labels and build the LightGBM classification model on top. If the generated clusters are of high quality, then classification model will be able to predict them with high accuracy. The generated clusters are evaluated by analyzing the distinctiveness of the clusters as well as the informativeness of the clusters.

The distinctiveness of the clusters is determined by computing the cross-validated F1-score. It is noteworthy that higher the CV score, meaningful and distinguishable the clusters are.

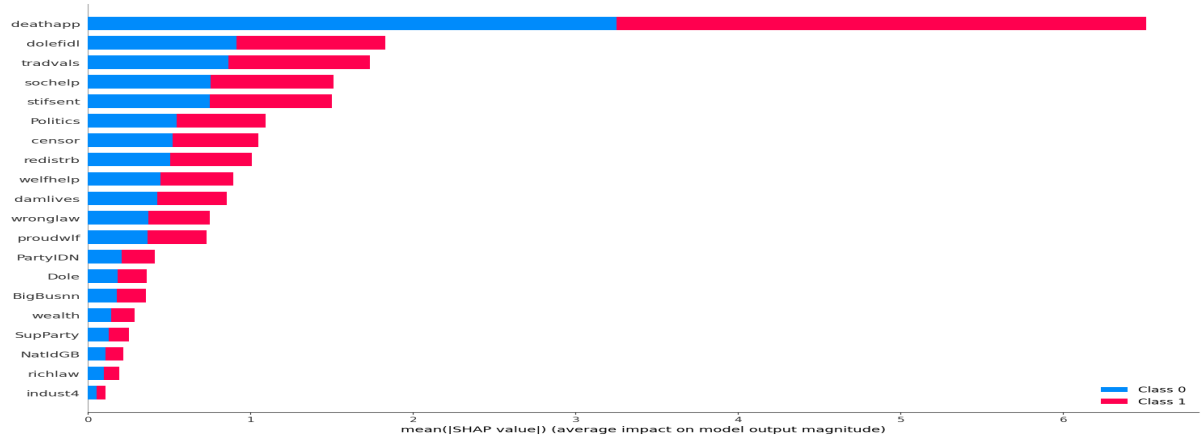
Table 5.3: Cross Validated F1 Score

Datasets	CV Score
Pre Brexit-Dataset	0.963
Post Brexit Dataset	0.972

The obtained CV score for both datasets make it evident the datapoints are grouped into distinguishable as well as meaningful clusters for both the datasets. This also justifies that the selection of k was an appropriate choice.

The informativeness of the clusters can determined by the Shap feature importance as discussed in the previous chapters. The summary plot provides an insight into the features that were important to classify the K-prototype clusters [26]. The questions corresponding to the variables are mentioned in the section 4.3.1.

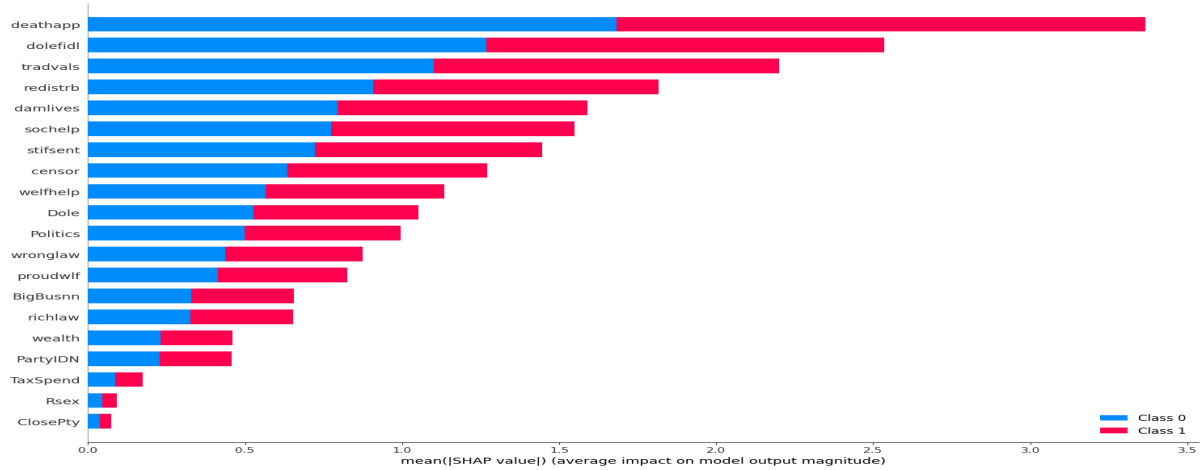
Fig 5.7: Summary plot of pre dataset



The above figure illustrates the important features in classifying the K-prototype clusters for the pre dataset. The plot summaries that 12 of the features have significant impact in cluster label classification.

Similarly, the summary plot for the post data set is also generated as well as important features are determined through the same.

Fig 5.8: Summary plot for post dataset



The above figure illustrates the important features in classifying clusters for post dataset. As per the SHAP plot, it is notable to know that here a total of 17 features has strong impact to classify the cluster labels correctly using LightGBM model.

The above SHAP summary plots provide an insight into the important features in classifying the clusters. It is intriguing that the ordinal features played major role in the classification process than most of the other categorical features. The summary plots provide an interesting image of features that can selected for further analysis of understanding the similarities as well as dissimilarities of the clusters formed between the two sets of time periods. The top of both the summary plots are constituted by ordinal variables. In both cases, "deathapp" which stands for "death penalty is the appropriate sentence" has the highest magnitude. It is notable that the first three features constituting the top of plots are similar. Likewise, the variable "redisturb" takes the 7th place in the pre dataset whereas it takes the 4th place in post dataset. The categorical variable such as "PartyIDN" can be seen in both datasets. The categorical variables such as partyIDN, closepty, taxspend, supparty have lower magnitude compared to the ordinal variable constituting the top of the SHAP plots.

Finally, the correlation matrices are developed to check for the similarities between the clusters by selecting the political attitude-based features that contributed strongly to classify the clusters from the summary plot. This includes the political attitude questions as they played as milestones in the clustering process.

Fig 5.9: Correlation matrices for pre dataset

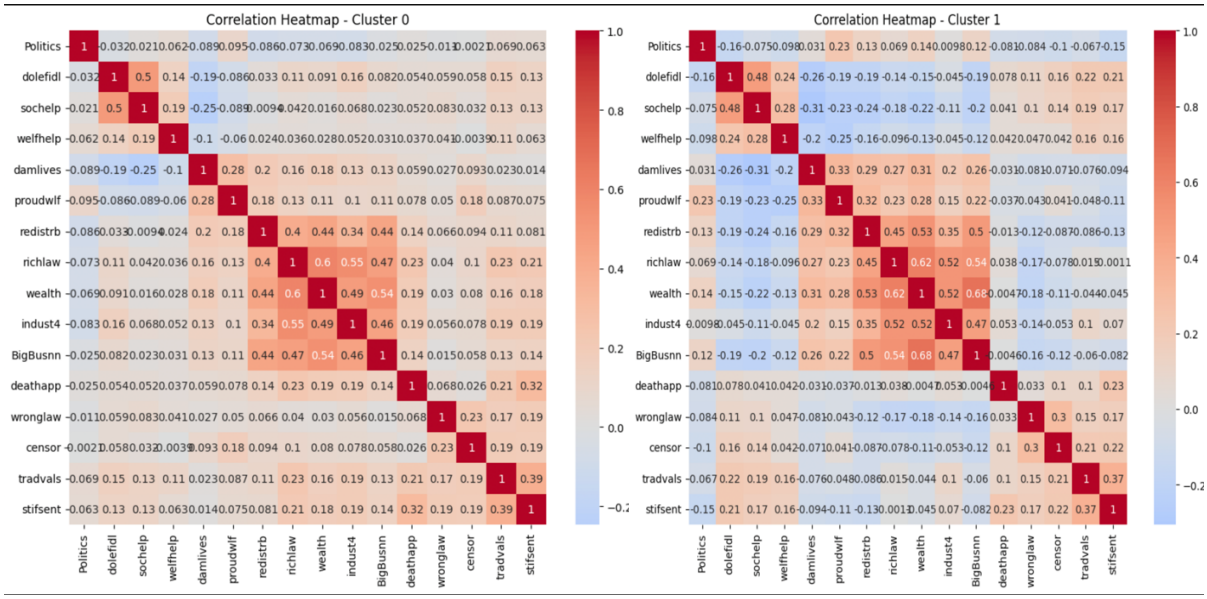
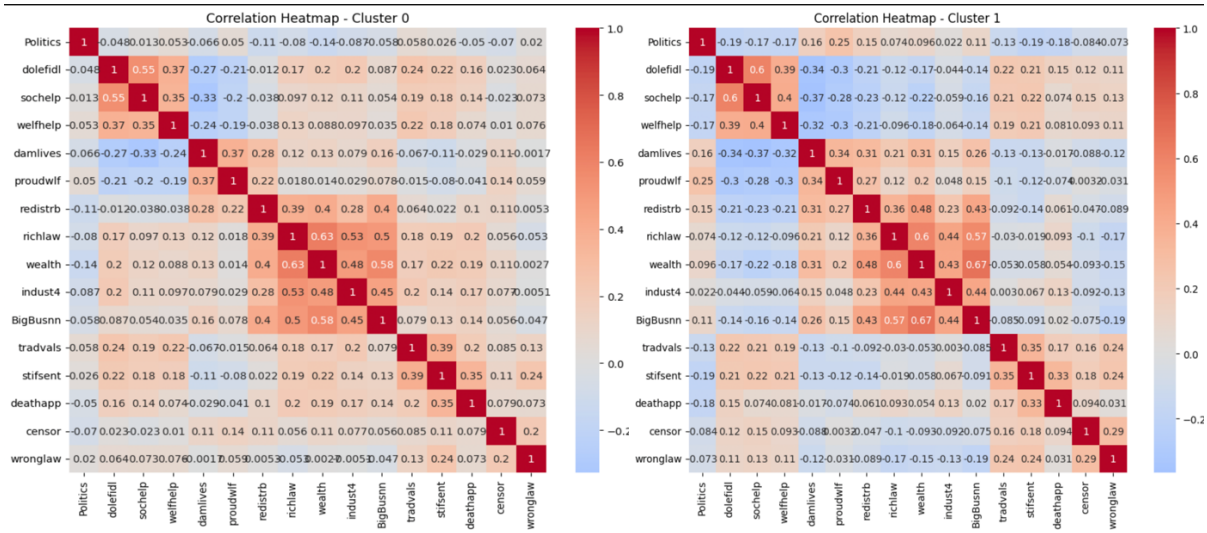


Fig 5.10: Correlation matrices for post dataset



The above heatmap representation of the correlation matrices shows that the clusters provide an approximate similar pattern. The respective clusters of the pre dataset showcase patterns that are very similar to the corresponding clusters in the post dataset respectively. This indicates that there was no reasonable change in the political attitude among the two datasets. Consider the cluster 0 for the two datasets, the variables redisturb, rich law, wealth, indust4, BigBusnn demonstrate strong positive correlations in both the datasets. A similar trend is observable in cluster 1 for both datasets. It is interesting to know that that variable Politics has the least correlation among all the variables in both cluster solutions. For instance, the variable redisurb which stands for government should redistribute income is positively correlated with richlaw (one law for rich another for poor), wealth (working people don't get their fair share of nation wealth), bigbusnn (big business benefit over workers). This trend is visible in cluster 0 of both datasets. Likewise, such similar characteristics are evident in the cluster 2 of both datasets.

The cluster profile difference implementation is a methodology that which computes the average value of political attitude features followed by the comparison of these features across the clusters. The values are computed based on the mode value as the feature set are of ordinal nature. Here we compare the clusters between the two datasets by computing the mode as we find the most common answer within each cluster. Here the ordinal variables are ranked from 1 to 5 such that 1 stands for Strongly agree,2 for agree, 3 for neutral etc. For instance, for variable in pre dataset in a cluster, the most common response may be "strongly agree" but for the next dataset, the most common response may be "agree". This indicates a shift in response.

Taking this into consideration, consider a cluster Z with variable X for two datasets pre and post. If the cluster profile difference for Z is 0 for X then the most common response in X is similar in both datasets. However, if it is -1, then the response in the cluster Z for post is 1 unit lower than pre such that if the most common response in preset was "agree" whereas in post set it is shifted to "neutral". Likewise, if it is 1 then the variable in cluster Z for post is one level higher than in predata, such that the most common political attitude in cluster Z for variable X was "neutral" for predata whereas it has shifted to "Agree" in post data. The survey questions corresponding to variables are mentioned in section 4.3.1.

Table 5.4: Cluster profile difference

Feature	Cluster 0	Cluster 1
welfhelp	0	0
sochelp	0	0
dolefidl	0	0
damlives	-1	0
proudwlf	-1	1
redistrb	-1	0
BigBusnn	0	1
wealth	0	1
richlaw	0	0
tradvals	0	-1
stifsent	0	0
deathapp	0	0
censor	0	-3
wronglaw	0	0
Politics	0	0
indust4	0	0

Based on the obtained cluster profile differences for Cluster 0, the profile difference values corresponding to the variables mostly constitute of 0 which indicates there has not been a change for the respective variables. However, the variables such as damlives, proudwelf, redisturb give a -1

indicating a shift in most common response such that it is one level lower than post dataset. Furthermore, in cluster 1, most of the variables gives a difference of 0. It also gives a difference of -3 for the variable censor which indicate a significant shift. It is evident that there has been shifts corresponding to cluster 1 whereas the cluster 0 is mostly similar in nature. Hence it is evident that cluster solutions illustrate an approximately similar pattern for cluster 0 whereas showcases shifts in cluster 1 across the two datasets.

From the obtained cluster solutions for both datasets, the cluster 1 mainly consists of responses with party identification as Labours, whereas cluster 0 consists of conservatives in both datasets. This indicates the trend of responses across the clusters. The cluster 0 with conservative party identification showcased strong similarity across the datasets whereas the cluster 1 with Labour party identification provided responses that which showcased shifts in responses for various political attitude questions across the two datasets.

5.1. Evaluation and Limitations

The quality of the dataset plays a major role in this experimentation. The selection of questions does play a major role as it is the critical milestone. During the research, the study dealt with the difficulty of finding survey questions that were available in both the time periods as the survey questions were more related to the relevant issues of the specific time-period. Hence, this limited the inclusion of wider perspectives in this analysis. It is also noteworthy to note that the nature of the data depicts the limited knowledge of respondents which led to the lack of responses for certain questions. Furthermore, the size of the data in both datasets are balanced. However, the availability of more data would have accelerated this analysis. For instance, the responses received related to the EU referendum were quite low. This may be due to the lack domain knowledge from the responders.

During the cluster analysis, the cluster solution indicates that the ordinal features played stronger role in the clustering process. The SHAP summary plot indicates more importance for ordinal features whereas the categorical questions including the party identification and the income inequalities are ranked low.

Chapter 6
Conclusion

Political attitudes are defined as the approach of citizens towards various perspectives of their social, economic as well as political dimensions. This research has taken a deep dive to understand the intriguing relationship between events and their impact on political attitudes. In the given study we chose the event of Brexit and COVID-19 as a set of events that played transformative role across the various aspects of British life. The two events had strong impact among the lives of British people due to their divisive nature and effect on the day to day lives of Britishers encompassing socio-economic as well as political perspectives.

The study commenced with the development of a dataset by the selection of appropriate survey questions from the British Social Attitude Survey organized by NatCen Social Research. It is then followed by the development of two datasets marking the pre and post Brexit era. It is noteworthy to know that post Brexit era is impacted by both covid as well as Brexit as the two events happened concurrently. Hence, a single dataset may reflect both the events. The prebrexit era encompasses the survey data from 2018 and 2019 whereas the post Brexit era include 2020 and 2021.The cluster solutions were then developed by the implementation of K-prototype clustering algorithm. We also evaluated the clusters by a tree-based classification model, LightGBM. The cluster solution generated were meaningful and distinguishable as they gave a cross-validated F1- score of approximately 0.94. It is then followed by the implementation of SHAP summary plot into determine the importance of features with respect to the clustering solutions. Finally, an analysis is done to determine if the cluster solutions were similar or not. The correlation matrices were plotted to determine the similarity between the clusters by the analysis of survey question with political attitudes. The developed correlation matrices for the clusters of the prebrexit time period were approximately similar to their respective counterparts in the dataset of post Brexit time period. This indicates that the cluster solutions were approximately similar and there has not been a significant impact on the political attitude of the people by the events of Brexit and Covid. Hence, based on the received survey responses of the people, there has not been a significant change in the political attitudes of the people.

6.1. Future Work

Furthermore, an investigation into the identified clusters to determine the changes in political attitudes can be performed. This involves a shift analysis in individual attitudes within the clusters before as well as after the events. This helps in identifying the micro-level changes that might be hidden in the broader cluster analysis. Inclusion of predictive models to understand the future trends can also be a viable thread for future scope.

The establishment of clearer causal link between the events and attitudes changes by the employment of advanced casual inference techniques such as prosperity score matching, or regression discontinuity analysis can also be of an ideal future scope. This helps to address potential features and strengthen the conclusions drawn. Exploration of various segmentation techniques such as latent class analysis or factor analysis could provide intriguing insights into the underlying structures.

Furthermore, an analysis into the social media data in conjunction with the survey data responses could offer a real-time perspective on how events potentially impact the political attitudes. Natural Language Techniques can be incorporated to extract and analyze the relevant contents. Incorporation of additional features such as media coverage or economic indicators may provide a more holistic approach in analyzing the political attitude influence. Finally, the conducted study can be validated using different datasets and methodologies. This helps in the enhancement of the robustness of resulting conclusions as well as reinforcing the generalizability of the respective conclusion

References

[1] Semetko, H. A., van der Brug, W., & Patti M. Valkenburg. (2003). The Influence of Political Events on Attitudes towards the European Union. *British Journal of Political Science*, *33*(4), 621–634. http://www.jstor.org/stable/4092199

[2] Mariya, Arix, 2017, https://www.brookings.edu/

[3] Gupta R, Hasan MM, Islam SZ, Yasmin T, Uddin J. Evaluating the Brexit and COVID-19's influence on the UK economy: A data analysis. PLoS One. 2023 Jun 15;18(6):e0287342. doi: 10.1371/journal.pone.0287342. PMID: 37319267; PMCID: PMC10270588.

[4] Applebaum, Anne. "Britain After Brexit: A Transformed Political Landscape". Journal of Democracy, vol. 28, no. 1, Jan. 2017, pp. 53-58.

[5] Euronews,2022, https://www.euronews.com/my-europe/

[6] Mustafa, G., Hussain, M., & Aslam, M. A. (2020). Political and Economic Impacts of Brexit on European Union. Liberal Arts and Social Sciences International Journal (LASSIJ), 4(2), 11–23. https://doi.org/10.47264/idea.lassij/4.2.2

[7] HENNESSY, B. (1970). A HEADNOTE ON THE EXISTENCE AND STUDY OF POLITICAL ATTITUDES. Social Science Quarterly, 51(3), 463–476. http://www.jstor.org/stable/42858636

[8] Riemann, Rainer & Grubich, Claudia & Hempel, Susanne & Mergl, Susanne & Richter, Manfred. (1993). Personality and Attitudes towards Current Political Topics. Personality and Individual Differences. 15. 313-321. 10.1016/0191-8869(93)90222-O.

[9] Evans, G., Heath, A., & Lalljee, M. (1996). Measuring Left-Right and Libertarian-Authoritarian Values in the British Electorate. *The British Journal of Sociology*, *47*(1), 93–112. https://doi.org/10.2307/591118

[10] Fleishman, John A., Types of Political Attitude Structure: Results of a Cluster Analysis, https://doi.org/10.1086/268990

[11] Hyehyun,2012, Public Segmentation and Government–Public Relationship Building: A Cluster Analysis of Publics in the United States and 19 European Countries, https://www.tandfonline.com/doi/epdf/

[12] Stattin, H., Amnå, E., 2022, Basic Values Transform Political Interest into Diverse Political Value, https://doi.org/10.1007/s10964-022-01654-w

[13] Joseph, Lindell, 2020, UK space of political attitude after Brexit referendum, https://journals.sagepub.com/doi/pdf/10.1177/1360780420965982

[14] GAL-TAN Scale, https://www.mobergpublications.se/continued/scale.htm

[15] Left-Right Scale, https://www.econstor.eu

[16] Lisa Windsteiger, William Sas, 2020, Covid-19 and social political attitudes. https://cepr.org/voxeu/columns/covid-19-and-socio-political-attitudes-europe-c

[17] Willi Klösgen and Jan M. Zytkow (Eds.). 2002. Handbook of data mining and knowledge discovery. Oxford University Press, Inc., USA.

[18] Anderberg, Michael R. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. Vol. 19. Academic press, 2014.

[19] ZHEXUE HUANG , 1998, Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, https://citeseerx.ist.psu.edu/

[20] Chaturvedi, Anil & Green, Paul & Caroll, J.. (2001). K-modes Clustering. Journal of Classification. 18. 35-55. 10.1007/s00357-001-0004-3.

[21] Aditya,2022,https://codinginfinite.com/k-prototypes-clustering

[22] Umargono, Edy & Suseno, Jatmiko & Gunawan, S.K. (2020). K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. 10.2991/assehr.k.201010.019.

[23] Silhouttescore, https://codinginfinite.com/silhouette-coefficient

[24] Ankita. 2023, https://www.analyticsvidhya.com/

[25] Shreyan, https://www.geeksforgeeks.org/lightgbm

[26] Veronica, 2021,https://medium.com/bricklane-tech/a-new-approach

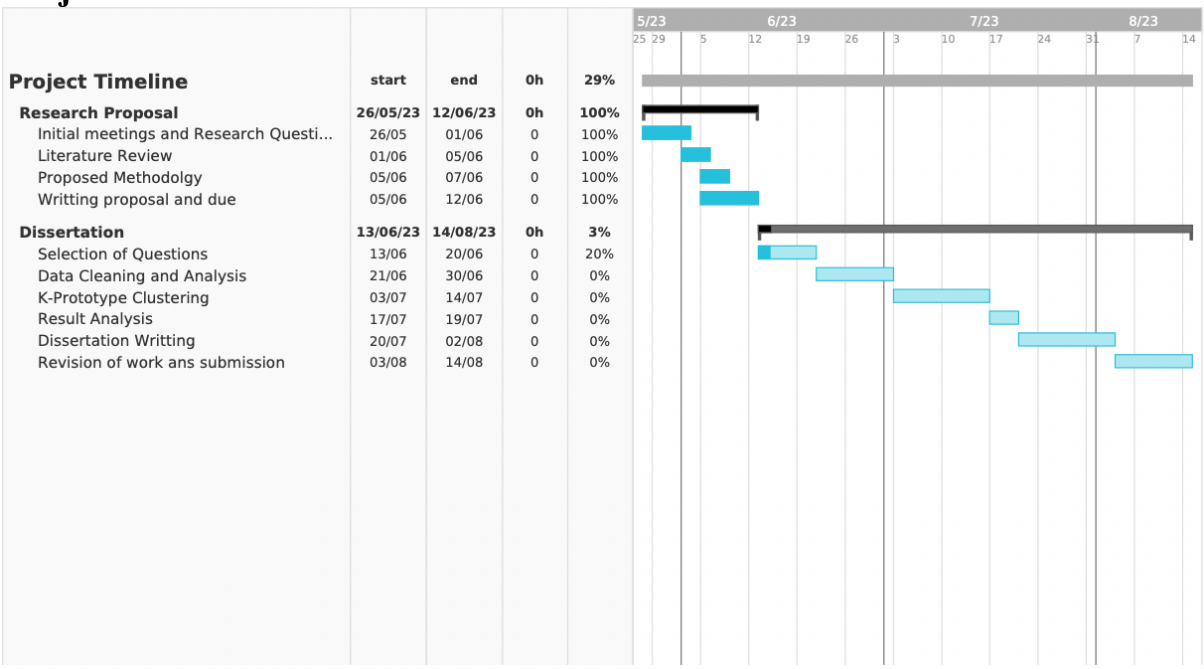
[27] NatCen Social Research, https://bsa.natcen.ac.uk/

[28] UK Data Services, https://ukdataservice.ac.uk/

[29] Isaac Quaye, 2016, https://scirp.org/journal/

Appendix

Project Plan



Project Timeline	start	end	0h	29%
Research Proposal	**26/05/23**	**12/06/23**	**0h**	**100%**
Initial meetings and Research Questi...	26/05	01/06	0	100%
Literature Review	01/06	05/06	0	100%
Proposed Methodolgy	05/06	07/06	0	100%
Writting proposal and due	05/06	12/06	0	100%
Dissertation	**13/06/23**	**14/08/23**	**0h**	**3%**
Selection of Questions	13/06	20/06	0	20%
Data Cleaning and Analysis	21/06	30/06	0	0%
K-Prototype Clustering	03/07	14/07	0	0%
Result Analysis	17/07	19/07	0	0%
Dissertation Writting	20/07	02/08	0	0%
Revision of work ans submission	03/08	14/08	0	0%