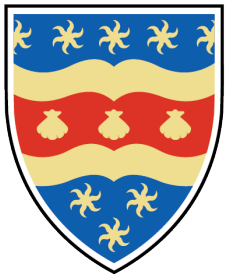

UNIVERSITY OF
PLYMOUTH

Topic Modelling on the application of British Telecommunication(BT) Data

Dhanshree Gaikwad(10800967)

September 2023

Supervised by: Dr. Malgorzata Wojtys

School of Engineering, Computing and Mathematics University of Plymouth

Abstract: In the contemporary digital era, enterprises are faced with an overwhelming volume of text data produced by their customers, users, and clients. Deriving valuable insights from this unstructured textual content has become a top priority for organizations spanning various sectors. Natural Language Processing (NLP) and text mining have risen as indispensable instruments for businesses aiming to leverage the potential of their text-based information. This dissertation delves into the domain of unsupervised text mining, investigating its importance in understanding user motivations and delivering actionable knowledge. The study employs a variety of advanced topic modelling algorithms, including Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA), to uncover latent themes and patterns within an extensive dataset sourced from the BT community forum. The dataset comprises unstructured textual discussions related to BT products and services, capturing the rich tapestry of user interactions. The primary objective of this research is to showcase how text mining models can play a pivotal role in unveiling the nuanced sentiments, concerns, and preferences of users, even in the absence of labelled data. By employing these algorithms, we aim to shed light on the latent topics and underlying structures within the vast repository of unlabelled customer service textual data. Through a comprehensive analysis, this dissertation illustrates the potential of unsupervised text mining techniques in providing businesses with a profound understanding of user intentions and actionable insights for informed decision-making. The findings of this study underscore the critical role of NLP and unsupervised text mining in today's data-driven landscape. It also serves as a testament to the transformative capabilities of text mining models and their potential to revolutionize how businesses interpret, engage with, and derive value from unstructured textual information.

Keywords: Topic identification, Topic modeling, LDA, NMF, LSA, British Telecomm(BT), Data analysis

Word Count:13141

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my sincere gratitude to Professor Malgorzata Wojtys for her unwavering support, guidance, and invaluable insights throughout the course of my research. Her expertise in the field of data science and topic modelling has been instrumental in shaping this project and elevating its quality. I am deeply thankful to the Data Analyst, Mr. Srihari Kalidas at BT for his constructive feedback and encouragement, which significantly contributed to the development and refinement of my work. I also extend my appreciation to the BT community forum users whose discussions and contributions formed the foundation of this research. Thanks to my fellow project members who were also working under the same domain(NLP) as me for sharing their experiences have enriched the dataset and made this study possible. Furthermore, I would like to acknowledge the support and encouragement of my friends and family, whose unwavering belief in my abilities has been a constant source of motivation.Finally, I express my heartfelt thanks to all those who, directly or indirectly, have played a role in shaping this thesis and in my academic journey as a whole.

Chapter 1

Introduction

In today's world, data has become indispensable for organizations. It serves as the lifeblood that drives businesses forward, as we generate massive amounts of data every second through corporate activities, sales figures, customer records, and stakeholder interactions. The scale of data we are currently processing and the untapped potential it holds for gaining deeper insights is remarkably distinct. With the exponential growth in internet usage across diverse cultures and educational backgrounds, businesses are facing mounting challenges in deciphering the sentiments and intentions expressed in consumer emails, chats, and social media comments. Moreover, the surge in e-commerce activities has led to a substantial increase in textual interactions between clients and businesses. This availability of massive unstructured data creates an opportunity for businesses to understand and analyse the data and strengthen decision-making. Data mining is the process of identifying patterns and extracting information from such big data sets using techniques that combine machine learning, statistics, and database systems. It encompasses a wide range of methods and techniques used to discover patterns, relationships, and insights from large datasets. Text mining being one of such techniques, specifically falls under NLP. Text mining methods range from basic keyword analysis to sophisticated topic modelling.

Text mining can be categorized into two primary approaches: supervised and unsupervised. In supervised text mining, algorithms are trained on labelled datasets, enabling them to make predictions or classifications based on patterns identified during training. On the other hand, unsupervised text mining relies on discovering hidden patterns and structures within unannotated data. Text mining aims to transform unstructured text into structured data that can be analysed, visualized, and used for decision-making. Topic modelling is one of the techniques of text mining or natural language processing (NLP) within data mining. The research focuses on data collected from the British Telecommunications (BT) community forum, where users post discussions related to BT products and services. BT, also known as British Telecom, is one of the United Kingdom's leading telecommunications companies. With a rich history dating back to the establishment of the Electric Telegraph Company in 1846, BT has played a pivotal role in shaping the telecommunications landscape. Today, BT provides a wide range of services,

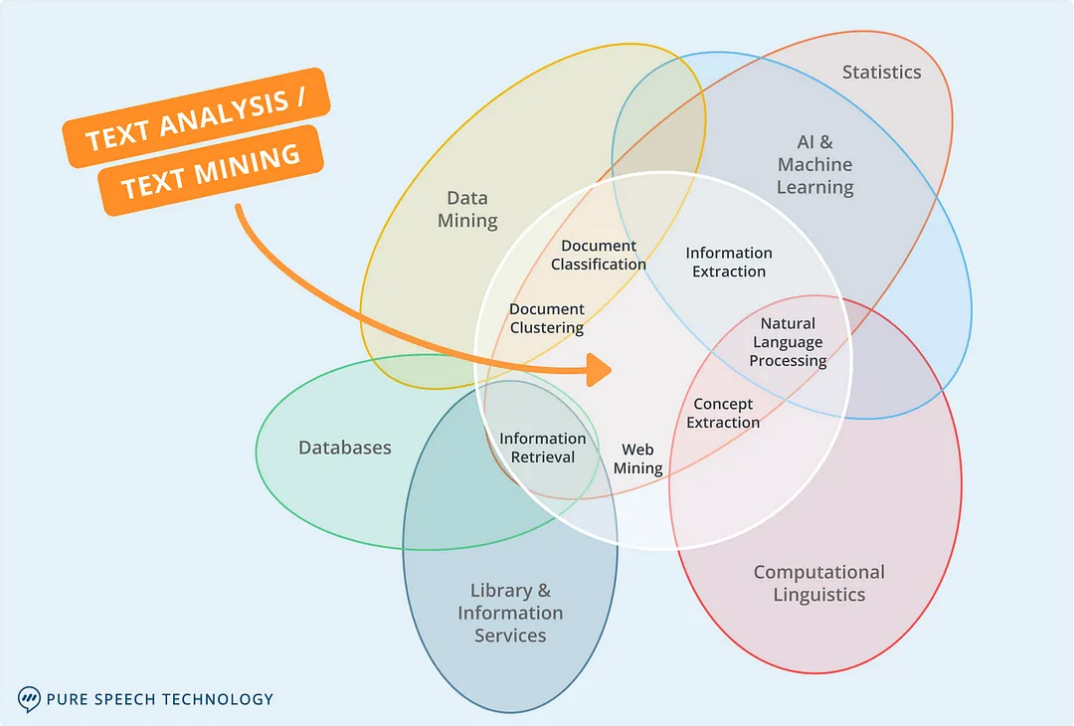

Figure 1.1: Venn diagram showing the intersection of Text Mining with six related fields (Michelle Chen, 2020)

including fixed-line and mobile telecommunications, broadband internet, and digital television. The BT community forum serves as a platform for BT customers to engage with the company, discuss their experiences, seek assistance, and share insights. Understanding user intent within this forum is crucial for BT, as it enables the company to enhance customer service, tailor its offerings, and improve the overall user experience.

In the domain of statistics, various probability models play a crucial role in text mining. These models include generative, probabilistic, and descriptive models. Generative models aim to generate new data samples that resemble the original data distribution. Probabilistic models focus on capturing probabilistic relationships between words and documents, enabling the modelling of document-topic associations. Descriptive models, on the other hand, aim to provide meaningful descriptions of the underlying data.

1.1 Goals and Objectives

The goal of this project is to leverage advanced topic modelling algorithms, including Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA), to uncover and understand user intent within the data downloaded from the BT Community Forum.

The objectives of this project are as follows

1. To develop and implement an effective data processing pipeline for the BT community dataset, with a focus on data collection, cleansing, transformation, and feature extraction,

to prepare the dataset for advanced text mining and topic modelling techniques, ultimately facilitating the exploration and extraction of meaningful insights from user discussions.

2. To explore the application of topic modelling algorithms, including Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA), for uncovering hidden themes in unstructured text data.

3. To evaluate the effectiveness of these topic modelling approaches in capturing user intent within discussions related to BT products and services in the BT community forum dataset.

4. To provide actionable insights for businesses, specifically the BT company, by interpreting the discovered topics and understanding user preferences and concerns.

1.2 Structure of the Implementation

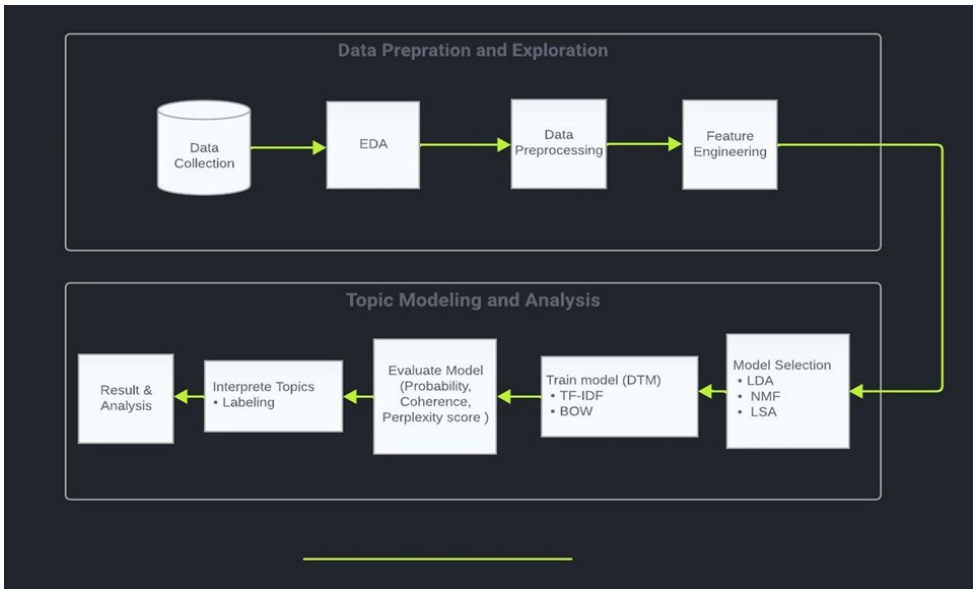

Figure 1.2: Illustration of Project Framework

Data Collection Data has been sourced from the official site of BT community forum.`https://community.bt.com/` **Exploratory Data Analysis (EDA)**Analyzing and understanding the characteristics of the dataset, including its size, distribution, and content. **Preprocessing**Preparing the text data for analysis by removing noise, stopwords, and irrelevant information. **Feature Engineering**Creating meaningful features or representations of the text data for modelling. **Model Selection**: Choosing appropriate unsupervised topic modelling algorithms, such as LDA, NMF, and LSA, based on the project's objectives. **Training the Model**Applying the selected topic modelling techniques to uncover latent themes and topics within the dataset. **Evaluating the Model**Assessing the performance of the models in capturing user intent and generating meaningful topics. **Result and Analysis**Interpreting the discovered topics to gain insights into user preferences and concerns related to BT products and services[1].

Chapter 2

Literature Review

This section of this report plays an essential role in providing a comprehensive examination of existing scholarly work, theories, and methodologies relevant to the subject matter. In this section, the focus is on systematically exploring and synthesizing pertinent literature to establish the theoretical framework and context for the study. Prior research and academic discourse are examined to identify gaps, challenges, and areas of opportunity within the field. This literature review not only offers an overview of key concepts and theories but also involves a critical analysis and comparison of various perspectives and findings from previous studies. Through this process, a robust foundation is constructed upon which the research objectives and methodologies are firmly anchored. As stated in the beginning of the report, the ever-increasing volume of data generated from various sources, including corporate transactions, social media interactions, and user-generated content, has prompted the need for effective techniques to extract valuable insights and knowledge from this wealth of unstructured textual data. This literature review explores various aspects of text mining, data cleaning, and topic modelling, providing a comprehensive background for the dissertation on Unsupervised Text Mining using topic modelling models.

2.1 Data Preprocessing

Effective text mining begins with high-quality data.[9] by Dasu Dasari and P. Suresh Varma emphasizes the significance of employing data cleaning techniques to enhance data quality using Python.

- Dasu Dasari and P. Suresh Varma[9] emphasize the importance of data cleaning techniques in enhancing data quality. Noisy data, which includes irrelevant characters, symbols, or formatting errors, can significantly impact the accuracy of topic modelling results. We followed their guidance to employ various data cleaning techniques using Python.
- A. Sharan and S. Siddiqi[18] propose a supervised approach to distinguish between keywords and stopwords using probability distribution functions. Leveraging their work,

we implemented stopwords removal to eliminate common words that do not contribute to topic identification. Additionally, insights from Sarica and Luo's [17]research on stopwords in technical language processing guided our preprocessing decisions.

- Christian McDonald [18], Cady Field [15] Christian McDonald's work on data cleaning with regular expressions and Cady Field's insights on string manipulation, regular expressions, and data cleaning were instrumental in implementing regular expressions for pattern matching and text manipulation during data preprocessing.

- Martin J. Ball [1], Henry Sweet [19], J. C. Wells [21] To ensure consistency in textual data, we also incorporated diacritics removal techniques based on the research of Martin J. Ball, Henry Sweet, and J. C. Wells. This involved eliminating diacritics such as accents and special characters from the text.

- Regular-Expressions.info[21] Cleaning and normalization of URLs within the text were accomplished by referencing Regular-Expressions.info, a valuable resource for working with regular expressions.

- For text tokenization, we relied on the Natural Language Toolkit (NLTK) library. The NLTK documentation provided essential guidance on tokenization techniques and best practices.

- Feature engineering plays a crucial role in preparing text data for topic modelling.[6] and [5] by Chen et al. and[11] Egger explore experimental approaches and deep NMF-based schemes for short text topic mining. These studies provide insights into how feature engineering choices can impact the performance of topic modelling techniques.

2.2 Topic Modelling and Its Significance

This section provides a comprehensive examination of existing scholarly work, theories, and methodologies relevant to the subject of topic modelling, with a focus on identifying popular methods and their suitability for analyzing BT community data.

- Topic modelling is a powerful technique for uncovering latent themes and patterns within textual data. While several methods exist, including Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA), the popularity of these methods varies depending on their strengths and applications.

2.3 Latent Dirichlet Allocation (LDA)

LDA is a widely adopted topic modelling technique known for its effectiveness in identifying topics within text documents. [2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan provides

a comprehensive overview of LDA, emphasizing its applicability in various domains. LDA has been used in analyzing user comments [16], aligning with the research focus on BT community data. LDA is particularly suitable for its ability to identify topics in a document collection and is widely applied for various text mining tasks.

2.4 Non-Negative Matrix Factorization (NMF)

NMF is another popular technique for topic modelling, known for its interpretability and ability to capture latent patterns. [8] by Jaegul Choo et al. presents an interactive NMF-based approach for topic modelling and document clustering. NMF is suitable for tasks where interpretability of topics is crucial, making it a valuable choice for understanding BT community discussions. DC-NMF, a divide-and-conquer variant of NMF, has been introduced for fast clustering and topic modelling [10].

2.5 Latent Semantic Analysis (LSA)

LSA remains relevant in text mining for its ability to capture semantic relationships between words and documents. [12] Hotho et al. emphasizes LSA's role in information retrieval. LSA can complement other topic modelling methods like LDA and NMF by providing insights into semantic associations within text data.

The choice of a topic modelling method for BT community data should consider the characteristics of the dataset and the research objectives. Here's an assessment of the suitability of these methods:

Please find below assessment of the suitability of these methods:

1. LDA is a versatile method suitable for a wide range of text mining tasks. It can be applied to analyze BT community data effectively, especially if the goal is to identify broad topics and trends within discussions. LDA's generative assumption aligns with the nature of user-generated content, making it a strong candidate.

2. NMF's interpretability makes it a valuable choice for understanding and explaining topics in BT community data. If the research aims to extract meaningful insights that can be easily interpreted by stakeholders, NMF is a suitable method. DC-NMF offers scalability for larger datasets, which can be advantageous for extensive BT community discussions.

3. LSA's strength lies in capturing semantic relationships, which can enhance the understanding of user-generated content.
   It complements other topic modelling methods and can be used in conjunction with LDA or NMF for a more comprehensive analysis.However, based on the project objectives the most suitable method would be Latent Dirichlet Allocation (LDA) due to following reasons.

4. LDA is well-suited for capturing user intent within discussions. It's particularly effective at identifying latent topics in a collection of documents. In the context of BT community discussions, LDA can help uncover the main themes and topics that users are discussing, which directly relates to understanding user intent.

5. LDA provides easily interpretable topics. This means that the topics discovered by LDA are represented as distributions of words, making it straightforward to understand the main themes of the discussions. This aligns with the objective of interpreting the discovered topics.

6. LDA's interpretability also extends to actionable insights. By analyzing the topics generated by LDA, you can gain insights into user preferences, concerns, and trends related to BT products and services. These insights can be directly used by BT or other businesses to make informed decisions and improvements.

7. LDA is a widely adopted and well-established method in the field of topic modelling. There is extensive support, libraries, and resources available for implementing LDA, making it a practical choice for research and business applications.
   While LSA and NMF are valuable methods, LDA's ability to effectively capture topics and provide interpretable results makes it the best choice for meeting the specified objectives in the context of understanding user intent and providing actionable insights for BT and similar businesses.Analysis of User Comments In the context of user comments, it's essential to consider studies that specifically address this area. Reference [4] by Stanik, Pietz, and Maalej discusses unsupervised topic discovery in user comments, which can be relevant to your research on BT community discussions. Additionally, reference [3] by Carreño and Winbladh analyzes user comments as an approach for software requirements evolution. These references highlight the importance of understanding and extracting insights from user-generated content, a key aspect of your research.

2.6 Evaluation Metrics

Evaluating the performance of topic modelling models is essential. References [7] and [20] discuss experimental explorations and deep NMF topic modelling, providing insights into metrics like coherence and perplexity. Reference [14] by Luo et al. introduces probabilistic non-negative matrix factorization, which can be used to assess topic modelling results. The paper titled "Applications of Topic Models" by Boyd-Graber, J., Hu, and Mimno (2017) provides an in-depth exploration of the practical applications of topic modelling techniques in various domains. This monograph addresses the challenge of understanding and making sense of large document collections, a problem that is increasingly prevalent in today's data-driven world. The paper's content is highly pertinent to the research topic modelling BT community topic modelling for several reasons.

1. The paper's opening question, "How can a single person understand what's going on in a collection of millions of documents?" resonates with the challenge of sifting through the vast amount of user-generated content in the BT community. It suggests that topic models can provide a solution by helping users grasp the general themes present in large document collections.

2. The paper discusses the effectiveness of topic models in information retrieval, which aligns with the objective of BT community research. It can aid in organizing and retrieving relevant information from user discussions.

3. Applications of Topic Models" goes beyond traditional applications and explores how topic models can facilitate qualitative analysis of text collections. This aspect is valuable for research, allowing me to gain deeper insights into user discussions and uncover nuanced themes.

4. Applications of Topic Models" goes beyond traditional applications and explores how topic models can facilitate qualitative analysis of text collections. This aspect is valuable for research, allowing me to gain deeper insights into user discussions and uncover nuanced themes.

5. The paper reviews the successful use of topic models in various domains, including scientific publications and political texts. While my research centres on BT community data, the paper's insights into applications across different domains can inspire innovative approaches or draw parallels to BT's dataset.

6. The paper is written for readers with a basic understanding of document processing and probability, making it accessible to research students. It provides a balanced introduction to topic models, aligning with the aim of my research. This is a valuable resource for the topic modelling research on BT community data. It offers insights into practical applications and qualitative analysis techniques, enhancing the effectiveness of my research by enabling a deeper understanding of the user-generated content within the BT community.

2.7 Case Study

Abdelmotaleb et al.'s GSDMM Model Evaluation Techniques with Application to British Telecom Data (Abdelmotaleb et al., 2023). Abdelmotaleb et al.'s paper focuses on evaluating topic models with the application to British telecom data using the Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) model. Here, we will explore the key aspects of this paper and identify points of comparison with our own research.

1. The referenced paper by Abdelmotaleb et al. conducts its research using short text datasets related to the British telecommunication industry, which may include diverse sources of

text such as customer feedback and reviews. In contrast, this study focuses exclusively on the BT community data, which comprises discussions and user-generated content specifically related to BT services and products. This data source is highly domain-specific and may possess unique characteristics compared to broader telecommunication industry datasets.

2. Both studies employ topic modelling techniques, but the choice of algorithms or adaptations may differ. Abdelmotaleb et al. specifically explore the Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) model for short text topic modelling. In this study, the utilization of topic modelling algorithms is tailored to the distinct characteristics of the BT community data, potentially involving variations or adaptations of topic modelling methods like LDA, NMF and LSA.

3. Abdelmotaleb et al. introduce novel evaluation methods that employ word embedding and measure within-topic variability and separation between topics. They focus on evaluating the GSDMM model and tuning its hyper-parameters. In this research, we may employ distinct evaluation methods or domain-specific metrics to assess the quality and relevance of topic models within the context of the BT community. These methods may differ from those presented in Abdelmotaleb et al.'s paper, as our study addresses unique data and research goals.

4. In summary, while both studies fall within the broader theme of text mining and topic modelling in the telecommunication industry, they differ in several key aspects, including data source, employed topic modelling techniques, evaluation methods, and specific research objectives. These distinctions underscore the uniqueness and relevance of our study to the BT community domain, as it seeks to extract valuable insights from a specialized dataset tailored to BT services and products.

2.8 Conclusion

This literature review has provided insights into various aspects of text mining, data cleaning, and topic modelling. It has highlighted the significance of these techniques in extracting valuable information from textual data and understanding user intentions. As we delve into the dissertation on "Unsupervised Text Mining" using topic modelling models, these foundational concepts will serve as a solid basis for further exploration and analysis.

Chapter 3

Methodology

The methodology section forms the bedrock of this research, presenting the structured approach employed to attain the objectives of this dissertation. Within this section, we elucidate the process of data collection, delve into exploratory data analysis (EDA), expound upon preprocessing procedures, describe the intricacies of feature engineering, explore the nuances of model selection, describe model training methodologies, discuss the metrics utilized for evaluation, and, in conclusion, present the intended analysis of results.

1. **Data Collection**
   For this study, data is sourced from the British Telecommunications (BT) community forum, a platform where users engage in discussions regarding BT products and services. The dataset encompasses a diverse array of textual data, including user comments, discussions, and queries. These unstructured textual inputs serve as the foundational material upon which our text mining and topic modelling analysis is constructed.

2. **Exploratory Data Analysis (EDA)**
   EDA represents a pivotal phase within the data preprocessing continuum. It entails acquiring an in-depth understanding of the dataset's structure, content, and attributes. Throughout the EDA process, we will conduct statistical assessments, visualize data distributions, and unearth crucial insights that will inform subsequent preprocessing stages. EDA is supposed to be conducted before preprocessing of the data which will help us explore data types of the columns present it dataset, identifying presence of null values, outlier detection and analysing the patterns of data.

3. **Preprocessing**
   Data preprocessing is a fundamental and often labour-intensive phase in any data analysis or text mining project. In the context of this research focused on BT community data, the preprocessing of the dataset emerged as a pivotal and time-consuming task. The BT community dataset, characterized by its noisy and unstructured nature, required extensive data processing to prepare it for subsequent analysis, particularly text mining and topic

modelling. This report delves into the data processing methods employed and their significance in the research process.

Data processing was identified as a central and critical component of the research due to several reasons:

- The BT community dataset exhibited a high level of noise, primarily stemming from user-generated content. Noise in the form of typographical errors, non-standard language, and irrelevant information had the potential to severely impact the accuracy and interpretability of the subsequent analysis. Thus, noise reduction was essential to ensure the reliability of the findings.
- To derive meaningful insights from the BT community data, the dataset needed to be cleaned and standardized. This involved addressing issues such as inconsistent casing, punctuation, and the presence of special characters. By improving data quality, the preprocessing phase aimed to set a solid foundation for subsequent analysis.
- Text mining and topic modelling techniques rely on well-structured and clean textual data. Tasks such as lowercasing, punctuation removal, and stop word elimination were essential to prepare the text for analysis. Additionally, lemmatization helped reduce word variations to their base forms, aiding in the identification of topics and patterns within the text.
- The BT community data contained various types of noise, including URLs, HTML tags, and numerical values, which were irrelevant to the research objectives. The preprocessing phase focused on systematically removing these elements to distil the dataset into its core textual content.

The preprocessing of the BT community data involved a series of systematic steps, including:

(a) All text data to be converted to lowercase to ensure consistency in word matching and analysis.

(b) Punctuation marks to be removed to eliminate unnecessary noise and to facilitate word tokenization.

(c) Common stopwords, which carry little semantic value, are to be removed to reduce noise, and enhance the relevance of the remaining words.

(d) Highly frequent words, which may not contribute to topic modelling insights, to be identified and removed.

(e) Perform words lemmatization to reduce inflectional forms to their base or dictionary forms, aiding in topic identification.

(f) Extraneous elements such as URLs and HTML tags to be systematically eliminated to maintain the integrity of the textual content.

(g) Numeric values, often non-relevant to text mining and topic modelling, to be excluded from the dataset.

Time Investment in Data Processing,

Data processing emerged as the most time-consuming phase of the research. Given the extensive nature of the BT community dataset, each preprocessing step required meticulous implementation and validation. The efforts invested in noise reduction, data quality enhancement, and text preparation were substantial. However, the time spent on data processing was considered a worthwhile investment, as it paved the way for more efficient and meaningful analysis in subsequent phases of the research. In conclusion, data processing played a pivotal role in this research focused on BT community data. The extensive noise and unstructured nature of the dataset made preprocessing a time-consuming yet indispensable step. The systematic application of techniques such as lowercasing, punctuation removal, stopwords elimination, and lemmatization, along with the removal of URLs, HTML tags, and numerical values, transformed the raw data into a clean and structured dataset ready for advanced text mining and topic modelling. This thorough data processing laid the foundation for the generation of meaningful insights and patterns within the BT community discussions, making it a crucial and valuable aspect of the research.

4. **Feature Engineering**

It's the process of transforming raw text data into a format suitable for modelling. In the context of text mining, this involves creating a numerical representation of the text data. Here, we can employ the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique, which assigns numerical values to words based on their importance in the corpus. Another popular technique is Bag of Words. The Bag of Words (BoW) is a text representation method where a document is represented as an unordered collection of words, disregarding grammar, word order, and context, while considering only the frequency of each word. In other words, it's like throwing all the words of a document into a bag, shaking it, and then counting how many times each word appears.

5. **Model Selection**

This research employs three distinct topic modelling techniques: Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA). Each of these algorithms serves a unique purpose in uncovering latent topics within the textual data.

(a) Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model that assumes each document is a mixture

of a set of topics. It helps identify underlying topics within the dataset and assigns topic probabilities to each document[13]

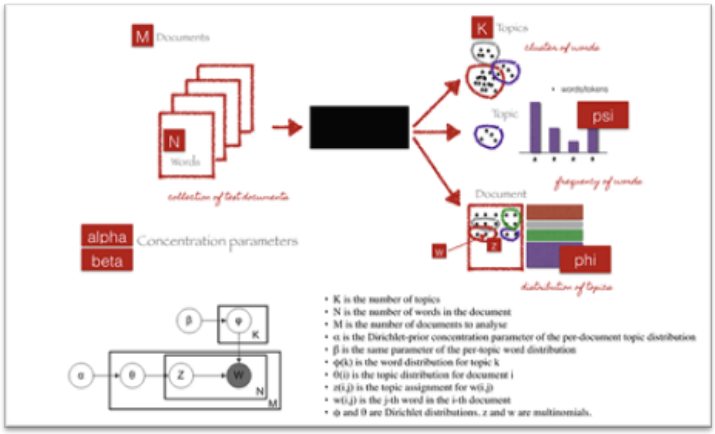

Figure 3.1: Topic Modelling by Latent Dirichlet Allocation (LDA) (Source: Shashank Kapadia, 2023)

We can depict the underlying process of Latent Dirichlet Allocation (LDA) as follows: given a set of M documents, N words in total, and an a priori specification of K topics, the model is trained to generate:

- ψ : This represents the distribution of words within each of the K topics.
- ϕ : This signifies the distribution of topics within each document i.

The Alpha parameter corresponds to the Dirichlet prior concentration parameter and reflects the density of topics within documents. Higher alpha values suggest that documents consist of a more diverse mixture of topics, leading to a more specific topic distribution within each document.

The Beta parameter, which is identical to the prior concentration parameter, characterizes the density of words within topics. Higher beta values imply that topics are comprised of a larger portion of the words, resulting in a more precise distribution of words for each topic.

(b) Non-Negative Matrix Factorization (NMF)
Unlike LDA, NMF takes a different approach as a decomposition, non-probabilistic algorithm that utilizes matrix factorization. It falls within the category of linear-algebraic algorithms[11] NMF operates on TF-IDF transformed data by breaking down a matrix into two matrices of lower rank. Specifically, TF-IDF serves as a measure for assessing the significance of words in a collection of documents. As illustrated in Figure 2.2, NMF decomposes its input, represented as a term-document matrix (A), into the product of a terms-topics matrix (W) and a topics-documents matrix (H). The values within W and H are iteratively adjusted, where W contains the foundational vectors, and H contains the corresponding weights [5]. It's crucial to

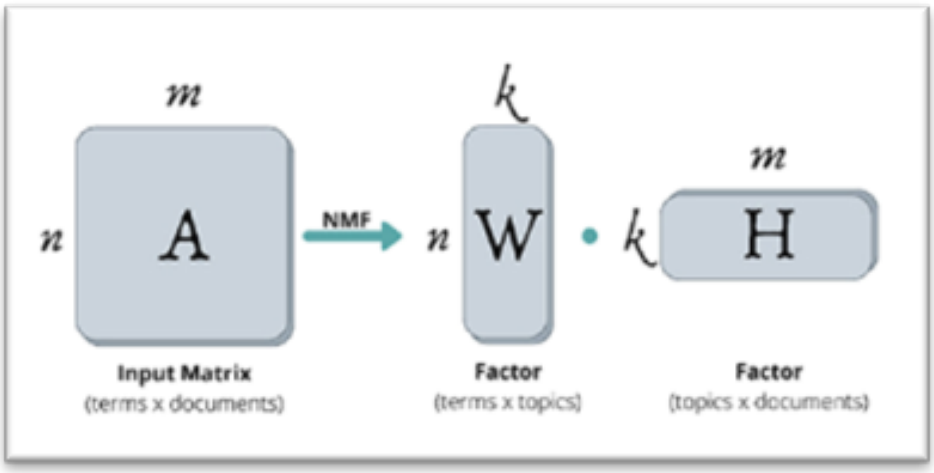

Figure 3.2: Intuition of NMF (Source : Egger, R. (2022b))

emphasize that all entries in W and H are constrained to be non-negative to facilitate the interpretation of topics; otherwise, the presence of negative values would pose interpretational challenges. Because NMF necessitates preprocessing of the data, several essential steps must be executed in advance. These include typical tasks in a classical Natural Language Processing (NLP) pipeline, such as converting text to lowercase, eliminating stopwords, performing lemmatization or stemming, and removing punctuation and numeric characters.

(c) Latent Semantic Analysis (LSA)

LSA, also known as Latent Semantic Indexing (LSI), uses singular value decomposition to discover patterns and relationships between words and documents. It reduces the dimensionality of the data while preserving semantic information.

Every natural language possesses its unique intricacies and subtleties that often prove challenging for machines to grasp and, at times, even for humans themselves. These intricacies encompass instances where different words may denote the same entity and situations where words with identical spellings carry distinct meanings.

For instance, consider the example of these two sentences:

”I thoroughly enjoyed Premchand’s latest novel.” ”They intend to embark on a novel marketing strategy.” In the first sentence, the term ’novel’ signifies a literary work, while in the second, it conveys the idea of something new or fresh. Humans effortlessly distinguish between these two interpretations by discerning the contextual nuances. Conversely, machines grapple with this task since they lack the ability to comprehend the contextual cues surrounding word usage. This is precisely where Latent Semantic Analysis (LSA) plays a pivotal role.

LSA endeavors to exploit the surrounding context of words to uncover latent concepts, referred to as topics. Simply mapping words to documents does not provide a comprehensive solution. What we truly need is the ability to extract these concealed

concepts or topics, and LSA is one such technique designed for this purpose.

Model Training

The selected algorithms are trained on the pre-processed and feature-engineered dataset. During training, the models learn the underlying topics within the textual data.

Evaluation Metric

To assess the performance of the topic modelling models, we utilize three key evaluation metrics: Probability Scores, Coherence Score and Perplexity Score: These metrics and scores are commonly used for evaluating the performance and quality of topic models, but their availability can depend on the specific algorithm and the software libraries or tools being used.

Below we present a brief explanation of each metric:

(a) Probability based metrics, such as Perplexity, are often associated with probabilistic topic models like Latent Dirichlet Allocation (LDA). These metrics measure how well the model predicts the observed data. Lower perplexity values indicate better model performance.

(b) Coherence(C_v) is a metric used to assess the interpretability of topics generated by a topic model. It measures the semantic similarity between the top words within a topic. Higher coherence scores indicate more interpretable and coherent topics.

We start with a corpus C with D documents and define the following quantities:

- δ_d document d represented by a bag of words.
- $|\delta_d|$ the number of words in document d,
- ω_d^C, i corpus word at index i, in document d,
- d document index in a corpus.d$\in \{1, 2, \ldots, D\}$,
- i word index in document d.i$\in 1, 2, \ldots, |\delta_d|$,

Also, we have trained a topic model with K topics and N most probable words per topic.The following quantities are:

- k topic index. k $\in 1, 2, \ldots, K$,
- n word index in a topic. n $\in 1, 2, \ldots, N$,
- W_k the set of N most likely words in topic k,
- ω_n^T, k topic word at index n in topic k,
- $\vec{\omega}_n, k$ vector to represent topic word at index n in topic k. Where$|\vec{\omega_n, k}| = N$

The C_v score is heavily based on the NPMI(Normalised Pointwise mutual Information) score, an advanced way to calculate the probability of two words co-occurring in a corpus.

$$NPMI(\omega', \omega^*) = \frac{log\frac{P(\omega', \omega^*)+\epsilon}{P(\omega')P(\omega^*)}}{-log(P(\omega', \omega^*) + \epsilon)} \tag{3.1}$$

The probability in question relies on a sliding window denoted by "s". Given "j" as the index representing the position of the sliding window within a document, the probabilities in the NPMI formula are determined through the following computation.

$$P(\omega_n, \omega_m) = \frac{\sum_{d=1}^{D} \sum^{|} \delta_d| - s_{j=1} b_d, j(\omega_n, \omega_m)}{\sum_{d=1}^{D} |\delta_d| - s} \tag{3.2}$$

(c) Perplexity

Perplexity is a measure of how well a probabilistic model predicts a sample. It is used to assess the quality of a probabilistic model, such as a topic model. It quantifies how well the model predicts a held-out or test dataset. Lower perplexity values indicate better model performance. The perplexity score for a topic model is typically calculated as follows:

Notable topic modelling algorithms like LDA and its variations, which are probabilistic models, often have perplexity scores associated with them. Coherence scores are widely used for evaluating the quality of topics across various topic modelling techniques. However, not all topic modelling algorithms are probabilistic or designed with these specific evaluation metrics in mind. For example, Non-negative Matrix Factorization (NMF), a non-probabilistic method, does not have associated perplexity scores. Evaluation for NMF may involve other techniques, such as manual inspection of topics or assessing topic quality based on the application's objectives.

6. **Result Analysis**

This phase represents the culmination of research efforts, where analysis of the outcomes derived from the application of topic modelling techniques and evaluation metrics will take place. This section plays a pivotal role in unravelling the latent topics and user intentions embedded within the BT community forum dataset.

Chapter 4

Exploratory Data Analysis (EDA)

This section of the thesis delves into the principles and practices of EDA, highlighting its significance in extracting meaningful insights from the dataset sourced from the BT community forum website. Through the systematic application of visualization, statistical summaries, and data manipulation techniques, EDA not only prepares the data for further analysis but also aids in formulating hypotheses and refining the research focus. Exploratory Data Analysis (EDA) is an essential component of data analysis where data is carefully examined and visualized to gain preliminary insights, detect patterns, and determine possible correlations. It plays a vital role in understanding the fundamental structure of the data, revealing significant characteristics, and facilitating informed decision-making for subsequent analysis. We have a data consist of 1,516 records related to the BT community. The dataset contains various columns such as 'Title', 'Link', 'Description', 'Day', 'Month', 'Year', 'Hour', 'Creator', and 'Topic'. The purpose of this exploratory data analysis (EDA) report is to gain insights, discover patterns, and understand the characteristics of the data.

4.1 Data Overview

In the "Data Overview" section, we provide a comprehensive description of the dataset used in this research, which has been sourced from the BT community forum website. This section includes valuable information about the dataset's metadata, such as its source, collection methods, and structure. Additionally, it outlines the specific columns and features present in the dataset, shedding light on the key elements that will be analysed in the subsequent sections of this thesis. A thorough understanding of the dataset's composition is essential to contextualize the text mining and topic modelling processes applied in this study. The dataset comprises diverse records from the BT community, covering topics such as announcements, guides, and community discussions. Each record contains information about the title, link, description, date (day, month, year), time (hour), creator/user, and topic. The dataset encompasses a range of dates from April 2018 to March 2022.

Description of Columns Present in Dataset

Title	The title of the topic discussed in the community.
Link	The URL link to the specific topic or discussion.
Description	A description or content related to the topic.
Day	The day when the topic was created or posted.
Month	The month when the topic was created or posted.
Year	The year when the topic was created or posted.
Hour	The hour of the day when the topic was created or posted.
Creator	The username or name of the creator who posted the topic.
Topic	The category or topic of the discussion.

Table 4.1: Description of columns present in dataset

Based on the following sample records, it appears that the dataset contains topics related to BT community announcements, guides, and discussions. The ”Title” column provides a summary of the topic, while the ”Link” column contains the URL to access the complete discussion. The ”Description” column provides additional details or content related to the topic. The date and time of each topic are split into separate columns, including ”Day,” ”Month,” ”Year,” and ”Hour.” These columns allow for easy filtering and analysis based on specific time periods. The ”Creator” column indicates the username or name of the individual who posted the topic. This information can be helpful in identifying active community members or tracking discussions from specific users. Finally, the ”Topic” column represents the category or topic under which the discussion falls. It helps in organizing and categorizing the various discussions within the BT community. The figure displayed below, Figure 1, illustrates a selection of records extracted from a dataset represented in CSV format.

	Unnamed: 0	Title	Link	Description	Day	Month	Year	Hour	Creator	Topic
0	0	Board topics	https://community.bt.com/t5/Announcements-Guides-Community/bd-p/Announcements	Board topics	28	Mar	2022	17	Announcements	Announcements
1	1	Profile settings	https://community.bt.com/t5/Announcements-Guides-Community/Profile-settings/m-p/2206884#M3713	P&g I am having an issue. My previous BT Community id is ******@gmail.com (which is active). This is before I left	17	Jan	2022	9	Drskhan	Announcements
2	2	Notice from BT?	https://community.bt.com/t5/Announcements-Guides-Community/Notice-from-BT/m-p/2204209#M3694	I received a notification today allegedly from BT asking me to confirm my acceptance of new terms and conditions by tomorrow (5 1 22) or I will lose access to my information - Question - is this genuine?	4	Jan	2022	11	nickm2	Announcements
3	3	Thank you to all our Community members	https://community.bt.com/t5/Announcements-Guides-Community/Thank-you-to-all-our-Community-members/m-p/2202055#M3686	Hi Commun I hope everyone is keeping safe and well. I can□□t quite believe we are just a few days off Christmas and not long now until we see the end of 2021, what a year it has b	21	Dec	2021	12	SeanD	Announcements
4	4	Email requesting setting up new direct debit - genuine?	https://community.bt.com/t5/Announcements-Guides-Community/Email-requesting-setting-up-new-direct-debit-genuine/m-p/2200426#M3667	I've had an email saying my direct debit 'is no longer active' and nees to be set up again. It contains the last 3 numbers of my account, but how can I check it's genuine?	13	Dec	2021	10	judex	Announcements

Figure 4.1: Sample records extracted from a dataset represented in CSV format

For an instance, the dataset contains records related to profile settings, with additional details such as the link to the community page discussing profile settings, a description mentioning an issue with a previous BT Community ID, the specific day, month, year, and hour of the record

creation, the creator identified as ”Drskhan,” and the topic labelled as ”Announcements.”

Data Type of Each Column Present in Dataset

The ”Title,” ”Link,” ”Description,” ”Creator,” and ”Topic” columns contain textual data or strings. The ”Day,” ”Month,” ”Year,” and ”Hour” columns contain numeric data, with ”Day,” ”Year,” and ”Hour” being integers, while ”Month” is represented as a text or string value.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1515 entries, 0 to 1514
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    1515 non-null   int64
 1   Title         1515 non-null   object
 2   Link          1515 non-null   object
 3   Description   1508 non-null   object
 4   Day           1515 non-null   int64
 5   Month         1515 non-null   object
 6   Year          1515 non-null   int64
 7   Hour          1515 non-null   int64
 8   Creator       1515 non-null   object
 9   Topic         1515 non-null   object
dtypes: int64(4), object(6)
memory usage: 118.5+ KB
```

Figure 4.2: Data type of each column present in dataset.

Overview of the central Tendency(Mean.Median),Spread(Standard Deviation),and Range(Max,Min)

	Unnamed: 0	Day	Year	Hour
count	1515.0	1515.0	1515.0	1515.0
mean	50.0	16.0	2022.0	14.0
std	29.0	8.0	1.0	5.0
min	0.0	1.0	2018.0	0.0
25%	25.0	9.0	2021.0	11.0
50%	50.0	16.0	2022.0	14.0
75%	75.0	23.0	2022.0	18.0
max	100.0	31.0	2022.0	23.0

Figure 4.3: General statistical characteristics of the dataset.

Identify Presence of null values in the dataset

The generated graph provides a visual representation of the presence of null values in different columns of the dataset. The x-axis is labeled as ”X - Column names” and represents the various

column names within the dataset. The y-axis is labeled as ”Y - Number of null values” and represents the count of null values in each column. Upon analyzing the graph, it has been determined that only one column, specifically the ”Description” column, contains null values. The graph shows that there are a total of seven occurrences of null values in the ”Description” column. This insight can be valuable for understanding the completeness of the dataset and identifying potential data gaps or missing information. Please refer to the figure number 3 below, which illustrates this finding.

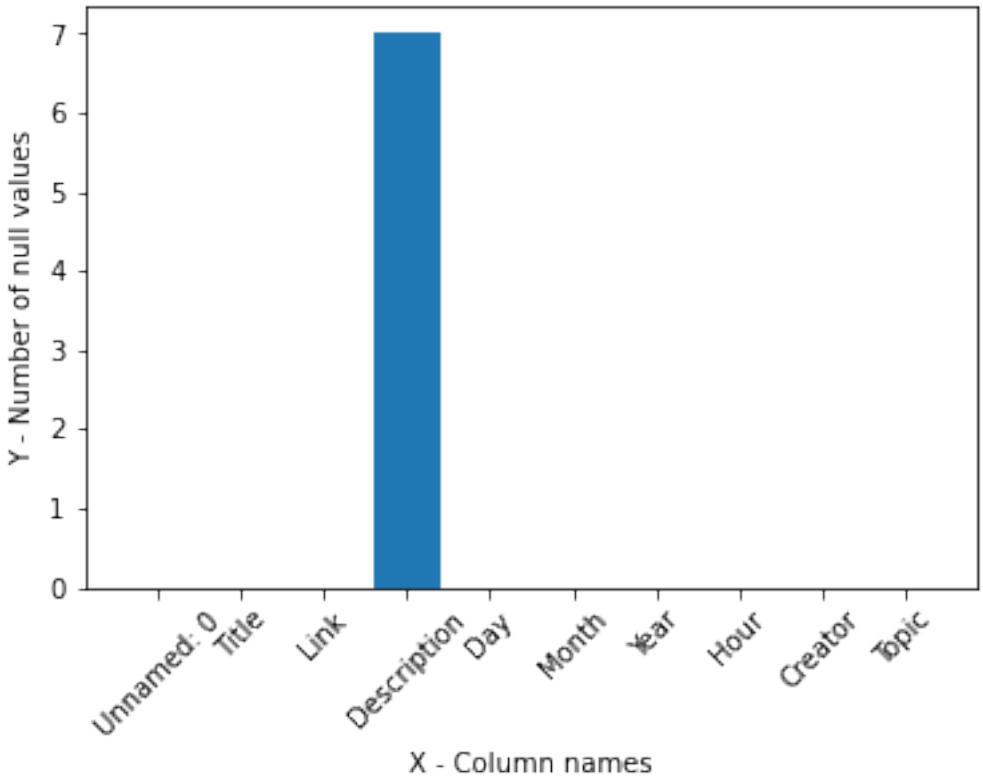

Figure 4.4: Identify presence of Null Values in the Dataset

Find FEATURES WITH CATEGORICAL DATA IN DATASET

The created graph illustrates the presence of categorical data within the dataset’s features. By examining the graph, we can observe that only the ”Topic” column or feature contains categorical data, while the remaining features appear to have non-categorical data.The x-axis of the graph represents the count of categories present within the ”Topic” field. Each bar on the x-axis corresponds to a specific category, indicating the frequency or occurrence of that category within the dataset. Notably, the count of records for each category is consistent, with all categories having a count of 101.This information suggests that the dataset’s ”Topic” feature comprises a set of distinct categories, where each category has an equal representation of 101 records.

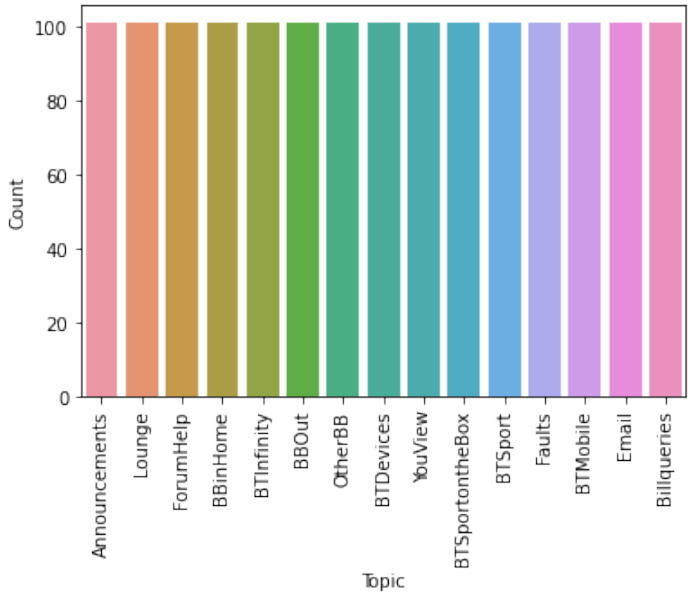

Figure 4.5: Features with categorical data in Dataset

NUMBER OF QUERIES RECEIVED PER YEAR, ARRANGED FROM HIGHEST TO LOWEST

The provided pie chart illustrates the number of queries received per year, with the years arranged from highest to lowest. Notably, the year 2022 has the highest count of queries, indicating a significant influx of queries during that year. On the other hand, the year 2018 accounts for only a minimal percentage of queries, representing a marginal portion of the total. In contrast, the year 2021 contributes to a substantial 22.2This pie chart effectively visualizes the distribution of queries across different years, emphasizing the prominence of the year 2022 and the relatively lower representation of 2018 while highlighting the significant portion attributed to 2021.

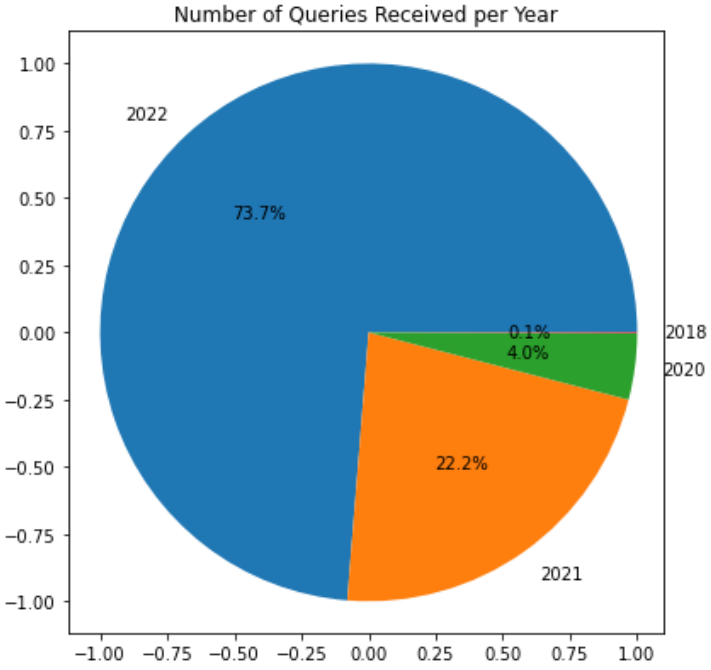

Figure 4.6: Percentage of queries received per year

TOP 5 USERS WITH HIGHEST NUMBER OF POSTS IN BT COMMUNITY FORUM

The graph represents the top five users with the highest number of posts in the BT community forum. The X-axis represents the frequency of posts, indicating the number of times each user has posted. The Y-axis displays the names of the users. According to the graph, the user "SeanD" has the highest post count, with a total of 49 posts. The remaining four users have posted below 10 times each. The graph highlights the significant difference in the number of posts between the top user and the others, indicating that SeanD is the most active contributor in the BT community forum.

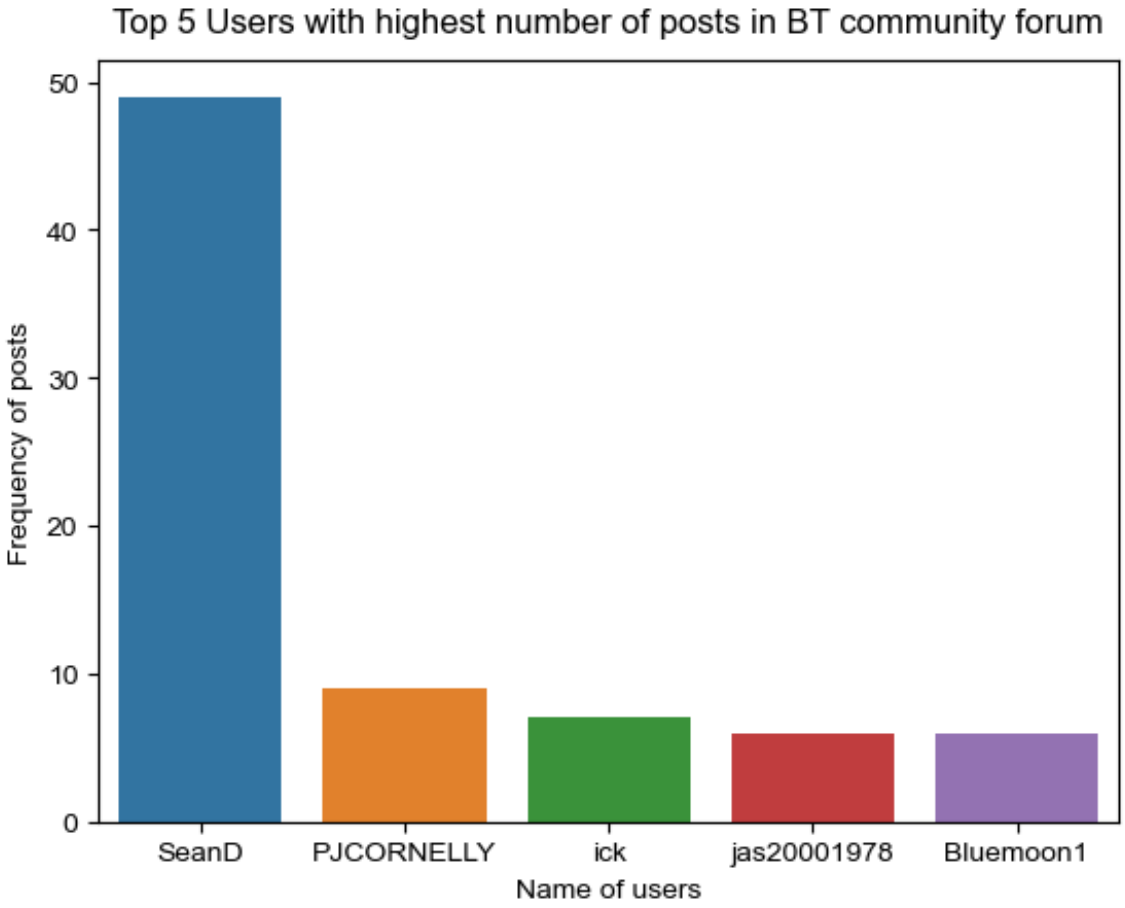

Figure 4.7: Top 5 Users with highest number of posts in BT community forum

HANDLING MISSING VALUES

According to the analysis depicted in Figure 3, it was observed that the 'Description' column contained missing values that required initial handling prior to the removal of stopwords. Missing values were eliminated from the dataset utilizing the dropna() function provided by the pandas library.

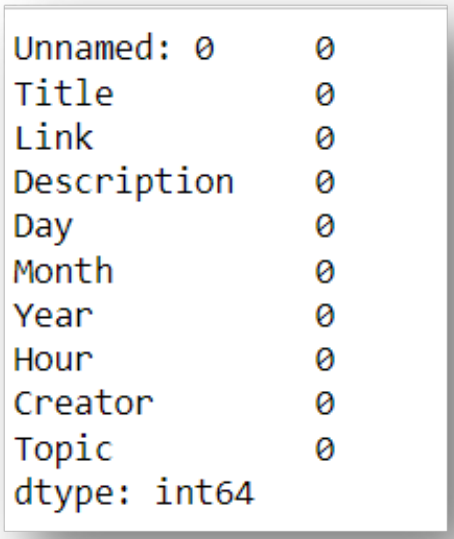

Figure 4.8: Records after removing missing values.

4.2 OUTLIER DETECTION

According to the paper by the Institute of Information Technology at Lodz University of Technology in Poland, outliers are data points, patterns, or data streams that differ significantly from other datasets or streams, or contain exceptionally rare and valuable information [x, y, ...]. Outliers are typically identified in datasets where values have a natural ordering and can be compared quantitatively. The detection of outliers is recognized as a fundamental problem in data analysis, aiming to identify objects that can be labeled as anomalous (Institute of Information Technology, Lodz University of Technology, ul. Wolczanska 215, 90-924 Lodz, Poland). Considering the nature of dataset being used, the traditional approach of outlier detection does not directly apply. The dataset includes columns representing Month, Date, and Year as integer values on which detecting outliers may not be as relevant or meaningful as it is for continuous numeric variables.

4.3 Given dataset has the presence of following.

The dataset encompasses a wide array of elements, including dates, timings, monetary amounts in pounds, special characters, stopwords, and numerical values. In essence, it encompasses a comprehensive collection of components that users typically employ when crafting comments or engaging in discussions. Before embarking on data preprocessing, let's take a brief glimpse at the various entities present within the dataset. These entities are enumerated as follows:

4.3.1 User Queries

A significant portion of the data consists of user queries and questions related to BT services, accounts, billing, and technical issues. These user-generated texts represent valuable insights into customer concerns and experiences. Please find sample posts below.

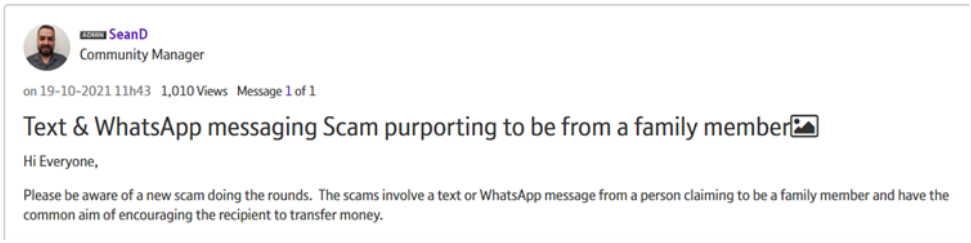

Figure 4.9: BT community Forum

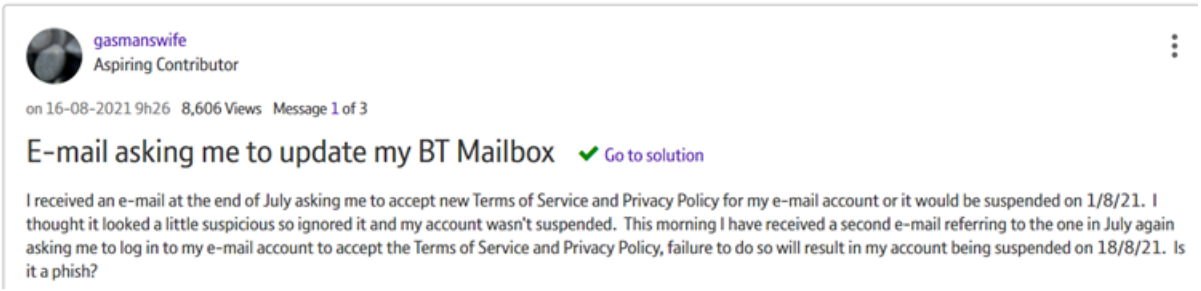

Figure 4.10: BT community Forum

4.3.2 Non ASCI /non-standard characters:

Non-ASCII characters are characters that are not a part of the ASCII character set, which only includes basic Latin letters, digits, and a few special symbols. Given dataset contains instances of those non-standard characters and accented marks. such as ”â€œBearsâ€,” likely result from encoding or decoding issues during data collection or storage. It’s important to address these encoding problems to ensure accurate text representation. In the context of text and data, an accented or diacritic mark is a symbol added to a letter or character to indicate a specific phonetic or linguistic feature. Diacritics can alter the pronunciation or meaning of a character in a particular language. They are commonly used in languages that have characters with different phonetic values or to indicate stress, tone, or other linguistic distinctions. (Perea, et al. 2016) For example, in the word ”résumé,” the diacritic mark is above the letter ”e” which changes its pronunciation, making it a two-syllable word rather than a one-syllable word ”resume. Please find following user’s comment from BT sample data vs how it is interpreted when downloaded in CSV format.

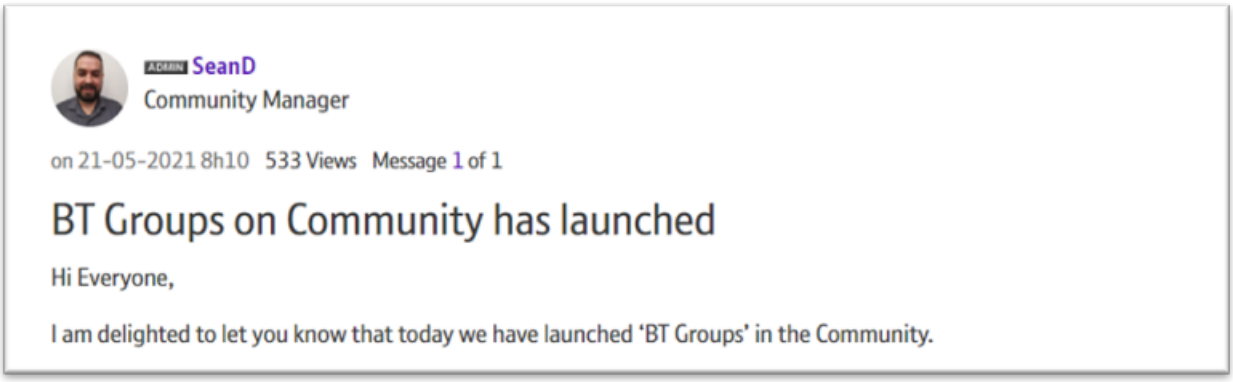

Figure 4.11: Samle Bt Posts

We can see that the punctuation (‘) mark has been decoded as “Â‰ÃÃ”

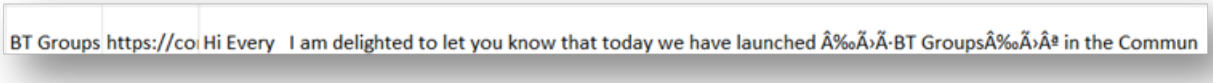

Figure 4.12: Text data after downloading into CSV format

4.3.3 HTML entities

The data has been gathered from the BT community forum website, where users engage in discussions and share their experiences. This data was collected and made available in a structured CSV (Comma-Separated Values) format, which is a common and easily accessible way to store tabular data. Each row in the CSV corresponds to a user’s post or comment on the forum, and each column represents a specific attribute or piece of information related to that post. The primary focus of this data is textual content, comprising user-generated comments, discussions, and interactions. Images, videos, or other multimedia content shared by users in the forum discussions are not directly included in the CSV data due to the limitations of the CSV format. However, an issue arises when it comes to handling images or any multimedia data shared by users. When the data was downloaded from the forum and saved into the CSV format, any image content provided by users was transformed into script-like code. This means that the CSV file interpreted images as script snippets, incorporating HTML-like tags and attributes. This transformation can be attributed to the CSV’s inability to directly store complex multimedia data like images

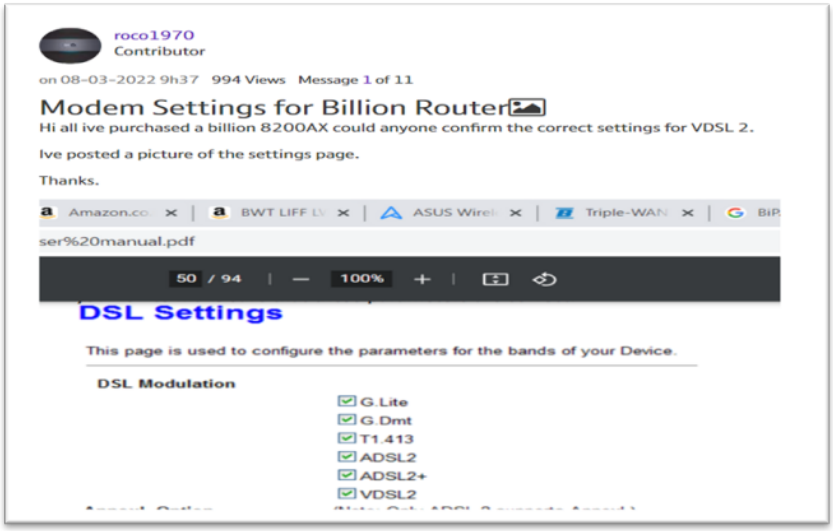

Figure 4.13: For instance, an image shared by a user might have originally appeared as follows in the forum discussion:

Source:https://community.bt.com/t5/BT-Fibre-broadband/Modem-Settings-for-Billion-Router/m-p/2219839

However, after the data was downloaded and saved in CSV format, it appears as:

Modem Settings for Billion Router	https://community.bt.com/t5/BT-Fibre-broadband/Modem-Settings-for-Billion-Router/m-p/2219839#M335211	Hi all ive purchased a billion 8200AX could anyone confirm the correct settings for VDSL 2. Ive posted a picture of the settings page. Thanks. span class="lia-inline-image-display-wrapper lia-image-align-inline" image-alt="Screenshot (21).png" style="width: 593px " img src="https:

Figure 4.14: Sample of HTML tags and attributes

4.3.4 Entries with URLs

The dataset includes a column named ”Link,” which comprises hyperlinks directing to specific user posts on the BT community forum. Furthermore, the ”Description” column contains a mixture of links interwoven with user comments. It’s worth noting that not all of these links possess validity; some are actually fragments of HTML scripts. To illustrate, consider the following example of a hyperlink present in the dataset:

¡a href=”https://www.ispreview.co.uk/index.php/2021/10/170-new-additions-to-openreachs-uk-fttp-rollout-programme.html” target=”_self”¿https://www.ispreview.co.uk/index.php/2021/10/170-new-additions-to-openreachs-uk-fttp-rollout-programme.html¡/a¿

In this example, the hyperlink points to an external webpage related to Openreach’s UK FTTP rollout program. However, please note that some of the hyperlinks in the dataset might also be fragments of HTML scripts, as mentioned earlier.

Chapter 5

Data Preprocessing

5.1 DATA CLEANING AND PREPROCESSING

In the world of information, data plays a key role in all fields of Science, Engineering and Technology. Numerous industries and services regularly handle vast amounts of data, which influences critical decisions shaping the future of organizations. However, it is common for the acquired data to contain inaccuracies (Dasu Varma, 2022). No dataset is entirely error-free, which necessitates the development of various tools to address issues like missing data, miscoding, null sets, spelling errors, and duplicate entries. Consequently, Data Cleaning emerges as a pivotal process to ensure the effectiveness and efficiency of subsequent data analysis. Cleaning the data is crucial as inaccurate data can yield unpredictable and unreliable outcomes, making it essential to prepare and refine the data prior to analysis

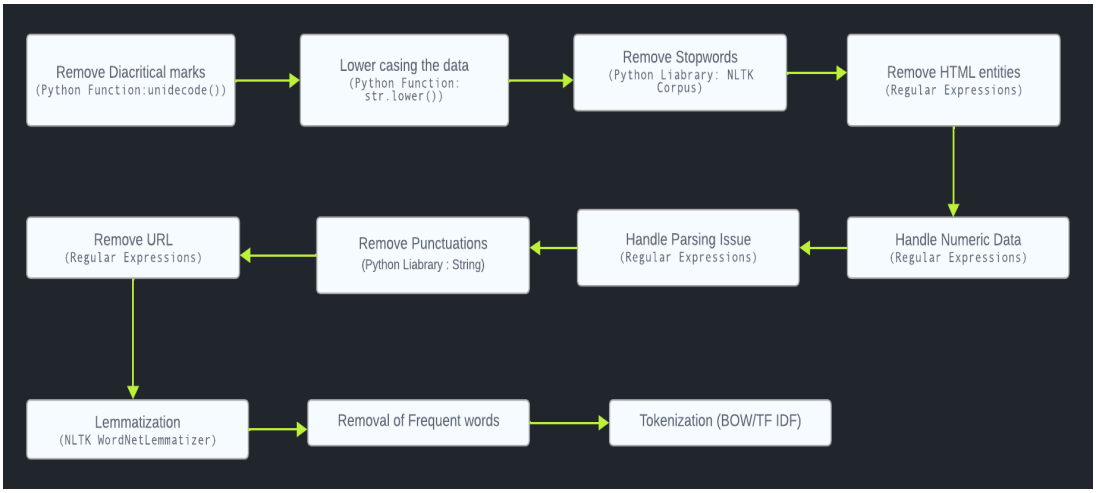

Figure 5.1: Pipeline used for the pre-processing of data

At the beginning of this phase, we will define the data cleaning tasks.

5.1.1 Remove Diacritical marks

Data cleaning process will commence by initially eliminating diacritical marks from the dataset. It's important to note that lowercasing the data prior to removing diacritical marks can influence their presence. Diacritics often carry distinct meanings in various languages and scripts, and converting characters to lowercase might modify their inherent significance or alter how diacritics function. For instance, in specific languages, the lowercase and uppercase versions of a character featuring a diacritic could convey disparate sounds or interpretations. Additionally, given that the dataset includes instances where punctuation marks coexist with diacritics (such as accents applied to punctuation symbols), the act of removing punctuations might inadvertently perturb the structural coherence of the text. To remove diacritical marks from the dataset I am using unicodedata library by python which is specifically designed to work with Unicode characters and provides accurate normalization and handling of diacritical marks. It helps change special letters and symbols from other languages into regular English letters. This is handy when working with letters that have accents or marks on them. It's like changing the way something sounds or is pronounced to make it easier to work with. Please locate the row provided from the dataset after applying the unidecode transformation to the data, both before and after removing diacritical marks.

Sample Raw Data	After removing diacritical marks
Hi Every I am delighted to let you know that today we have launched □Û÷BT Groups□Ûª in the Commun	Hi Every I am delighted to let you know that today we have launched U/BT GroupsUa in the Commun

Figure 5.2: Sample data before and after removing Diacritical marks

5.1.2 Lower casing the data

Lowercasing all text ensures uniformity and consistency throughout the dataset. Since text data is inherently unstructured, the same word can appear in various forms (e.g., "Topic," "topic," and "TOPIC"). By converting all text to lowercase, we standardize the representation of words, thereby treating them as equivalent regardless of their capitalization. This consistency is essential to prevent the same word from being counted multiple times and to improve the accuracy of frequency-based analyses.

5.1.3 Remove stopwords

Moving forward, our next action involves the elimination of stopwords. Stopwords are commonly occurring words that are often excluded from text data during tasks related to natural language processing. This exclusion stems from the fact that stopwords usually carry limited meaningful information. These words appear frequently in the text and don't significantly enhance our understanding or analysis of the content (Sharan and Siddiqi 2014). Using a standard stopword

list, often provided by widely used libraries like the Natural Language Tool Kit (NLTK), has become a norm in NLP research and industry for data preprocessing (Sarica S Luo J, 2020). I'll be utilizing the NLTK corpus to eliminate stopwords. NLTK, or the Natural Language Toolkit, is a widely used Python library that serves various purposes in natural language processing (NLP) tasks. Its effectiveness stems from its wealth of resources, encompassing a wide range of corpora, tools, and datasets that significantly enhance NLP projects. The NLTK corpus is particularly noteworthy for its effectiveness in stopword removal, and the following highlights explain why it's the one I chose: Standardized Stopword Lists: NLTK comes with ready-made lists of common words for different languages, like English. These lists are carefully put together based on language studies, so they work well for various tasks involving language processing. Effortless Integration: The NLTK corpus seamlessly blends into the NLTK library, simplifying the process of accessing and employing stopwords for efficient removal. Community Support: Since NLTK is used by many people and has a lot of people who contribute to it, there are plenty of helpful things available. I found lots of resources, guides, and conversations that could help make the most of NLTK's corpus in my projects. Please find below list of words from corpus.

```
from nltk.corpus import stopwords
", ".join(stopwords.words('english'))
```

```
"i, me, my, myself, we, our, ours, ourselves, you, you're, you've, you'll, you'd, your, yours, yourself, yourselves, he, him, h
is, himself, she, she's, her, hers, herself, it, it's, its, itself, they, them, their, theirs, themselves, what, which, who, wh
om, this, that, that'll, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing,
a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during,
before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when,
where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very,
s, t, can, will, just, don, don't, should, should've, now, d, ll, m, o, re, ve, y, ain, aren, aren't, couldn, couldn't, didn, d
idn't, doesn, doesn't, hadn, hadn't, hasn, hasn't, haven, haven't, isn, isn't, ma, mightn, mightn't, mustn, mustn't, needn, nee
dn't, shan, shan't, shouldn, shouldn't, wasn, wasn't, weren, weren't, won, won't, wouldn, wouldn't"
```

Figure 5.3: List of stopwords from NLTK Corpus

Sample Raw Data	Data after removing stopwards
P&g I am having an issue. My previous BT Community id is ******@gmail.com (which is active). This is before I left	pg issue previous bt community id which active left
I received a notification today allegedly from BT asking me to confirm my acceptance of new terms and conditions by tomorrow (5 1 22) or I will lose access to my information - Question - is this genuine?	received notification today allegedly bt asking confirm acceptance new terms conditions tomorrow lose access information question genuine
Hi Commun I hope everyone is keeping safe and well. I canÃƒÂƒÃ‚Â‚ÃƒÂ‚Ã‚Â‰ÃƒÂƒÃ‚ÂƒÃƒÂ‚Ã‚Â›ÃƒÂƒÃ‚Â‚ÃƒÂ‚Ã‚Âªt quite believe we are just a few days off Christmas and not long now until we see the end of 2021, what a year it has b	hi commun hope everyone keeping safe well canuat quite believe days christmas long see end year b
I've had an email saying my direct debit 'is no longer active' and nees to be set up again. It contains the last 3 numbers of my account, but how can I check it's genuine?	ive email saying direct debit is longer active nees set again contains last numbers account check genuine
Hello. If this is not the correct place for the enquiry please feel free to move it. I received a legit looking email stating my direct debit is no longer active. It contains actual BT links but BT confirmed on call it was not sent by them. Also checked BT account and bank account online and DD still active. Issue is that mail correctly quotes last 4 digits of my BT account, which is concerning. Assuming this is all any scammer has but still an issue I have forwarded to phishing@bt com. Has anyone else had this?	hello correct place enquiry please feel free move it received legit looking email stating direct debit longer active contains actual bt links bt confirmed call sent them also checked bt account bank account online dd still active issue mail correctly quotes last digits bt account concerning assuming scammer still issue forwarded phishingbt com anyone else this

Figure 5.4: Data sample before and after removing stopwords

Here is the data provided both before and after removing stopwords from the text

5.2 Remove HTML entities

The dataset includes HTML scripting code, as mentioned in section 3, point 3.3. The dataset obtained from the BT community forum contained a wealth of valuable information, but it was essential to preprocess the data before performing text mining and topic modelling tasks. A significant challenge in this process was the presence of HTML content within the forum posts. Unlike structured and well-defined HTML documents, the forum data consisted of user-generated content with varying degrees of HTML tags, formatting, and inconsistencies. Traditional methods, such as employing Beautiful Soup or HTML tag removal with TextHero, proved to be less effective due to the heterogeneous nature of the HTML content. To address this issue, a custom data cleaning approach was implemented, leveraging regular expressions. Regular expressions provided the flexibility needed to identify and remove HTML tags and associated content, regardless of their specific format or inconsistencies. This approach allowed for a more thorough cleaning of the data, ensuring that the subsequent text mining and topic modeling processes were not influenced by HTML artifacts. The regular expressions used in this cleaning process were designed to match patterns commonly found in HTML content, such as opening and closing tags, attributes, and text enclosed within tags. By systematically searching for these patterns and replacing them with empty strings, we were able to effectively strip away the HTML components, leaving behind the raw textual content of the forum posts. While regular expressions provided the necessary flexibility and control over the cleaning process, it's important to note that this method required careful crafting of regex patterns to handle the specific idiosyncrasies of the dataset. Please find below sample data before and after removing HTML tags.

Sample raw data	After removing html tags
span class="lia-inline-image-display-wrapper lia-image-align-center" image-alt="SFT Staying safe online_Hero side-by-side Full 1920x800 px.jpg" style="width: 999px " img src="https: community.bt.com t5 image serverpage image-id 74858i7B10DA2A47760FA1 image-size large?v=v2&amp amp px=999" role="button" title="SFT Staying safe online_Hero side-by-side Full 1920x800 px.jpg" alt="SFT Staying safe online_Hero side-by-side Full 1920x800 px.jpg" span &amp n	n
FONT size="4" STRONG SPELLCHECKER. Having a day off ? STRONG FONT span class="lia-inline-image-display-wrapper lia-image-align-center" image-alt="OFHC-LJ-500ani.gif" style="width: 500px " img src="https: community.bt.com t5 image serverpage image-id 76216i3FEED0B88A799A2B image-size large?v=v2&amp amp px=999" role="button" title="OFHC-LJ-500ani.gif" alt="OFHC-LJ-500ani.gif" span FONT size="4" STRONG I wonder if the Prime Minister is invited to the party ? STRONG FONT	strong spellchecker day strong strong wonder prime minister invite party strong
data-contrast="auto" Hi Everyone, data-ccp-props="{&amp quot 201341983&amp quot :0,&amp quot 335559739&amp quot :160,&amp quot 335559740&amp quot :259}" data-contrast="auto" We data-contrast="auto" data-contrast="auto" would like your help! data-contrast="auto" As the community manager, I data-contrast="auto" am part data-contrast="auto" data-contrast="auto" of the Get help team in our Digital department. data-contrast="auto" We data-contrast="auto" data-contrast="auto" w data-contrast="auto" ant data-contrast="auto" to share data-contrast="auto" ou data-contrast="auto" r plans for the upcoming data-contrast="auto" months data-contrast="auto" - data-contrast="auto" and data-contrast="auto" get data-contrast="auto" your data-contrast="auto" questions and data-contrast="auto" feedback data-contrast="auto" to help us data-contrast="auto" keep the focus on the right things. data-ccp-props="{&amp quot 201341983&amp quot :0,&amp quot 335551550&amp quot :1,&amp quot 335551620&amp quot :1,&amp quot 335559739&amp quot :160,&amp quot 335559740&amp quot :259}"	everyone help community manager part help team digital department w ant share ou r plan upcoming month question feedback help u keep focus right thing

Figure 5.5: data before and removing HTML entities

5.2.1 Remove URL

In addition to handling HTML content, the dataset collected from the BT community forum contained a plethora of URLs. These URLs were embedded within the forum posts and often served as references or sources cited by forum members. However, the URLs were far from uniform in format, and some were not even valid, leading to issues that needed to be addressed before further analysis. To effectively clean and standardize the URLs within the dataset, a custom URL cleaning process was devised. Regular expressions were once again employed to tackle the variability in URL formats and to identify and remove invalid or non-functional URLs. Customized Regular Expressions: Customized regular expressions were developed to address the diverse range of URL formats present in the data. These regular expressions were designed to recognize and capture URLs regardless of their structure, including those with varying use of protocols (e.g., ”http://” or ”https://”), domain names, subdomains, and query parameters. Handling Invalid URLs: Some URLs within the dataset were not functional and did not lead to valid webpages when accessed. To ensure the integrity of the data and prevent issues during further analysis, regular expressions were employed to identify and remove these invalid URLs. The criteria for identifying invalid URLs were based on common patterns of non-functional or erroneous URLs, such as those containing typographical errors or missing essential components.

It's important to note that while regular expressions proved to be a powerful tool for URL cleaning, the specific regular expressions used were tailored to the characteristics and idiosyncrasies of the dataset. Adjustments and fine-tuning were performed iteratively to handle any unique challenges that arose during the cleaning process.

Sample raw data	After removing URLs
More places added to FTTP rollout. A href="https: www.ispreview.co.uk index.php 2021 10 170-new-additions-to-openreachs-uk-fttp-rollout-programme.html" target="_self" https: www.ispreview.co.uk index.php 2021 10 170-new-additions-to-openreachs-uk-fttp-rollout-programme.html A	places added fttp rollout
Hi This is an odd one but hopefully I can get help, many years ago I use to have a BT account, with this account I was given some space for free to make a basic website, which I did, the address was. A href="https: www.btinternet.com ~myusername" target="_blank" https: www.btinternet.com ~myusername A A few years ago it disappeared, is there any way I can retrieve the files i uploaded to this website, it has some nostalgia on there that I would like to retrieve, or are they lost forever?	hi odd one hopefully get help many years ago use bt account account given space free make basic website did address was https myusername years ago disappeared way retrieve files uploaded website nostalgia would like retrieve lost forever
Hi Every We have seen reports of a new A href="https: community.bt.com t5 Announcements-Guides-Community Be-aware-of-scams-What-to-look-out-for-and-how-to-spot-a-scam td-p 1979456" target="_self" Smishing A (Phishing by SMS) whereby BT customers are being targe	hi every seen reports new smishing phishing sms whereby bt customers targe
Hi Im hoping someone can help me and point me in the right direct regarding BT block my website from users. 18 months ago (maybe longer) our website was hacked with a "phishing" virus (grrr have people nothing better to do) We had a new site built, moved to a new host and had it confirmed many times as safe, but I believe BT still has it marked as Unsafe and people say they can't get access to our site. The site is A href="http: www.bigwoodys.co.uk" target="_blank" www.bigwoodys.co.uk A Virustotal.com has no one listing us as a threat	hi im hoping someone help point right direct regarding bt block website users 18 months ago maybe longer website hacked phishing virus grrr people nothing better do new site built moved new host confirmed many times safe believe bt still marked unsafe people say cant get access site site one listing us threat

Figure 5.6: Sample data before and after removal of URL

5.2.2 Punctuation removal

In the process of preparing dataset for topic modelling, the handling of punctuations presented a unique challenge. Unlike conventional text preprocessing pipelines, where punctuations are often one of the first elements to be removed, we chose to defer punctuation removal until after addressing HTML tags and URLs. This decision was driven by the specific characteristics of our dataset and the use of regular expressions in the cleaning process. Please find below list of punctuations. Regular expressions were a key tool for cleaning HTML tags and URLs.

```
List of punctuation characters in string library:
['!', '"', '#', '$', '%', '&', "'", '(', ')', '*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']',
'^', '_', '`', '{', '|', '}', '~']
```

Figure 5.7: Punctuations

Punctuation symbols such as ‘¡’, ‘¿’, ‘: ‘, and ‘ / ‘ were integral in crafting regular expression patterns that could effectively identify and capture these elements within the dataset. As such, removing punctuations prior to applying regular expressions would have compromised the ability of our expressions to accurately match and clean HTML tags and URLs. Punctuation symbols, especially those commonly used in HTML and URLs, have specific roles in defining the structure and syntax of these elements. Removing punctuations prematurely could have resulted in the fragmentation of these patterns, making it challenging to identify and clean them effectively. By keeping punctuations intact until the HTML tags and URLs were addressed, which helped preserve the integrity of these structural patterns within the dataset. Given data preprocessing pipeline followed a sequential approach, where HTML tags and URLs were addressed before punctuations. This ensured that the most complex and context-sensitive cleaning tasks were performed first, gradually simplifying the data cleaning process. By the time punctuations were handled, the dataset had already undergone significant transformations, making it easier to manage and maintain consistency in the textual content. Removing punctuations is a standard text preprocessing task that enhances the quality of textual data for various natural language processing tasks, including topic modelling. The decision to defer punctuation removal until after addressing HTML tags and URLs was a deliberate choice to maintain the integrity of regular expression patterns and to simplify the overall data preprocessing pipeline.

5.2.3 Handling Parsing Issue

While handling diacritical marks, an unexpected issue arose where the pound symbol (£) was parsed as ”aPS.” This parsing error introduced inconsistencies in the textual data and posed a potential challenge for our subsequent text mining and topic modelling analysis. While the exact amount represented by the pound symbol was not a primary focus of our analysis, the discussions related to billing and financial matters were important for research. To address this parsing issue and ensure that the textual data accurately reflected the conversations about monetary topics in the BT forum, developed a tailored solution. A custom regular expression

designed to systematically identify and replace instances of ”aPS” with the word ”money” throughout the dataset. The regular expression crafted for this purpose targeted the specific string ”aPS.” It was designed to match and replace ”baPS” with ”money” wherever it appeared in the text. By using regular expressions in this manner, we effectively restored the context and meaning associated with the pound symbol (£) within the dataset.

Raw Data sample	After replacing to word pounds
service want bt line installed working bt game town here connect broadband supplier need phone mobile broadband good get 3g plans upgrade 3g years continue wait possible pay aPS140 line installation get services another provider stuck statebacked provider thanks	service want bt line installed working bt game town here connect broadband supplier need phone mobile broadband good get 3g plans upgrade 3g years continue wait possible pay pounds line installation get services another provider stuck statebacked provider thanks
paid aPS25 subscription bt sports specifically want watch porier vs mcgregor fight sunday morning france moment still stream here	paid pounds subscription bt sports specifically want watch porier vs mcgregor fight sunday morning france moment still stream here
hi broadband internet either wired ethernet cable without 09 mbps 003 mbps evening since 730 pm possible usually speed range 6 mbps 11 mbps rely internet connection work many thanks comments hightly frustrating pay entire aPS 2499 service	hi broadband internet either wired ethernet cable without 09 mbps 003 mbps evening since 730 pm possible usually speed range 6 mbps 11 mbps rely internet connection work many thanks comments hightly frustrating pay entire pounds service

Figure 5.8: Sample data before and after handling Parsing issue

5.2.4 Handling numeric data

In the next step of data preprocessing, removed all numeric data from the ”Description” column. This decision was made to ensure that the analysis is focused solely on the textual content of user comments and discussions. Numeric data, while valuable in other contexts, could introduce noise and potentially reduce the performance of text-based models. The primary objective was to maintain the textual relevance of our dataset. Numeric data, such as date, billing amount, data speed, mobile numbers, and amount might not contribute meaningfully to discussions but could disrupt the analysis. By removing such data, aimed to keep the essence of the discussions intact. The process of removing numeric data was systematic and involved the use of regular expressions. These expressions allowed us to identify and replace numerical values and patterns with empty strings within the ”Description” columns. The outcome was a dataset that contained only text, simplifying subsequent text-based analyses.

After replacing to word pounds	Raw Data sample
received notification today allegedly bt asking confirm acceptance new terms conditions tomorrow 5 1 22 lose access information question genuine	received notification today allegedly bt asking confirm acceptance new terms conditions tomorrow lose access information question genuine
hi commun hope everyone keeping safe well canuat quite believe days christmas long see end 2021 year b	hi commun hope everyone keeping safe well canuat quite believe days christmas long see end year b
ive email saying direct debit is longer active nees set again contains last 3 numbers account check genuine	ive email saying direct debit is longer active nees set again contains last numbers account check genuine
delete profile amp amp data bt	delete profile amp amp data bt
probably 64 million dollar question anybody aware plans bt release x box thanks	probably million dollar question anybody aware plans bt release x box thanks
imagealtsft staying safe onlinehero full 1920x800 pxjpg stylewidth 999px img srchttps t5 image serverpage imageid 74858i7b10da2a47760fa1 imagesize largevv2amp amp px999 rolebutton titlesft staying safe onlinehero full 1920x800 pxjpg altsft staying safe onlinehero full 1920x800 pxjpg span amp n	imagealtsft staying safe onlinehero full x pxjpg stylewidth px img srchttps t image serverpage imageid ibdaafa imagesize largevvamp amp px rolebutton titlesft staying safe onlinehero full x pxjpg altsft staying safe onlinehero full x pxjpg span amp n

Figure 5.9: Sample data after handling numeric values

5.2.5 Lemmatization

In the subsequent stages of our data preprocessing, we opted for lemmatization over stemming. Stemming is a technique that reduces words to their root or base form by removing suffixes, which can sometimes lead to the creation of non-standard or non-English words. For instance, in the case of the words ”walks” and ”walking,” stemming would reduce both words to ”walk,” which is a valid English word. However, when we encounter words like ”console” and ”consoling,” stemming may produce ”consol,” which is not a recognized English word. To address this issue and ensure that our text remained in proper English, we employed lemmatization. Lemmatization, unlike stemming, reduces words to their lemma or dictionary form, ensuring that the resulting words are valid English words. Lemmatization helps to maintain the linguistic integrity of the text while still reducing words to their base forms. This choice was essential for our text mining and topic modelling tasks, as it allowed to analyse the data with a focus on meaningful English words, enhancing the accuracy and interpretability of our results. Used WordNetLemmatizer from NLTK library to lemmatize sentences. Please find below sample text after lemmatization.

Sample Raw Data	After Lemmatization
pg issue previous bt community id which active left	pg issue previous bt community id which active left
received notification today allegedly bt asking confirm acceptance new terms conditions tomorrow lose access information question genuine	receive notification today allegedly bt ask confirm acceptance new term condition tomorrow lose access information question genuine
hi commun hope everyone keeping safe well canuat quite believe days christmas long see end year b	hi commun hope everyone keep safe well canuat quite believe day christmas long see end year b
ive email saying direct debit is longer active nees set again contains last 3 numbers account check genuine	ive email say direct debit be longer active nees set again contains last number account check genuine
hello call indian gentleman purporting bt forceful said getting right speed router tried get go bt speed checker via google smalt rat hung up looking number called non existent local number blocked it joined bt upgraded fttp last week potentially plausible knew name coincidence scammers hacked bt me anybody got thoughts info many thanks john	hello call indian gentleman purport bt forceful say get right speed router try get go bt speed checker via google smalt rat hang up look number call non existent local number block it join bt upgrade fttp last week potentially plausible knew name coincidence scammer hack bt me anybody get thought info many thanks john
hi every please aware new scam rounds scams involve text whatsapp message person claiming family member common aim encouraging recipient transfer mo	hi every please aware new scam round scams involve text whatsapp message person claim family member common aim encouraging recipient transfer mo

Figure 5.10: Data sample before and after removing Lemmatizing text

5.2.6 Removal of Frequent words

In the subsequent phase of data preprocessing, took measures to filter out frequent words from domain-specific corpus that were deemed less relevant to analysis. Stopwords function has the specific list of words in its corpus and may not remove all the least important words hence created a customized list of words These words were identified and designated for removal to refine the dataset and focus on more informative content. The list of words designated for removal

included: ’set,’ ’cant,’ ’look,’ ’able,’ ’also,’ ’like,’ ’please,’ ’whole,’ ’say,’ ’new,’ ’recently,’ ’still,’ ’know,’ ’back,’ ’smart,’ ’anyone,’ ’and,’ ’get,’ ’hi,’ ’would,’ ’go,’ ’thanks,’ ’im,’ ’ive,’ ’please,’ ’amp,’ ’seem,’ ’see,’ ’also,’ ’like,’ ’could,’ ’it,’ ’however,’ ’make,’ ’fttp,’ and ’re.’ By removing these frequently occurring but less informative words from our corpus, aimed to reduce noise and focus our analysis on the most meaningful and relevant textual content. This process helped improve the quality and accuracy of text mining and topic modelling efforts, allowing us to extract valuable insights from the domain-specific data.

5.2.7 Handling missing values

In the subsequent preprocessing step, conducted a careful examination of the data to identify and handle null values. It was observed that, during the preprocessing procedures, certain rows lost all their content and became empty. To maintain the dataset’s integrity and ensure meaningful analysis, took the necessary action to remove these empty rows. To identify and remove the empty rows, we utilized the isnull() method. This method allowed us to efficiently detect rows with missing or null values within the dataset.

5.3 Data Exploration after Preprocessing

Figure 5.11: Word cloud representation of user comments

1. In this analysis, the word cloud generated from your community data reveals key topics and themes that dominate user discussions. ”BT” is the most prominent term, highlighting that discussions related to BT services and experiences are a primary focus. Users are actively engaged in addressing problems, as indicated by the frequent appearance of ”try.” ”Fibre” signifies significant interest in fiber-optic broadband, while ”TV” suggests a focus on television services. The presence of ”issue” underscores discussions around challenges and concerns, and ”home” and ”email” point to home network and email-related topics. ”Broadband” highlights discussions about broadband services, and ”switch” implies users exploring service or provider alternatives. These terms collectively provide valuable

insights into user interests, issues, and areas of discussion within the community. Further context analysis can offer a comprehensive understanding of user experiences and needs.

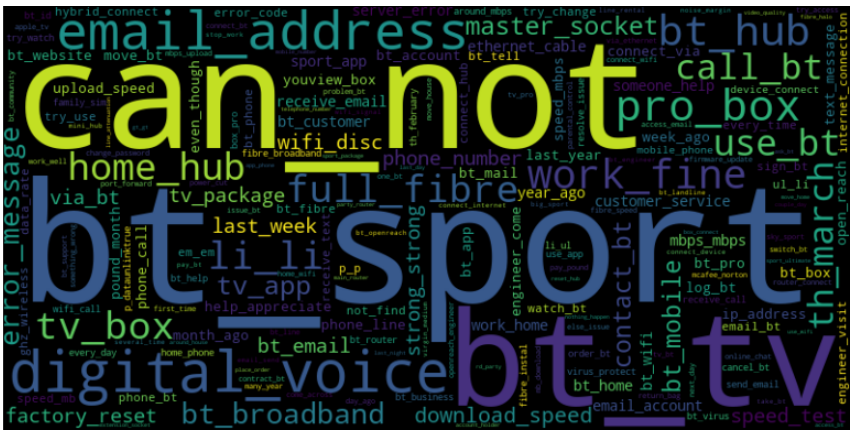

Figure 5.12: Bigram representation user comments

2. In this analysis, the extraction of bigrams from the Description column has provided a deeper understanding of the context within community data. Several significant bigrams stand out, shedding light on specific user discussions and concerns. ”bt_sport” indicates a focus on BT’s sports offerings, suggesting discussions related to sports content and subscriptions. ”can_not” is a prominent bigram, indicating that users often discuss issues or challenges they cannot resolve. ”bt_tv” suggests that users frequently mention BT’s television services, possibly discussing packages, channels, or viewing experiences. ”digital_voice” implies a focus on digital voice services, which could involve discussions about landlines or VoIP. ”email_address” highlights discussions related to email addresses and potential issues users encounter. ”Work_fine” suggests users discussing their experiences with BT’s services, potentially indicating satisfaction or issues. ”pro_box” could relate to conversations about BT’s set-top boxes and related features. ”bt_broadband” indicates frequent discussions about BT’s broadband services, encompassing speed, reliability, and performance. Lastly, ”error_message” signifies users encountering and discussing error messages, which could be related to various BT services or products. These prominent bigrams offer valuable insights into the specific topics and issues that engage in BT community, providing a foundation for deeper analysis and addressing user needs effectively.

3. The analysis of trigrams from the Description column reveals critical insights into the most prevalent discussions and concerns within community data. Notably, ”bt_sport_app” signifies extensive conversations regarding BT’s sports application, suggesting a keen interest in sports content and related app usage among users. ”bt_tv_app” emphasizes discussions about BT’s television app, indicating user engagement with the app’s features and functionalities. ”can_not_find” is a significant trigram, indicating users often face difficulties in locating specific information or resolving issues, warranting potential improvements in user guidance and support. ”bt_tv_box” suggests conversations about BT’s

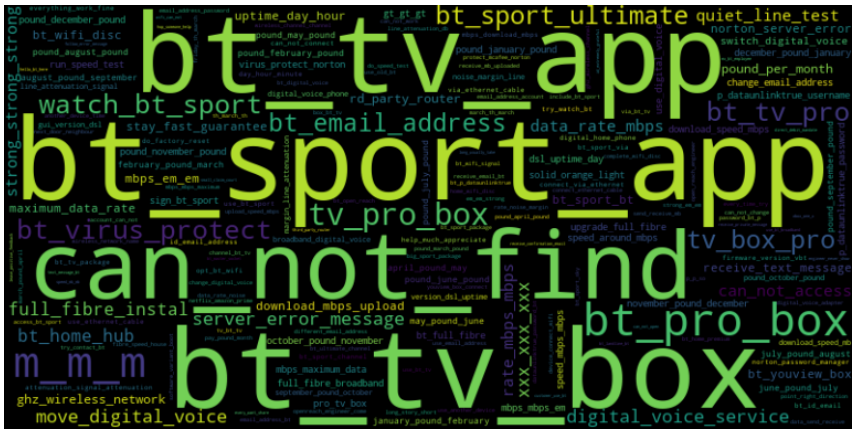

Figure 5.13: Representation of user comments using Trigram

TV set-top boxes, possibly related to setup, usage, or troubleshooting. These prominent trigrams provide valuable insights for enhancing user experiences, tailoring support, and addressing common pain points effectively within BT community.

Figure 5.14: Word cloud representation of the Topics discussed in community

4. The word cloud generated from the Title column of the community data offers a snapshot of the most frequently occurring terms, providing insights into the prevalent topics or subjects of discussion within the community. ”Sport” emerges as a prominent term, indicating a substantial interest in sports-related content or services among community members. ”Email” signifies a focus on email-related topics, which may encompass issues, solutions, or queries concerning email services. The terms ”box,” ”wifi,” ”home,” and ”broadband” suggest that discussions about home network setups, broadband services, and associated devices are central to the community’s discourse. ”Pro,” ”hub,” and ”voice” potentially point to discussions regarding advanced BT products and services, while ”alert,” ”line,” and ”norton” could imply conversations about security alerts, connectivity issues, or Norton antivirus software. This analysis underscores the diverse range of topics and user interests within the community, offering valuable insights for content creation, support, and community management efforts.

5. The word cloud derived from the bigrams in the Title column of the community data pro-

Figure 5.15: Bigram representation of the Topics discussed in community

vides a deeper understanding of the context and topics initiated by community members when starting conversations. ”BT Sport” and ”BT TV” emerge as the most prevalent bigrams, indicating a significant focus on BT’s sports and television services. ”Digital Voice” suggests discussions related to voice services, while ”Scam Alert” implies a shared concern about potential scams within the community. ”Full Fibre” likely signifies conversations about high-speed broadband options, and ”Pro Box” might relate to discussions regarding advanced BT set-top boxes or equipment. ”Virus Protect” points to dialogues concerning virus protection and cybersecurity measures, while ”Sport App” and ”BT Virus” could be associated with discussions about BT’s sport-related applications and antivirus software. The presence of ”Error Code” suggests that members often seek help or discuss issues by referencing error codes. This analysis illuminates the diverse array of topics initiated by community members and highlights their specific areas of interest and concern, offering valuable insights for community management and content planning efforts.

Chapter 6

Results and Discussion

This section serves as the heart of this report, where the findings, detailed data analysis, and a meaningful discussion are presented to extract insights from the results. This section includes various graphs and visual representations that illuminate patterns, trends, and relationships within the dataset. Through meticulous analysis, the significance of these findings is explored, drawing connections to the research objectives and hypotheses. By scrutinizing the data and exploring its implications, a comprehensive understanding of the underlying phenomena and their potential real-world applications is provided.

6.1 Topic modelling using LDA

The analysis of the LDA model applied to the BT community forum dataset provides insights into the prominent themes and discussions within the forum. Given the nature of the dataset being sourced from the BT community forum, the identified topics can be closely linked to BT services and user interactions. Each topic is characterized by a set of representative words. Let's delve into the interpretation of each topic: 5 topics generated from model.

```
Topic 1:
bt, tv, phone, box, hub, use, work, broadband, number, connect

Topic 2:
phishing, launch, report, target, style, xxx, identical, annex, recycle, vm

Topic 3:
channel, ultimate, guide, broadcast, eurosport, sport, hdr, cast, scam, discovery

Topic 4:
email, norton, try, account, bt, error, address, password, message, nan

Topic 5:
dect, voicemail, compatible, hit, alarm, pgt, head, highlight, round, payroll
```

Figure 6.1: Representation of 5 topics generated by LDA model

1. Topic 1 BT Services and Connectivity This topic revolves around discussions related

to various BT services and connectivity concerns. Representative words such as ”bt” (British Telecom), ”tv,” ”phone,” ”broadband,” and ”connect” strongly suggest conversations about BT service offerings, device compatibility, troubleshooting connectivity issues, and optimizing the usage of BT services.

2. Topic 2 Security and Online Threats The second topic centers on matters of security and potential online threats. The presence of words like ”phishing,” ”report,” ”target,” and ”xxx” indicates discussions related to identifying and addressing phishing attempts, sharing reports of suspicious activities, and providing advice on safeguarding personal information within the context of BT services.

3. Topic 3 Broadcasting and Entertainment This topic pertains to discussions encompassing broadcasting, entertainment, and content available through BT services. Representative words such as ”channel,” ”eurosport,” ”sport,” and ”broadcast” imply conversations about TV channel options, sports events coverage, and entertainment choices accessible through BT offerings.

4. Topic 4 Technical Support and Account Management The fourth topic focuses on technical support and account management matters. The inclusion of words like ”email,” ”norton,” ”error,” ”address,” and ”password” indicates conversations related to technical issues users encounter with BT services, seeking solutions for email-related problems, account access errors, and password recovery assistance.

5. Topic 5 Communication Devices and Technology The fifth topic is centered around discussions about communication devices and technology. Representative words such as ”dect,” ”voicemail,” and ”compatible” point to conversations about the compatibility of communication devices, setting up voicemail services, and exploring advancements in communication technology within the context of BT services.

6.2 Word probability distribution

In this analysis, we generated five topics using the LDA model, and for each topic, we calculated the word probabilities to understand the most characteristic words associated with that topic. Word probability distribution in the context of topic modelling, such as Latent Dirichlet Allocation (LDA), refers to the likelihood or probability of individual words being associated with specific topics within a given dataset. Each word in the dataset is assigned a probability score for its association with each topic. High word probabilities indicate the importance of specific words within a topic.

Topic 1: bt, tv, phone, box, hub, use, work, broadband, number, connect Word Probabilities: 'bt': '0.205', 'tv': '0.110', 'phone': '0.102', 'box': '0.099', 'hub': '0.093', 'use': '0.085', 'work': '0.083', 'broadband': '0.076', 'number': '0.074', 'connect': '0.073' This topic is strongly associated with telecommunications and broadband services, as indicated by the high word probabilities for terms like 'bt,' 'tv,' 'phone,' 'broadband,' and 'hub.' Users discussing topics related to their BT services, such as TV, phone, and broadband, are likely contributing to this topic. The presence of words like 'connect' and 'number' suggests discussions on technical issues and customer support.

Topic 2: phishing, launch, report, target, style, xxx, identical, annex, recycle, vm Word Probabilities: 'phishing': '0.158', 'launch': '0.138', 'report': '0.128', 'target': '0.095', 'style': '0.090', 'xxx': '0.089', 'identical': '0.082', 'annex': '0.074', 'recycle': '0.074', 'vm': '0.072' This topic revolves around online security and potential threats like phishing, as indicated by high word probabilities for terms like 'phishing,' 'report,' and 'target.' Discussions in this topic may involve users sharing their experiences with phishing attempts or seeking advice on how to stay safe online. Words like 'style' and 'recycle' might suggest discussions related to identifying phishing attempts and protecting personal information.

Topic 3: channel, ultimate, guide, broadcast, eurosport, sport, hdr, cast, scam, discovery Word Probabilities: 'channel': '0.152', 'ultimate': '0.144', 'guide': '0.120', 'broadcast': '0.111', 'eurosport': '0.111', 'sport': '0.081', 'hdr': '0.076', 'cast': '0.075', 'scam': '0.066', 'discovery': '0.064' This topic is likely cantered around broadcasting and sports, with high word probabilities for terms like 'channel,' 'ultimate,' 'eurosport,' and 'sport.' Users may be discussing TV channels, sports events, and broadcasting technologies within this topic. The presence of 'scam' suggests that users might also discuss potential scams related to sports or broadcasting services.

Topic 4: email, norton, try, account, bt, error, address, password, message, nan Word Probabilities: 'email': '0.182', 'norton': '0.136', 'try': '0.110', 'account': '0.093', 'bt': '0.091', 'error': '0.088', 'address': '0.082', 'password': '0.077', 'message': '0.075', 'nan': '0.066'

This topic is associated with email and account management, with high word probabilities for terms like 'email,' 'norton,' 'account,' and 'password.' Users may discuss email-related issues, account security, and troubleshooting within this topic. The presence of 'error' suggests that users may seek assistance with email or account-related errors.

Topic 5: dect, voicemail, compatible, hit, alarm, pgt, head, highlight, round, payroll Word Probabilities: 'dect': '0.168', 'voicemail': '0.134', 'compatible': '0.101', 'hit': '0.092', 'alarm': '0.092', 'pgt': '0.089', 'head': '0.083', 'highlight': '0.081', 'round': '0.081', 'payroll': '0.080'

This topic likely encompasses discussions related to technology and devices, as indicated by high word probabilities for terms like 'dect,' 'voicemail,' 'compatible,' 'alarm,' and 'head.' Users

might be sharing their experiences with various devices and seeking advice on compatibility and troubleshooting. The presence of 'payroll' could suggest discussions related to payroll software or processes.

6.3 Conclusion

In conclusion, the application of LDA topic modelling to the BT Community dataset has revealed five distinct topics, each with its own characteristic words. These topics cover a range of subjects, including telecommunications, online security, broadcasting, email management, and technology. Understanding these topics can be valuable for community moderators and researchers to categorize and analyze user-generated content effectively.

6.4 Topic probability distribution

Topic probability distribution refers to the likelihood or probability of a document or a piece of text belonging to each of the predefined topics within an LDA (Latent Dirichlet Allocation) model. The topic probability distribution provides insights into how relevant each of these topics is to a given document or text. LDA model assigns a probability score to each of the five topics. These probability scores indicate the degree to which the document aligns with each topic. A document is often associated with multiple topics, but the probability distribution quantifies the strength of association with each topic. Topic Probabilities: Topic 1: 0.0558 Topic 2: 0.0551 Topic 3: 0.0554 Topic 4: 0.7786 Topic 5: 0.0551

6.5 Observations and Analysis

- : From the probability scores, it is evident that Topic 4 stands out as highly associated with the BT Community dataset, with a probability score of approximately 0.7786. This suggests that a significant portion of the content within the dataset is predominantly related to Topic 4.

- Topics 1, 2, 3, and 5 all exhibit relatively low probability scores, each around 0.055. This indicates that these topics have a minimal presence in the dataset or are less dominant in the discussions compared to Topic 4.

- The high probability score of Topic 4 suggests that most of the content within the dataset is closely related to this topic. It would be essential to further investigate the characteristics and content of Topic 4 to understand the specific subject matter it represents.

- Given the dominance of Topic 4, it is essential to explore its content and characteristics to gain insights into the primary concerns or interests of the BT Community forum users.

```
Topic 1:
bt, tv, phone, use, try, box, email, work, hub, help

Topic 2:
commun, platform, parental, discovery, used, phishing, dect, hdmi, happy, outage

Topic 3:
db, pm, firestick, mode, cricket, min, cast, stutter, employee, regular

Topic 4:
nan, reward, mcafee, uninstall, norton, photo, car, antivirus, software, forward

Topic 5:
voicemail, spam, active, round, pas, domain, imap, email, sender, present

Topic 6:
email, mail, inbox, spam, office, outlook, essential, flash, gmail, eurosport

Topic 7:
bag, return, atmos, label, dolby, round, equipment, sms, transcript, pg

Topic 8:
netflix, http, opt, firefox, credit, instruction, whatuas, pgt, john, ancient

Topic 9:
feedback, apple, member, pas, positive, customer, phishing, report, iphone, compliment

Topic 10:
road, large, itv, postcode, chromecast, ps, match, gfast, impossible, rugby
```

Figure 6.2: Representation of 10 topics generated by LDA model

Analyzing the characteristic words and discussions within Topic 4 can help identify the central themes.

Content Recommendation and User Experience: Understanding the dominance of Topic 4 can also be leveraged for content recommendations and enhancing the user experience. Providing users with relevant content related to the highly associated topic can improve their satisfaction and engagement. In conclusion, the Topic Probability Distribution reveals that Topic 4 is highly associated with the BT Community dataset. This finding underscores the importance of analyzing and understanding the content within Topic 4 to address the community's central concerns effectively. Additionally, it highlights the diversity of topics within the dataset, each representing specific areas of interest and discussion among BT users.

6.6 Comparative Analysis of LDA Topics: 5 Topics vs. 10 Topics

By progressing from 5 to 10 topics, we aimed to unravel a more comprehensive understanding of user interactions and issues deliberated within the forum. The analysis below elucidates how the expansion from 5 to 10 topics has furnished us with more profound insights.

1. Mirroring the 5-topic model's theme, Topic 1 in the 10-topic model underscores user engagement with BT's technological aspects, from TV to broadband. The focus remains on problem-solving and connectivity enhancement.

2. A new facet emerges in the 10-topic model with Topic 2. It encapsulates discussions about the community platform itself, parental control features, and even mentions of "phishing." This underscores an evolving concern for user safety within the platform.

3. Topic 3 dives into diverse areas, touching on firestick usage, cricket discussions, and employee-related matters. This demonstrates an expansion into broader conversations surrounding user interests and platform engagement.

4. Topic 4 captures software-centric discussions encompassing rewards, McAfee, Norton, and antivirus software. The theme extends into software management and optimization, revealing users' concerns about security and user experience.

5. The 10-topic model sheds light on email communication dynamics. Topics 5 and 6 dissect email usage intricacies, including spam management, inbox organization, and integration with Microsoft Outlook and Gmail.

6. An entirely new dimension unfolds with Topic 7, which delves into product returns, streaming equipment, and even technical transcripts. This highlights users' interactions beyond the primary services.

7. The 10-topic model expands entertainment insights, with Topic 8 discussing streaming platforms like Netflix, browser choices, and internet browsing behavior, indicating a more holistic engagement with technology.

8. The introduction of Topic 9 reflects an increased focus on user opinions, Apple product discussions, and customer feedback. This suggests BT's proactive approach in understanding user preferences.

9. The final topic, Topic 10, draws attention to geographical relevance, entertainment preferences, and even technological feasibility. It unveils a nuanced understanding of user diversity and localized interests.

Both sets of topics reflect different aspects of discussions within the BT community. The 5-topic model seems to capture broader themes like services, phishing, broadcasting, email, and technical terms. On the other hand, the 10-topic model provides more granularity and covers a wider range of topics including device modes, rewards, email-related issues, customer feedback, and entertainment content.

6.7 Coherence and perplexity value of the topics generated by LDA model.

The image shows the results of a topic modelling experiment, where the number of topics is varied, and the coherence and perplexity scores are calculated. Coherence measures how semantically related the words in a topic are. A higher coherence score indicates that the words

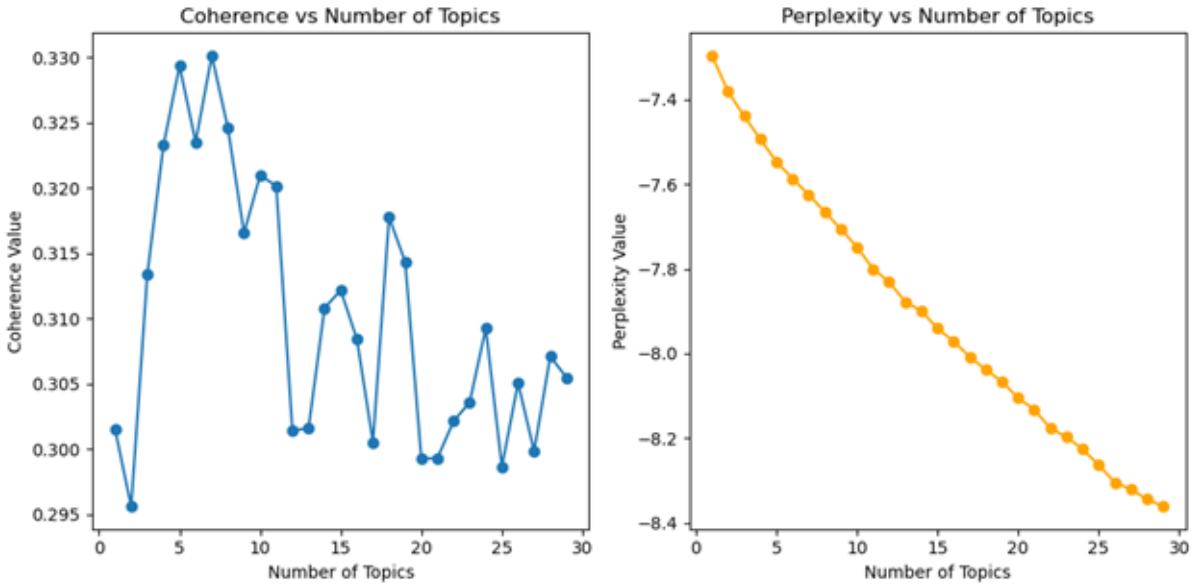

Figure 6.3: Performance of LDA model based on coherence and perplexity.

in the topic are more closely related. Perplexity measures how well the topic model predicts the words in the corpus. A lower perplexity score indicates that the model is better at predicting the words. The graph on the left shows the coherence score as a function of the number of topics. The coherence score increases as the number of topics increases, but the rate of increase slows down as the number of topics gets higher. This is because it becomes more difficult to find topics that are semantically coherent when there are a lot of topics. The graph on the right shows the perplexity score as a function of the number of topics. The perplexity score decreases as the number of topics increases. This is because a model with more topics can better capture the different patterns in the corpus. The optimal number of topics is the one that strikes a balance between coherence and perplexity. In this case, the optimal number of topics is around 4 or 5. This is because the coherence score is still relatively high at this point, and the perplexity score is not too high. Overall, the results of this experiment suggest that the optimal number of topics for this corpus is around 4 or 5. This is based on the trade-off between coherence and perplexity.

6.8 NMF Model

1. Topic 1: Phone and Mobile Service This topic seems to be focused on phone-related discussions, including numbers, digital services, mobile phones, and landlines. 2: TV and Entertainment This topic revolves around television services, including channels, sports, apps, and packages, suggesting discussions related to entertainment options.

2. Topic 3: Email and Account Issues Discussions in this topic are centered around email and account-related issues, including addresses, passwords, and error messages.

3. Topic 4: Home Network and Devices This topic involves discussions about home networks, Wi-Fi, routers, hubs, and connectivity problems for various devices.

Figure 6.4: Representation of 5 topics generated by NMF model

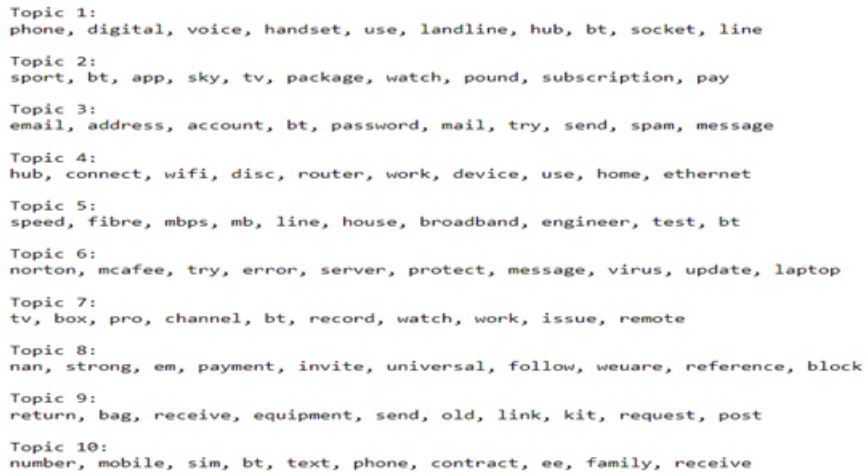

Figure 6.5: Representation of 10 topics generated by NMF model

4. Topic 5: Broadband Speed and Testing The focus here is on broadband speed and performance, including testing, speed values in Mbps, and related issues.

1. Topic 1: Phone and Digital Services Like the previous set, this topic encompasses discussions about phone services, digital handsets, landlines, and hubs.

2. Topic 2: Sports and TV Packages This topic maintains discussions about sports, TV packages, apps, and related subscription-based services.

3. Topic 3: Email and Account Management Similar to the 5-topic model, this topic deals with email, account details, passwords, and related issues.

4. Topic 4: Home Network and Connectivity This topic continues to include discussions on home networks, Wi-Fi, routers, devices, and connectivity challenges.

5. Topic 5: Broadband Speed and Performance Consistent with the previous set, this topic focuses on broadband speed, Mbps values, and performance evaluation.

6. Topic 6: Cybersecurity and Virus Protection This topic introduces cybersecurity discussions, including antivirus software like Norton and McAfee, server errors, and virus protection.

7. Topic 7: TV Services and Issues Discussions about TV services, channels, watching options, and related issues are clustered within this topic.

8. Topic 8: Miscellaneous and Technical Issues This topic covers various issues, including payments, references, blocks, and more, suggesting a collection of diverse technical discussions.

9. Topic 9: Equipment and Returns This topic centres around equipment, returns, and sending old items back, indicating discussions about returning BT-related equipment.

10. Topic 10: Mobile Services and Contracts This topic is focused on mobile services, contracts, text messages, and related discussions involving mobile phones.

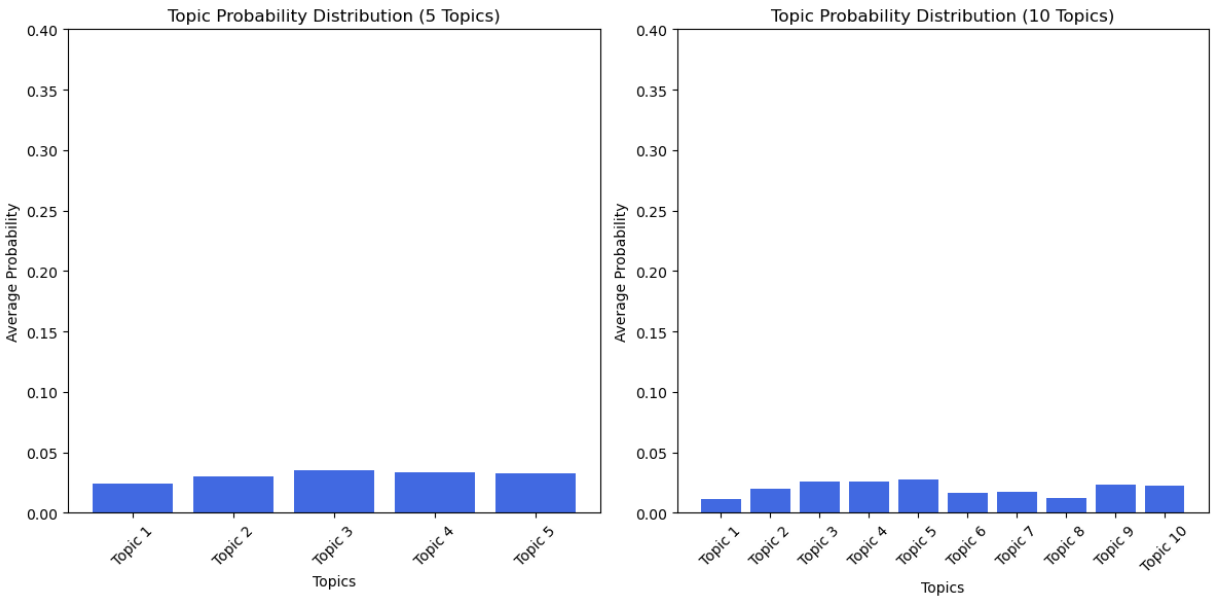

Figure 6.6: Distribution of 5 VS 10 topics generated by NMF model

The NMF model effectively captures varied themes in the BT Community data, ranging from telecommunication services to technical troubleshooting and entertainment preferences. The 10-topic model introduces more specialized discussions, including cybersecurity, equipment returns, and diverse technical challenges. Both models showcase user interests, concerns, and inquiries that are commonly encountered in a community forum setting.

6.9 Latent semantic Allocation(LSA)

In this analysis, we have generated five as well ten topics using the LSA model. Based on the topics generated using this model, Topic 1 seems to be related to customer support or technical

issues, with terms like ”bt,” ”tv,” ”email,” ”phone,” and ”hub” suggesting discussions about troubleshooting problems with services and devices.

Topic 2 appears to revolve around television and entertainment services, with terms like ”tv,” ”box,” ”sport,” ”channel,” and ”watch” indicating discussions about TV packages, sports channels, and streaming apps.

Topic 3 seems to be centered on email and account-related issues, with terms like ”email,” ”account,” ”norton,” and ”password” suggesting discussions about email problems, account settings, and security concerns.

Topic 4 appears to focus on technical issues related to networking and connectivity, with terms like ”norton,” ”connect,” ”wifi,” ”router,” and ”error” indicating discussions about network setups, connectivity problems, and device configurations.

Topic 5 appears to be related to internet speed and broadband performance, with terms like ”speed,” ”fibre,” ”mbps,” and ”download” suggesting discussions about internet speed tests, broadband packages, and network performance. The Y-axis represents the words or terms that

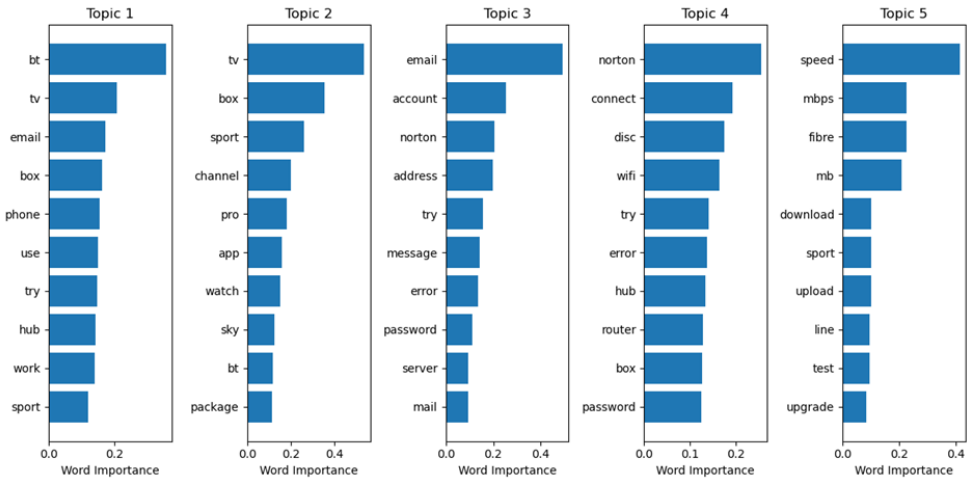

Figure 6.7: LSA Topic-Terms Distribution

are associated with each topic. Each horizontal bar in the graph corresponds to a word, and the Y-axis labels indicate those words. Specifically, for each topic (Topic 1, Topic 2, etc.), the Y-axis displays the top 10 words that are most strongly associated with that topic based on their importance scores within that topic. These words are the most distinctive terms for each topic.

The X-axis represents the importance or weight of each word within its corresponding topic. It quantifies how significant each word is in defining the topic. The X-axis labels are not explicitly shown in the code because it represents the importance values. The more to the right a bar is on the X-axis, the more important that word is within the topic. These probabilities represent the likelihood that Document #1 belongs to each of the topics. The highest probability is associated with Topic 1, suggesting that Document #1 is most strongly related to discussions about topics such as ”bt,” ”tv,” ”email,” ”phone,” and ”hub.” On the other hand, the lowest probability is associated with Topic 2, indicating that this document is less relevant to discussions about ”tv,” ”box,” ”sport,” and related terms. The other topics fall in between, with varying degrees of

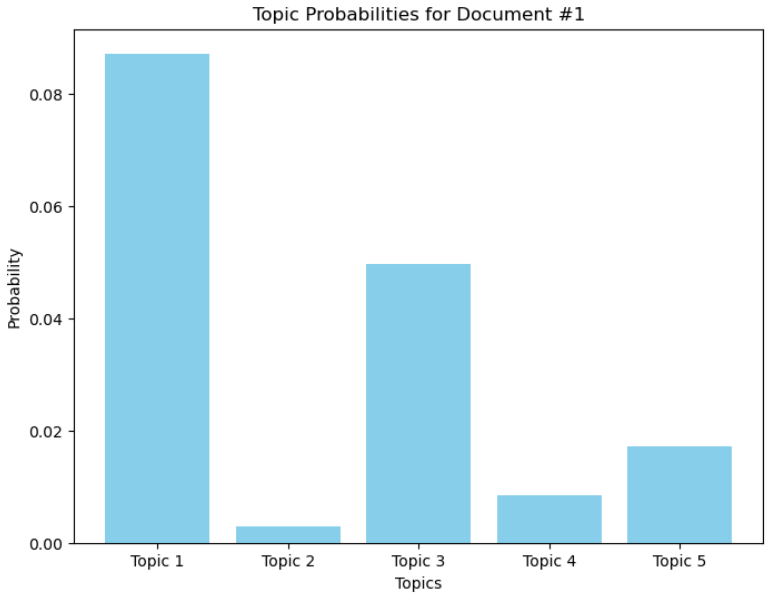

Figure 6.8: LSA topic probabilities for document

relevance to the document's content.

Chapter 7

Conclussion

In conclusion, the analysis of the BT community forum dataset using various topic modeling techniques, including LDA, NMF, and LSA, has provided valuable insights into the discussions and themes prevalent within the forum. While all three models have their strengths and advantages, the NMF (Non-Negative Matrix Factorization) model, particularly in its 10-topic configuration, has demonstrated its ability to provide comprehensive and specialized insights based on the BT community data.

The 10-topic NMF model excels in uncovering diverse and nuanced themes, ranging from cybersecurity and equipment returns to mobile services and contracts. These specialized topics offer a deeper understanding of the multifaceted nature of user-generated content within the forum. This granularity and specificity make NMF a robust choice for analyzing complex and multifaceted datasets like those found in community forums.

While LDA and LSA have also contributed valuable insights, NMF's capacity to identify highly specific topics and its ability to capture the unique intricacies of the BT community discussions make it the model that has provided the most comprehensive and insightful results for this particular dataset.

Ultimately, the choice of the most suitable topic modeling technique should align with the specific goals and requirements of the analysis. In the context of the BT community data, NMF's ability to uncover specialized topics and its comprehensive insights make it the preferred model for extracting meaningful information from user-generated discussions.

References

[1] Martin J Ball. "A diacritic range for phonation types". In: *Journal of the International Phonetic Association* 18.1 (1988), pp. 39–40.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[3] Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. "Applications of topic models". In: *Foundations and Trends® in Information Retrieval* 11.2-3 (2017), pp. 143–296.

[4] Laura V Galvis Carreño and Kristina Winbladh. "Analysis of user comments: an approach for software requirements evolution". In: *2013 35th international conference on software engineering (ICSE)*. IEEE. 2013, pp. 582–591.

[5] Yong Chen et al. "Affinity regularized non-negative matrix factorization for lifelong topic modeling". In: *IEEE Transactions on Knowledge and Data Engineering* 32.7 (2019), pp. 1249–1262.

[6] Yong Chen et al. "Experimental explorations on short text topic mining between LDA and NMF based Schemes". In: *Knowledge-Based Systems* 163 (2019), pp. 1–13.

[7] Yong Chen et al. "Experimental explorations on short text topic mining between LDA and NMF based Schemes". In: *Knowledge-Based Systems* 163 (2019), pp. 1–13.

[8] Jaegul Choo et al. "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization". In: *IEEE transactions on visualization and computer graphics* 19.12 (2013), pp. 1992–2001.

[9] Dasu Dasari and P Suresh Varma. "Employing Various Data Cleaning Techniques to Achieve Better Data Quality using Python". In: *2022 6th International Conference on Electronics, Communication and Aerospace Technology*. IEEE. 2022, pp. 1379–1383.

[10] Rundong Du et al. "DC-NMF: nonnegative matrix factorization based on divide-and-conquer for fast clustering and topic modeling". In: *Journal of Global Optimization* 68 (2017), pp. 777–798.

[11] Roman Egger and Joanne Yu. "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts". In: *Frontiers in sociology* 7 (2022), p. 886498.

[12] Marti Hearst. "What is text mining". In: *SIMS, UC Berkeley* 5 (2003).

[13] Shashank Kapadia. “Evaluate topic models: Latent Dirichlet allocation (LDA)”. In: *Towards Data Science* (2019).

[14] Minnan Luo et al. “Probabilistic non-negative matrix factorization and its robust extensions for topic modeling”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[15] Archana Patel and Narayan C Debnath. *Data Science with Semantic Technologies: New Trends and Future Developments*. CRC Press, 2023.

[16] David Ramamonjisoa. “Topic modeling on users’s comments”. In: *2014 third ICT international student project conference (ICT-ISPC)*. IEEE. 2014, pp. 177–180.

[17] Serhad Sarica and Jianxi Luo. “Stopwords in technical language processing”. In: *Plos one* 16.8 (2021), e0254937.

[18] Aditi Sharan and Sifatullah Siddiqi. “A supervised approach to distinguish between keywords and stopwords using probability distribution functions”. In: *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. 2014, pp. 1074–1080.

[19] Henry Sweet. *A handbook of phonetics*. Vol. 2. Clarendon Press, 1877.

[20] JianYu Wang and Xiao-Lei Zhang. “Deep NMF topic modeling”. In: *Neurocomputing* 515 (2023), pp. 157–173.

[21] John C Wells. “Orthographic diacritics and multilingual computing”. In: *Language problems and language planning* 24.3 (2000), pp. 249–272.

[22] Data Mining: Statistics and More?, D. Hand, American Statistician, 52(2):112-118.

[23] R. Baeza-Yates and B. Ribeiro-Neto. "Modern Information Retrieval", second edition, Addison-Wesley, 2011.

[24] Kröll, M. Strohmaier, M. (2009). Analyzing Human Intentions in Natural Language Text. In Gil, Y., Fridman Noy, N. (Eds.), Proceedings of the 5th International Conference on Knowledge Capture (pp. 197-198). New York, NY, USA: ACM.

Sudalai Rajkumar(2018) Kaggle. https://www.kaggle.com/code/sudalairajkumar/getting-started-with-text-preprocessing/notebook [Accessed on 1/8/2023]

Michelle Chen(2020) A Guide: Text Analysis, Text Analytics Text Mining tds. https://towardsdatascience.com/a guide-text-analysis-text-analytics-text-mining-f62df7b78747 [Accessed on 8/9/2023]

Pure Speech technology, TextAnalysisTextMiningTextAnalyticsDifference. https://www.purespeechtechnology analysis-text-analytics-text-mining/ [Accessed on 8/9/2023]

Alnusyan, R., Almotairi, R., Almufadhi, S., Shargabi, A. A., and Alshobaili, J. (2020). "A semi-supervised approach for user reviews topic modelling and classification," in 2020 International Conference on Computing and Information Technology (Piscataway, NJ: IEEE), 1–5. doi: 10.1109/ICCIT-144147971.2020.9213721

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Available online at: https://www.wired.com/2008/06/pb-theory/ [Accessed 10/8/2023].

Cai, T., and Zhou, Y. (2016). What should sociologists know about big data. https://esymposium.isaportal.org/re should-sociologists-know-about-big-data/ [Acessed on 8/9/2023]

Grootendorst, M. (2020). BERTopic: Leveraging BERT and c-TF-IDF to Create Easily Interpretable Topics. Zenodo. doi: 10.5281/zenodo.4430182

[32]Wang, J., and Zhang, X.-L. (2021). Deep NMF Topic Modelling. Available online at: http://arxiv.org/pdf/2102.12998v1 [Accessed on 8/9/2023).

David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," Journal of machine Learning research.

Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park, "Utopian: User-driven topic modelling based on interactive nonnegative matrix factorization," IEEE transactions on visualization and computer graphics.

Nicolas Gillis and Stephen A Vavasis, "Fast and robust recursive algorithmsfor separable nonnegative matrix factorization," IEEE transactions on pattern analysis and machine intelligenc.

Abhishek Kumar, Vikas Sindhwani, and Prabhanjan Kambadur,"Fast conical hull algorithms for near-separable non-negative matrix factorization," in International Conference on Machine Learning.

Nicolas Gillis, "Successive nonnegative projection algorithm for robust nonnegative blind source separation," SIAM Journal on Imaging Sciences.

Y. Chen, J. Wu, J. Lin, R. Liu, H. Zhang, and Z. Ye, "Affinity regularized non-negative matrix factorization for lifelong topic modelling," IEEE Transactions on Knowledge and Data Engineering.

39] Tan, A.H., 1999, April. Text mining: The state of the art and the challenges. In Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases (Vol. 8, pp. 65-70).

40] Hotho, A., Nürnberger, A. and Paaß, G., 2005. A brief survey of text mining. Journal for Language Technology and Computational Linguistics, 20(1), pp.19-62.

41]Hearst, M., 2003. What is text mining. SIMS, UC Berkeley, 5. Jo, T., 2019. Text mining (Vol. 45). Cham, Switzerland: Springer International Publishing.

42] Vayansky, I. and Kumar, S.A., 2020. A review of topic modelling methods. Information Systems, 94, p.101582. Hu, Y., Boyd-Graber, J., Satinoff, B. and Smith, A., 2014. Interactive topic modelling. Machine learning, 95, pp.423-469.

43] Churchill, R. and Singh, L., 2022. The evolution of topic modelling. ACM Computing Surveys, 54(10s), pp.1-35. Barde, B.V. and Bainwad, A.M., 2017, June. An overview of topic modelling methods and tools. In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 745-750). IEEE.

44] Yu, H. and Yang, J., 2001. A direct LDA algorithm for high-dimensional data—with application to face recognition. Pattern recognition, 34(10), pp.2067-2070.

45] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modelling: models, applications, a survey. Multimedia Tools and Applications, 78, pp.15169-15211.

46] Martinez, A.M. and Kak, A.C., 2001. Pca versus lda. IEEE transactions on pattern analysis and machine intelligence, 23(2), pp.228-233.

47] Chen, L.F., Liao, H.Y.M., Ko, M.T., Lin, J.C. and Yu, G.J., 2000. A new LDA-based face recognition system which can solve the small sample size problem. Pattern recognition, 33(10), pp.1713-1726.

48] Egger, R. and Yu, J., 2022. A topic modelling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in sociology, 7, p.886498.

49] Kuang, D., Choo, J. and Park, H., 2015. Nonnegative matrix factorization for interactive topic modelling and document clustering. 50] Partitional clustering algorithms, pp.215-243.

Mifrah, S. and Benlahmar, E.H., 2020. Topic modelling coherence: A comparative study between LDA and NMF models using COVID'19 corpus. International Journal of Advanced Trends in Computer Science and Engineering, pp.5756-5761.

51] Luo, M., Nie, F., Chang, X., Yang, Y., Hauptmann, A. and Zheng, Q., 2017, February. Probabilistic non-negative matrix factorization and its robust extensions for topic modelling. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

52] Ramamonjisoa, D., 2014, March. Topic modelling on users's comments. In 2014 third ICT international student project conference (ICT-ISPC) (pp. 177-180). IEEE.

53] Du, R., Kuang, D., Drake, B. and Park, H., 2017. DC-NMF: nonnegative matrix factorization based on divide-and-conquer for fast clustering and topic modelling. Journal of Global Optimization, 68, pp.777-798.

54] Newman, D., Bonilla, E.V. and Buntine, W., 2011. Improving topic coherence with regularized topic models. Advances in neural information processing systems, 24.

55] Qaiser, S. and Ali, R., 2018. Text mining: use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181(1), pp.25-29.

56] Laender, A.H., Ribeiro-Neto, B.A., Da Silva, A.S. and Teixeira, J.S., 2002. A brief survey of web data extraction tools. ACM Sigmod Record, 31(2), pp.84-93.

57]mention also, Friedl, J.E., 2006. Mastering regular expressions. " O'Reilly Media, Inc.".

58] C., Pietz, T. and Maalej, W., 2021, September. Unsupervised topic discovery in user comments. In 2021 IEEE 29th International Requirements Engineering Conference (RE) (pp. 150-161). IEEE.

59] Carreño, L.V.G. and Winbladh, K., 2013, May. Analysis of user comments: an approach for software requirements evolution. In 2013 35th international conference on software engineering (ICSE) (pp. 582-591). IEEE.

60] Boyd-Graber, J., Hu, Y. and Mimno, D., 2017. Applications of topic models. Foundations and Trends® in Information Retrieval, 11(2-3), pp.143-296.

61] Abdelmotaleb, H., Wojtys, M. and McNeile, C., 2023, April. GSDMM model evaluation techniques with application to British telecom data.