Abstract

This study delves into an in-depth examination of discussions related to cancer on Twitter, covering the years from 2009 to 2022. Its primary objective is to understand the emotions, behaviors, and trending topics related to cancer within the Twitter community. The research adopts topic modeling techniques to identify prevalent themes. Intriguingly, the results demonstrate the dynamic nature of cancer discourse on Twitter, which has evolved in response to new medical advancements, changing societal attitudes, heartfelt personal stories, and significant global events, notably the COVID-19 pandemic. The study's methodology is comprehensive, encompassing data collection, thorough data preparation, and the deployment of the LDA model for topic discovery. Emphasis is also placed on the nuances of model tuning to achieve optimal analytical outcomes. The findings underscore Twitter's pivotal role as a platform for disseminating health information, facilitating meaningful connections among patients, medical professionals, and the wider public. By presenting a holistic overview of cancer-related conversations on social media, this research offers invaluable insights that hold significance for healthcare professionals, policymakers, and researchers, setting a foundation for subsequent studies exploring the interplay between digital platforms and global health narratives.

Keywords: Social media analytics,Text mining, Topic modeling, LDA

Contents

List of Figures

List of Tables

List of Abbreviations

SMPCS	School of Mathematical, Physical and Computational Sciences
LDA	Latent Dirichlet allocation

Chapter 1

Introduction

Cancer, as the world's second largest cause of mortality, has had a profound impact on global health, leading to approximately 10 million deaths in 2020 alone (World Health Organization, 2022). Over the past two decades, the number of cancer diagnoses has nearly doubled, rising from an estimated 10 million in 2000 to 19.3 million in 2020, as stated by (Global Cancer Observatory, 2020). Among these, breast cancer, lung cancer, and prostate cancer stand out as the most prevalent types. Each of these cancers has its unique set of risk factors, treatments, and outcomes. Beyond the medical implications, cancer's impact is profound, affecting the social, psychological, and economic aspects of individuals' lives. For many, a cancer diagnosis brings emotional distress and financial challenges. On a broader scale, the increasing healthcare costs associated with cancer treatments strain national healthcare systems.

This widespread influence of cancer on individual and societal levels underscores the importance of platforms that facilitate open discussions about the disease. Enter Twitter. Launched in 2006, Twitter has grown into a global digital platform where individuals can interact, disseminate information, and voice their opinions in real-time. As of now, there are approximately 4.2 billion active social media users, which equates to about 53% of the global population (*ZDNET*, n.d.). With its concise format, limited to 280 characters, Twitter offers a unique blend of text, images, and links, making it a compelling platform for public discourse. Given its real-time nature and global reach, Twitter serves as a valuable resource for researchers aiming to study public sentiment, behaviors, and trends on a myriad of topics, including cancer.

In the realm of research, one technique stands out for its ability to dissect and understand the vast amounts of data generated on platforms like Twitter: topic modeling. This is an unsupervised technique which emerges to be a pivotal blend between machine learning and text mining designed to uncover abstract topics within extensive textual datasets. Rather than focusing on singular documents, this method emphasizes discerning broad themes that permeate the entire dataset. Topics are characterized by word distributions, and each document is perceived as a mix of these topics. This approach allows researchers to condense vast amounts of textual data, making it more digestible and highlighting patterns that might not be immediately obvious. The true potential of topic modeling is realized when dealing with large volumes of unstructured data, a common characteristic of social media content.

Topic Modeling when applied to health discourse, especially discussions surrounding cancer on platforms like Twitter becomes an instrumental tool. It allows researchers to sift through millions of tweets to identify prevalent themes, concerns, and sentiments. Whether it's about

understanding patients' experiences, identifying common misconceptions, or gauging public sentiment about new treatments or research breakthroughs, topic modeling offers a structured approach to make sense of the vast and varied conversations happening online.

In conclusion, the strength of topic modeling can be used to get essential core information from Twitter and such insights derived from real-world conversations, hold the potential to significantly influence public health strategies, awareness campaigns, and even medical research directions.

1.1 Problem statement

In today's digital age, the vast digital footprints left by users on platforms like Twitter offer an unparalleled opportunity to gauge public sentiment, knowledge, and concerns on various topics. One such critical topic, both in terms of its medical and socio-economic impact, is cancer. As one of the leading causes of death globally, conversations about cancer on Twitter encompass a spectrum of themes, ranging from personal experiences and emotional support to discussions about treatments, research breakthroughs, and myths.

However, given the magnitude of data generated on Twitter—millions of tweets daily—extracting meaningful and specific information about cancer becomes an challenge. Traditional methods are often limited in their scope and might either miss out on significant, tangential conversations that don't explicitly mention the predefined keywords or, conversely, include irrelevant data due to the broad or ambiguous nature of certain keywords. Additionally, the dynamic and real-time nature of Twitter conversations, combined with the short length of tweets, further compounds the challenge. The content often contains slang, abbreviations, and is influenced by current events, making it a difficult target to study with static analytical tools.

Moreover, while the sheer volume of this data presents a goldmine for researchers and policymakers, it also poses substantial challenges in terms of storage, preprocessing, and analysis. Even after successfully aggregating cancer-related tweets, understanding the multifaceted and layered conversations requires advanced techniques. Furthermore, the global and diverse user base of Twitter means that conversations about cancer are influenced by cultural, regional, and demographic factors. Understanding these variances is crucial, as they can offer insights into disparities in awareness, access to healthcare, or prevailing myths in different communities or regions. However, segmenting and analysing the data to unearth these insights is a challenging process.

In summary, while Twitter holds the potential to provide invaluable insights into the global discourse on cancer, several challenges stand in the way. There is a pressing need for a robust, comprehensive, and nuanced methodology that can navigate the complexities of Twitter data, accurately capture the multifaceted conversations about cancer, and translate them into actionable insights. This research aims to rise to this challenge and delve into the current state of the UK's discussion on cancer through Twitter.

1.2 Aims and objectives

The primary aim of this study is to conduct a comprehensive analysis of discussions related to cancer on Twitter, with a particular emphasis on the UK demographic. The research seeks to delve into the multifaceted nature of these conversations, capturing a broad range of topics. This includes understanding individual behaviors and prevalent public knowledge about the disease, as well as uncovering heartfelt narratives of those affected. Furthermore, the study intends to investigate the effects of existing support mechanisms and communities on the

platform, assessing their role in aiding individuals and families navigating through the challenges of cancer. Through this holistic approach, we aspire to gain a richer understanding of the digital discourse surrounding cancer within the UK's Twitter landscape.

Objectives:

- Gather a comprehensive collection of tweets from the UK related to cancer, capturing both medical information and individual stories.
- Utilize an extensive set of keywords, ensuring that the research captures the breadth of the discourse, including region-specific including local terms and phrases.
- Process and refine the collected data by eliminating irrelevant content and noise ensuring a focused dataset for analysis.
- Determine the most suitable parameters for the analytical model to guarantee precise and meaningful results.
- Employing topic modeling tools to identify main themes from the tweets. This includes topics on habits linked to cancer, people's emotions, and discussions about support like helplines.
- Rigorously assess the adopted methodologies to ensure their robustness and accuracy in capturing the essence of the discourse.

With these objectives in mind, the study seeks to provide a detailed understanding of how the UK discusses cancer on Twitter, shedding light on both factual information and personal experiences, as well as the support structures available.

1.3 Solution approach

To achieve the set aims and objectives of the project, a systematic solution approach is set for implementation. The initial phase involves identifying the most optimal sources for data acquisition. After these sources have been clearly determined, a comprehensive data collection process is initiated, consolidating information from each source. This ensures that the data not only remains consistent but also retains its integrity. Following the collection phase, the data undergoes a meticulous filtration process, where unrelated tweets are excluded. Subsequently, the data is subjected to an in-depth cleaning and transformation process, ensuring it's in the ideal state for modeling. With the data prepared, attention is then directed towards the modeling phase. Here, the model's hyperparameters are adjusted and optimized to ensure its peak performance. After the execution of the model, the results are first manually labeled to provide context to the findings. Following the labeling, the model's output undergoes a rigorous evaluation process. This evaluation employs both quantitative metrics and a handson qualitative assessment to ascertain the quality of the topics generated. The culmination of this process sees the creation of visual aids, designed to offer a clear visual representation of the patterns and trends evident in the UK's Twitter discussions about cancer.

1.4 Organization of the report

The report dives deep into a comprehensive Literature Review, exploring the evolving landscape of health trends on social media, emphasizing Twitter's role as a pivotal health discussion platform. The nuances of topic modeling, its applicability to Twitter data, and pertinent findings

from prior research are thoroughly examined, laying the groundwork for our investigation. Following this foundation, the Methodology segment thoroughly details our research trajectory. It encompasses the stages of data acquisition, integration, and refinement, leading up to the intricate procedures of topic modeling, from data preparation to model evaluation and topic labeling. With the methods firmly established, the Results section showcases our core findings, enriched with illustrative tables and figures, leading to a concise summary of the outcomes. The narrative then progresses to the Discussion and Analysis phase, where our findings are critically examined, contextualized against existing knowledge, and their broader significance is highlighted. Potential limitations and their implications are also discussed, offering a balanced view of our research. Drawing towards a close, the Conclusion and Future Work segment encapsulates the research's key insights, offering reflections on the potential next steps in this domain. Rounding off the report, the Reflection chapter provides an introspective look into the research journey, encapsulating the experiences, challenges, and learnings gleaned throughout the process.

Chapter 2

Literature Review

In today's interconnected world, platforms like Twitter have significantly revolutionized the manner in which health information is disseminated, consumed, and discussed. This influential microblogging site, renowned for its real-time updates, has rapidly emerged as a central hub for diverse health dialogues. These conversations span a broad spectrum, from general wellness tips and preventive measures to in-depth discussions about cutting-edge medical research and innovations. As Twitter continues to grow in popularity, an ever-increasing number of healthcare professionals, researchers, enthusiasts, and the general public are turning to it (*Social media use among healthcare professionals*, 2022). This collective engagement transforms the platform into a dynamic mirror, accurately reflecting global health sentiments, concerns, emerging trends, and public perceptions.

However, the vastness and rapid pace of Twitter also introduce a formidable challenge: the task of extracting relevant, meaningful, and actionable health information from the overwhelming sea of daily tweets. This is where advanced analytical techniques come into play. Topic modeling, a sophisticated statistical method, is tailored to distill and interpret such expansive datasets (Saxton, 2018). By meticulously categorizing extensive text into distinct, well-defined themes, topic modeling offers a structured and organized summary of the most prevalent health discussions on Twitter. This analytical approach not only aids in capturing and understanding the current health narratives but also adeptly pinpoints areas of misinformation, misconceptions, or topics that demand further exploration and attention.

As the scope of health discourse on Twitter continues to expand encompassing critical areas, the indispensability and relevance of analytical tools like topic modeling become even more pronounced. By harnessing the insights derived from such comprehensive analysis, stakeholders can shape impactful public health initiatives, inform and direct research endeavors, and guide evidence-based policy decisions. Ultimately, the goal is to ensure that the vast, collective voice on Twitter is not just heard, but also effectively interpreted and acted upon, benefiting society at large. Certainly! Let's review the provided content in the context of our conversation to ensure there are no repetitions and that the content is coherent.

2.1 Social Media Health Trends

The revolutionary power of social media in health discourse is undeniable. Digital platforms have emerged as essential mediums for the communication and distribution of health-related knowledge, providing a setting for dynamic interactions between healthcare providers and the public. These channels have proven vital in sharing timely and critical health information, as seen by the current epidemic. According to (Hagg et al., 2018), social media plays an important role in tackling global health issues such as COVID-19, making it a crucial tool for public health communication.

In recent years, the significance of social media in health discussions, particularly among young adults, has been highlighted. A study by Peng Wu and Ran Feng titled ”Social Media and Health: Emerging Trends and Future Directions for Research on Young Adults” underscores the intersection of social media and health. The research emphasizes the role of social media in health-related fields and constructs a theoretical model of factors affecting the continuous use intention of health-related social media among young adults (?). This study provides evidence of the growing reliance on social media platforms for health information, especially among the younger demographic, and the potential implications for health communication strategies.

Additionally, the importance of social media data for health research is becoming acknowledged by the scientific community. The tendency of mining social media information to find health narratives and generate valuable insights is emphasised (Korda and Itani, 2011). Such analyses can shed light on public health concerns, patterns, and concerns, offering an innovative perspective that traditional research may overlook.

As the digital landscape continues to evolve, the role of social media in shaping health narratives and influencing public perceptions cannot be understated. It offers a unique platform for real-time interactions, dissemination of health advisories, and gauging public sentiment, making it an invaluable tool for health professionals, researchers, and policymakers alike.

2.2 Twitter as a Health Discussion Platform

The advent of social media has produced a revolutionary environment for healthcare discussion, with Twitter emerging as a particularly significant platform. Twitter’s real-time stream of news, personal accounts, and other medical facts provides patients and healthcare practitioners with an interactive medium (Sugawara et al., 2012). Twitter is a crucial medium for conveying critical health information due to its immediacy and accessibility, particularly during global health emergencies such as the COVID-19 pandemic (Hagg et al., 2018).

Celebrity health disclosures on Twitter have been shown to significantly influence public health discourse. When high-profile individuals share their health diagnoses or experiences, it often sparks widespread conversation and awareness on the platform. Studies have shown that these celebrity disclosures can lead to a surge in Twitter conversations, emphasizing the platform’s potential to shape public health discourse and possibly influence health behaviors (Vos et al., 2019). Such events underscore the platform’s ability to amplify health-related messages and the broader implications of celebrity influence in health communication.

Moreover, the platform has been used to monitor mental health discussions, providing insights into public sentiment and concerns related to mental well-being (McClellan et al., 2017). Such

studies underscore the potential of Twitter as a tool for real-time health surveillance and sentiment analysis.

However, while Twitter provides an expansive platform for health discussions, it is crucial to approach its content with discernment. Ensuring the quality and evidence-based nature of the information disseminated is paramount, especially given Twitter's significant role in health discourse (Pershad et al., 2018).

The platform's potential for real-time health surveillance has been explored in various studies. For instance, researchers have utilized Twitter to monitor public sentiment and misinformation regarding vaccines, providing valuable insights into public health campaigns and strategies (Blankenship et al., 2015).

The recent global health crisis further underscored Twitter's significance in health communication. During the COVID-19 pandemic, Twitter played a pivotal role in disseminating information, tracking public sentiment, and combating misinformation. The platform became a primary source of updates, guidelines, and public health advisories, emphasizing its importance in global health communication (Chew and Eysenbach, 2020).

Therefore, Twitter's role in health discourse is multifaceted. It serves as a platform for information dissemination, sentiment analysis, and real-time health surveillance. However, the challenges posed by misinformation and the rapid spread of content necessitate a cautious and informed approach to its content, especially in the realm of health

2.3 Topic Modeling in Contemporary Research

Topic modeling has emerged as a computational technique pivotal for identifying latent thematic structures within vast text datasets. By analyzing the inherent patterns and relationships between words, topic modeling algorithms can uncover the underlying topics that pervade a collection of documents. This process is invaluable in the digital age, where the sheer volume of available text can be overwhelming, making manual analysis impractical (Blei, 2012).

Beyond mere identification, the true strength of topic modeling lies in its ability to structure and categorize vast amounts of unstructured data. By organizing content into clear topic clusters, it not only facilitates easier information retrieval but also provides insights into the dynamics and interplay of different themes within the data. Such organization aids researchers and analysts in understanding the nuances and trends present in large textual datasets (George and Birla, 2018).

The adaptability of topic modeling is evident in its diverse applications. In the realm of management research, its potential in uncovering hidden trends and patterns has been recognized, proving invaluable for informed decision-making (Stor´opoli et al., 2020). Furthermore, its role in comparative research is noteworthy. Topic modeling serves as a bridge in such studies, seamlessly integrating qualitative and quantitative analyses. This integration offers a robust framework for drawing comparisons across varied datasets, enriching the research landscape (Roberts et al., 2018).

An intriguing aspect of topic modeling is its intersection with other research methodologies.

The interplay between grounded theory and topic modeling, for instance, has been explored, revealing both convergences and divergences. Such explorations contribute to the methodological discourse in social research, offering fresh perspectives (Mohr and Bogdanov, 2013).

While the technique's dynamism is evident, especially when traditional methods are integrated with lexical resources like WordNet synset for candidate labels (Rahman et al., 2020), challenges persist. The evolving nature of topic modeling necessitates continuous refinement. Ensuring the method's reliability and consistency remains paramount, as underscored by ongoing research in the field (Mulunda et al., 2018).

2.4 Mathematical Principles of LDA

Latent Dirichlet Allocation (LDA), a groundbreaking probabilistic topic model, was introduced by Blei, Ng, and Jordan in 2003 (Blei et al., 2003). Their seminal work provided a comprehensive mathematical foundation for the LDA model, emphasizing its capability to discover latent topics in large text corpora. The model operates on the principle that documents are mixtures of topics, and topics are mixtures of words. The mathematical intricacies of the model, as detailed in their paper, revolve around the Dirichlet distribution, which serves as the prior for the topic proportions in documents and the word distributions in topics. The paper elucidates the generative process of LDA, where each document is modeled as a finite mixture over an underlying set of topic probabilities. The authors' rigorous approach to the model's formulation and their subsequent Bayesian inference methods have since become foundational in the realm of topic modeling.

Building upon the foundational work of Blei and colleagues, the paper on "Bayesian Parameter Estimation in LDA" delves deeper into the Bayesian aspects of LDA (Liu et al., 2015). The Bayesian framework provides a natural way to address the uncertainty in parameter estimation, which is inherent in any statistical model. In the context of LDA, Bayesian methods offer a principled approach to estimate the posterior distribution of the model parameters given the observed data. The paper underscores the importance of Gibbs sampling, a Markov Chain Monte Carlo (MCMC) technique, in estimating these parameters. By iteratively sampling from the posterior distribution, Gibbs sampling facilitates the estimation of topic-word and document-topic distributions. The Bayesian perspective, as articulated in this paper, not only enhances the robustness of the LDA model but also provides a deeper understanding of its underlying mathematical structure.

2.5 Topic Modeling on Twitter Data

With the fast expansion and diversity of data on Twitter, it is difficult to manually read through all of the tweets; thus, improved approaches for successfully analysing the material to identify key themes are required. Advanced approaches such as topic Modeling may be used by implementing various algorithms such as LDA, NMF, Top2Vec, BERTopic, etc. (Egger and Yu, 2022) performed a thorough comparison of the most common topic Modeling strategies such as LDA, NMF, Top2Vec, and BERTopic. Their extensive investigation shed light on the strengths and weaknesses of each approach. They have done crucial research that is essential for researchers investigating to select the best topic Modeling method for their datasets.

Adding to the diverse field of research, (Islam, 2019) explored the interrelationship between Yoga and Veganism on Twitter. Utilizing the LDA method for topic modeling, the researchers unearthed a significant overlap in discussions related to the two subjects. This intersection was further peppered with themes revolving around health, wellness, and sustainable living, suggesting that these health-oriented practices might appeal to similar demographics or share underlying philosophies.

(Montenegro et al., 2018) embarked on a localized approach, leveraging LDA to mine insights from Twitter datasets of Dumaguete City. Their efforts highlighted the utility of LDA in micro contexts, unearthing themes centred around the city's activities, sentiments, and events. Furthermore, (Hidayatullah et al., 2019) also utilized LDA topic Modeling to address meteorological issues. They discovered trends relating to weather, climate, and other natural occurrences by analysing tweets from the dataset published on BMKG's official account. This study demonstrated LDA's adaptability in a variety of sectors.

(Yang and Zhang, 2018) addressed the value of (LDA) in analysing Twitter datasets. They further extended it a step forward by integrating topic Modeling with sentiment analysis, providing a comprehensive perspective of the data's dominating topics and inherent sentiments. This dual strategy made it possible to understand the Twitter landscape in more detail. (Pratama et al., 2022), on the other hand, used the power of LDA to distil public impressions of the Telkom University during the chaotic days of the Covid-19 outbreak. Their study emphasised LDA's ability to sift through massive amounts of tweets to provide a comprehensive picture of public attitude towards the institution throughout the epidemic. Both studies, in essence, highlight the potential of LDA in understanding large-scale Twitter data, whilst with distinct focus areas and purposes.

The studies of (Ostrowski, 2015) and (Habbat et al., 2021) both delve into the world of mining Twitter data, emphasising the complexities of getting significant insights from such a vast and unstructured digital environment. Ostrowski's work largely promotes the benefits of LDA as a powerful tool for addressing the issues connected with comprehending the enormity of Twitter datasets. This support for LDA is seen in the study done by Habbat, Anoun, and Hassouni. In their investigation of tweets from Moroccan users, they not only used LDA but also compared its effectiveness to that of another approach, Non-negative Matrix Factorization (NMF). Their comparative approach reinforced the preeminence of LDA, particularly in achieving superior topic coherence.

(Negara et al., 2019) showcased LDA's adaptability in the Indonesian linguistic context, categorizing diverse topics seamlessly. This reaffirms LDA's versatility across varied linguistic datasets. Complementing this, (Srinivasan and K, 2021) demonstrated LDA's finesse in context-specific studies, analyzing Elon Musk's tweets to discern key themes.

Apart from using traditional LDA to get insights from twitter data of various backgrounds, researchers like (Sasaki et al., 2014) expanded upon the conventional Latent Dirichlet allocation (LDA) model by proposing the Twitter-TTM, which assesses the dynamics of user interests and topic trends online. Their innovative approach aimed to provide a more accurate representation of tweet generation and dissemination. Similarly, (Resnik et al., 2015) investigated the potential of topic Modeling in health and wellness by investigating the automated detection of depression-related language on Twitter. Their creative approach extended beyond classic LDA,

utilising more complex models such as SLDA and SANCHOR. Their findings highlighted the potential of these models for recognising and comprehending mental health-related themes.

2.6 Critique of Findings and Implications

Topic modeling on platforms such as Twitter has garnered significant attention in contemporary research. The collection of papers reviewed showcases the recurrent use of the Latent Dirichlet Allocation (LDA) model. Chosen for its capability to effectively manage the expansive data generated by Twitter, LDA emerges as a dominant theme across these studies.

In the context of the LDA model's efficacy, it's noteworthy that its strength lies in its potential to identify and categorize topics within the intricate web of Twitter data. This data often comes with its set of challenges, from varying linguistic patterns to the frequent use of emojis and abbreviations. Despite these challenges, LDA's performance has been largely commendable, as depicted in multiple studies.

However, a crucial observation arising from a deeper analysis is the inconsistent application methodologies across the research spectrum. Specifically, a glaring gap in many studies is the lack of comprehensive hyperparameter tuning for the LDA model. Hyperparameter tuning is not just a supplementary step but a fundamental process that ensures the optimization of the model. Its absence or lack of emphasis can critically influence the accuracy and reliability of the model's outcomes, potentially leading to misleading results.

Furthermore, another significant area of concern is the limited focus on quantitative evaluation. While several papers present their findings, there's an observable trend of under-emphasizing or entirely omitting the use of robust quantitative metrics such as coherence score or perplexity measure for evaluation. Relying solely on qualitative evaluations might result in an incomplete picture that lacks the empirical precision that metrics provide. For a field as dynamic and vast as topic modeling on Twitter, the integration of quantitative metrics is not just beneficial but essential. It offers a structured, objective lens to assess the performance, reliability, and generalizability of the findings.

Moreover, in the broader landscape of research ethics and practical implications, two primary considerations emerge. The first revolves around the ethical nuances of mining Twitter data, especially in an era where data privacy and user consent are paramount. The second pertains to the tangible, real-world applications of the research findings. While theoretical contributions are important, elaborating on their practical consequences in areas like as marketing, sentiment analysis, and public policy can considerably increase the research's relevance and usefulness. The following table gives a pinpoint summary of the literature review.

In conclusion, the ensemble of reviewed papers provides a comprehensive overview of the state of topic modeling on Twitter, with the LDA model at its core. While the insights and advancements are notable, the collective findings also underscore the pressing need for more refined methodologies, particularly in the realms of hyperparameter tuning and quantitative evaluation. As the field progresses, these considerations, coupled with ethical and practical deliberations, will be pivotal in shaping the future trajectory of research in this domain.

Paper Title	Key Findings	Limitations

Using Latent Dirichlet Allocation for Topic Modelling in Twitter	Effective utilization of LDA for topic extraction from Twitter data	Limited quantitative evaluation and dataset specifics
Online topic model for Twitter considering dynamics of user interests and topic trends	Considered user interests and dynamic topic trends for a more evolved topic modeling approach	Complexity in methodology; might not be transparent for all readers
Telkom University Opinion Topic Modeling on Twitter Using Latent Dirichlet Allocation During Covid-19 Pandemic	Highlighted the utility of LDA for sentiment analysis during a specific event (Covid-19)	Narrow scope due to focus on a single university's opinions
Data Aggregation Of Tweets And Topic Modelling Based On The Twitter Dataset	Emphasized the importance of data aggregation and preprocessing for effective topic modeling	Less emphasis on postmodeling analysis and interpretation
Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter	Explored models beyond LDA for specific language related to depression on Twitter	Reliance on labeled data; might not be scalable for larger datasets
A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts	Comprehensive comparison of multiple topic modeling techniques	Potential for overwhelming readers unfamiliar with all models; lack of decisive conclusions
Topic Modeling of Weather and Climate Condition	Applied LDA for extracting topics related to weather and climate conditions	Narrow theme, which might limit broader applicability
Using latent dirichlet allocation for topic modelling in twitter	Detailed exploration of LDA's strengths in handling Twitter data's intricacies	More generalized findings; could benefit from deeper real-world applications
Topic Modelling Twitter Data with Latent Dirichlet	Detailed categorization of topics using LDA, showcasing its breadth in application	Limited quantitative metrics and broader real-world implications

Table 2.1: Summary of the literature review

Chapter 3

Methodology

The methodology employed in this study is crafted to offer a broad yet insightful understanding of thematic structures within an expansive dataset that covers the years from 2009 to 2022. The journey begins with the Data Acquisition phase, where a diverse range of tweets are systematically gathered, laying the groundwork for the subsequent stages of the study. This foundational dataset, rich in content and scope, transitions into the Preliminary Analysis phase. Here, the initial layers of the data are peeled back, revealing prominent themes and patterns using a blend of analytical tools, without diving into granular specifics. Given the vastness of the dataset and the potential for inconsistencies, the Data Cleansing phase plays an instrumental role. This phase dedicates itself to refining and streamlining the data, ensuring it's primed for the deeper analytical processes that follow. The In-depth Thematic Analysis phase, while still maintaining a veil of overview-level discretion, delves a step further into the data, exploring the intricacies of the identified themes and unearthing deeper connections. The methodology's capstone is the Validation and Refinement phase. Emphasizing the study's commitment to accuracy and relevance, this phase integrates various validation processes, cross-referencing, and iterative refinements, ensuring the insights derived are both robust and anchored in authenticity. Collectively, these phases craft a methodological tapestry that guides the study in its mission to navigate and elucidate the thematic landscape of the dataset, all while preserving a sense of overarching clarity in its overview.

3.1 Data Acquisition and Integration

Data acquisition is a pivotal and foundational step in any data analysis process, setting the tone for all subsequent stages. The quality, depth, and relevance of the data procured play a cardinal role in shaping the eventual outcomes, influencing the robustness and credibility of the results. It's imperative to recognize that data doesn't function in isolation; the synthesis and integration of multiple datasets provide a multi-dimensional perspective, ensuring a holistic understanding that encapsulates the myriad facets of the topic under investigation.

For the purposes of this study, a selection of datasets was retrieved from Kaggle. These datasets, while not directly linked to cancer, were significant in that they contained tweets specifically from the UK and spanned the research timeline of 2009-2022. The specificity of these UK-centric tweets added a geographical nuance to the study, allowing for a more localized understanding of the subject matter. However, given the disparate origins and varied structures of these datasets, combining them presented a unique challenge. To transform these varied sets into a cohesive dataset, a diligent merging methodology was adopted. This procedure was not merely about collating data; it was a meticulous exercise in preserving data integrity. Utmost

care was exercised to ensure accuracy, reconcile any structural differences, and eliminate potential redundancies or overlaps. The culmination of this rigorous data acquisition and

integration process was a singular, unified dataset, enriched with UK-specific tweets and primed for in-depth analytical exploration.

3.1.1 Data Consistency and Redundancy Management

Data refinement is pivotal in guaranteeing the caliber of the dataset, particularly when the source is a dynamic platform like Twitter. The inherent nature of such platforms means raw data often comes with a mix of relevant and irrelevant information. Refinement serves as a critical filtering process, honing in on pertinent details while discarding unnecessary elements, thus ensuring a cleaner, more focused dataset devoid of inconsistencies or extra data.

In the context of this study, a judicious approach was adopted to distill the data to its essence. Out of the multitude of fields available in the raw data, only two were identified as paramount for the analysis: 'text', representing the actual content of the tweet, and 'time', indicating when the tweet was posted. This narrowed focus streamlined the dataset, making the analysis more targeted. An absence of values in these crucial fields was deemed unacceptable. As a result, any record lacking a 'text' or 'time' entry was promptly excised, ensuring that only complete data points were considered for further analysis.

A significant challenge arose from the potential of having overlapping data, especially since tweets from similar timeframes could be present across multiple datasets. A meticulous procedure was implemented to identify and remove such duplicate entries, reinforcing the dataset's integrity. Another key aspect of the refinement was the standardization of data formats. Timestamps, being crucial for chronological analyses, were uniformly formatted to adhere to the "YYYY-MM-DD HH:MM:SS" structure, ensuring consistency and ease of analysis.

Post these rigorous refinement steps, the dataset was whittled down to a curated collection of 37,983 tweets. This refined dataset not only encapsulated the study's focal points but also set a solid groundwork, primed and ready for the next stages of in-depth data processing and analysis.

3.2 Data Preprocessing

Data Preprocessing is a vital phase in the data analytics pipeline, designed to transform raw, unstructured data into a ready-to-analyze format. Particularly with textual data, like that from Twitter, the raw content often contains a mix of relevant information, noise, and possible inconsistencies. The challenge of preprocessing becomes even more pronounced when considering the spontaneous and diverse nature of tweets, with myriad topics and sentiments expressed in concise formats.

In the realm of machine learning, preprocessing holds paramount importance. It ensures that the models and algorithms, which rely heavily on the quality of input data, are fed with information that's both relevant and structured. Considering a platform like Twitter, where users discuss a vast array of subjects and respond to global trends in real-time, it becomes imperative to filter out data that might detract from the primary research focus.

The aim is to sift through the vast volumes of tweets, isolating and retaining only those that align with the research theme, while discarding potential noise. This might include unrelated trending topics, retweets without additional commentary, or tweets that lack context. Such rigorous preprocessing not only streamlines the dataset but also amplifies the accuracy and relevancy of

the subsequent analysis, ensuring that the insights derived are both meaningful and aligned with the study's objectives.

3.2.1 Keyword Filtering

The core aim of this study was to comprehensively explore the multifaceted dialogues surrounding cancer on Twitter, capturing not just the surface-level conversations but also the underlying nuances. Recognizing that discussions about cancer encompass more than just direct mentions, the study adopted a broad lens. Rather than merely looking for tweets with the term 'cancer', the research embraced a wide array of related terms. This included references to potential causes, specific symptoms, available treatments, preventive habits, prescribed medications, and even the emotional and social aspects tied to cancer, such as support mechanisms and personal experiences.

This expansive keyword list, detailed in the document in appendix, served as a vital tool in filtering the data. Each tweet in the original dataset underwent meticulous scrutiny, being cross-referenced against this list. The objective was clear: to ensure that only the most pertinent tweets, those resonating with the study's core focus, made the cut. Any tweet that didn't resonate with at least one keyword was promptly excluded from the dataset.

Such rigorous filtration yielded a refined collection of 12,354 tweets, distilled from the initial pool of 37,983. This methodical approach underscored the study's commitment to precision, ensuring that every retained tweet had intrinsic relevance to the overarching theme of cancer discussions on Twitter. The resultant dataset, thus, provided a robust base for the subsequent phases of analysis, primed to offer insights deeply aligned with the research objective.

3.2.2 Text Refinement

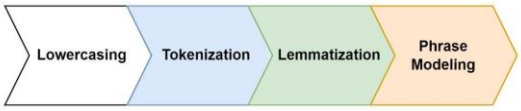

Figure 3.1: Text refining process

With the thematically aligned dataset in place, the next phase focused on refining it to extract meaningful insights. This dataset, predominantly comprised of tweets, represents a rich reservoir of information, emotions, and specific details. The primary challenge was to clean and organize this data, ensuring it was well-structured for in-depth analysis. Refining text stands as a pivotal task in the realm of natural language processing. This vital procedure transforms raw and unstructured data, gradually unveiling hidden patterns. By adjusting the text consistently and methodically segmenting it into simpler components, the stage is set for clearer insights, facilitating deeper examinations and discoveries in subsequent analytical phases. As this structured dataset takes shape, it lays a strong foundation for the subsequent analytical stages. With the groundwork set, the research is poised to delve deeper into further analysis, leveraging advanced techniques and tools to unearth even more profound insights.

Lowercasing: One of the foundational steps in textual data processing is ensuring uniformity in case usage. By converting all the textual data into lowercase, the study ensured a consistent representation across the dataset. This simple yet impactful step meant that words like

”Cancer”, ”cancer”, and ”CANcer” would be universally recognized and treated as ”cancer”. Such uniformity is more than just cosmetic; it plays a pivotal role in reducing data fragmentation. A significant advantage of this process is the reduction in the dataset’s dimensionality. By ensuring that different case variations of the same word are treated uniformly, the number of unique tokens in the dataset is considerably reduced. This simplification facilitates smoother subsequent analyses and minimizes computational strain.

Tokenization: To truly grasp the essence of the tweets, it was essential to deconstruct them to their fundamental components. This is where tokenization comes into play. Tokenization is akin to segmenting a continuous stream of text into its constituent parts, usually words. By employing delimiters such as spaces or punctuation, the study utilized the ‘word tokenize‘ function from the ‘nltk‘ library. This method transformed cohesive strings of text in tweets into lists of individual words or ’tokens’, paving the way for granular, word-level analysis.

Lemmatization: The intricacies and variations inherent in language can sometimes pose analytical challenges. A single concept or action in language can often be represented by several word variants. For instance, ’running’, ’ran’, and ’runs’ all allude to the fundamental action of ’run’. Lemmatization, a process that reduces words to their base or root form, was employed to tackle this challenge. Using the ‘WordNetLemmatizer‘ from the ‘nltk‘ library, word variants were harmonized to their canonical forms. This method not only trimmed potential redundancies but also ensured that the dataset maintained the depth and richness of its semantic content.

Phrase Modeling: Words, while powerful on their own, often derive additional meaning from their associations. Certain word pairs or combinations can evoke specific contexts or sentiments. Recognizing this, the study ventured into phrase modeling. By harnessing the ‘Phrases‘ class from the ‘gensim‘ library, models were constructed to identify and analyze bigrams (two-word combinations) and trigrams (three-word combinations). With preset criteria for thresholds and frequency, this process identified 75 distinct bigram phrases and 15 trigram combinations. This meticulous approach ensured that the multi-word expressions, which often carry nuanced meanings, were not overlooked, thereby enriching the overall analytical context.

3.3 Topic Modeling Preparation

Before moving to the advanced stages of topic modeling, it’s essential to first ensure a solid base for the data. This early stage is crucial because it confirms that the data is clean and well-organized, which is necessary for effective modeling later on. To achieve this quality, the study highlighted two important steps. The first step was the creation of a data dictionary. This dictionary provides a clear and detailed description of the dataset’s features, helping researchers understand its contents better. The next step was the formation of a structured corpus. This is a way of organizing the text data so that it’s easier to analyze. By taking these two steps seriously, the study ensures that the data is in the best possible shape for the topic modeling process. This careful preparation means that the results from the modeling will be more reliable and useful.

3.3.1 Creating Data Dictionary

The data dictionary is a foundational element in many natural language processing tasks. At its core, it serves as a bridge between human language and computational algorithms. By assigning

unique IDs to words, the dictionary abstracts the textual data into a format that machines can efficiently process. This abstraction is not just a matter of convenience; it's a necessity. In large datasets, words can appear millions of times. Processing them as strings would be computationally expensive and slow. By converting words into unique IDs, algorithms can operate faster and more efficiently. Moreover, this systematic mapping ensures that the semantic uniqueness of each term is preserved, allowing for more accurate topic modeling outcomes.

Using the gensim library, the data dictionary was developed to provide a systematic mapping of word IDs to their respective words from the dataset. Assigning unique IDs to words ensures a standardized reference for each term, facilitating efficient computational processes in subsequent stages of analysis.

3.3.2 Creating Corpus

Simultaneously, the study turned its attention to the formation of a corpus. Corpus is a mathematical representation of a text dataset, emphasizing the frequency of word occurrences. This frequency-based representation is pivotal for several reasons. Firstly, it reduces the dimensionality of the data, making computational processes feasible. Secondly, by focusing on word frequencies, the corpus captures the underlying thematic structures in the dataset. Words that frequently appear together are likely to be contextually or thematically related. The transformation of raw text data into this frequency-based format is a critical step in ensuring the success of the topic modeling endeavor. By harnessing the capabilities of the gensim library, the dataset was transformed to reflect the frequency of each word.

3.4 Theoretical Foundations of the Model

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that categorizes underlying topics in text corpora. At its core, LDA assumes that documents are probabilistic mixtures of topics, and these topics themselves are probabilistic mixtures of words. The model employs a three-level hierarchical Bayesian structure. Each document in a collection is modeled as a finite mixture over an underlying set of topic probabilities. The Dirichlet process determines the mixture of topics for each document and the mixture of words for each topic.

From a mathematical perspective, for a document d, the topic distribution θ_d is drawn from a Dirichlet distribution:

$$\theta_d \sim \text{Dir}(\alpha)$$

For each word w in document d, a topic $z_{d,w}$ is drawn from a multinomial distribution defined by θ_d, and the word itself is drawn from a multinomial distribution defined by topic $z_{d,w}$ and word distribution $\varphi_{zd,w}$:

$$z_{d,w} \sim \text{Multinomial}(\theta_d)\ w$$

$$\sim \text{Multinomial}(\varphi_{zd,w})\ \text{Two}$$

pivotal hyperparameters

in LDA are α and β. The parameter α is the Dirichlet prior on the per-document topic distributions, and β is the prior on the per-topic word distribution. These hyperparameters influence the granularity of topics discovered by the model. A higher value of α results in documents being composed of more topics, and a higher β value makes topics to be composed of a larger set of words.

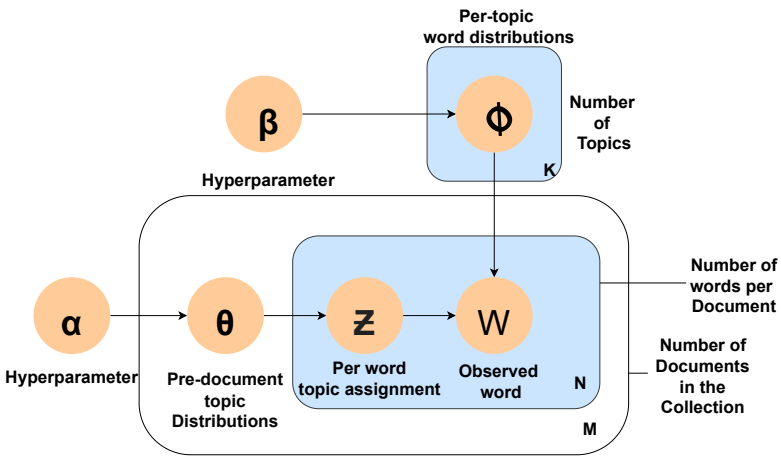

Figure 3.2: LDA GRAPHICAL MODEL, adapted from Montenegro et al. (2018).

LDA's capability to uncover hidden thematic structures in a text corpus is unparalleled, making it a standout choice among various topic modeling algorithms. Its unsupervised nature, which doesn't necessitate prior annotations or labeling, is especially advantageous for extensive datasets, ensuring efficiency and adaptability. The probabilistic foundation of LDA adeptly captures the inherent uncertainty, nuances, and subtleties, especially in datasets like Twitter

that often contain ambiguities. The model's proven efficacy and successful applications in similar research scenarios further reinforce its selection for this study.

In the context of this research, which focuses on analyzing Twitter data, the LDA model's parameters are crucial in determining the granularity and specificity of the topics extracted from the tweets. Here's a breakdown of the parameters and their roles:

1. Documents (D): Each tweet is treated as a document. Given the concise nature of tweets, the model is expected to identify more specific and nuanced topics compared to longer texts.

2. Vocabulary (V): The set of unique words across all tweets. This would include common terms used in Twitter language and other Twitter-specific lexicons.

3. Topics (K): The number of topics to be extracted from the dataset. This is a hyperparameter that needs to be set before running the model. The optimal number can be determined using methods like perplexity scoring or coherence measures.

4. Alpha (α): As previously mentioned, α is the Dirichlet prior on the per-document topic distributions. In the context of Twitter data, a slightly higher α might be preferred given the diverse range of topics a user might tweet about.

5. Beta (β): The Dirichlet prior on the per-topic word distribution. Given the brevity of tweets, a lower β might be suitable to ensure topics are not too broad.

Mathematically, the LDA model operates by iterating over each word in the tweets and assigning it to a topic based on the current topic assignments of other words and the words' co-occurrence patterns. The probability of word w being assigned to topic k in tweet d is given by:

$$p(z_{d,w} = k) \propto \frac{n_{d,k} + \alpha}{n_d + K\alpha} \times \frac{n_{k,w} + \beta}{n_k + V\beta}$$

Where:

- $n_{d,k}$ is the number of words in tweet d assigned to topic k.
- n_d is the total number of words in tweet d.
- $n_{k,w}$ is the number of times word w is assigned to topic k across all tweets.
- n_k is the total number of words assigned to topic k across all tweets.

This probability is computed for each topic, and the word is then reassigned to the topic with the highest probability. The model undergoes this iterative process of topic assignment multiple times until it converges to a stable set of topic assignments. Through this methodology, documents (in this case, tweets) are effectively modeled as mixtures of topics, and these topics are in turn modeled as mixtures of words. This ensures that the underlying thematic structures in the Twitter dataset are accurately and comprehensively captured.

3.5 Model Optimization

In the process of optimizing the LDA model, setting the right hyperparameters is vital. The selection isn't merely about achieving higher numerical values but ensuring that the model genuinely captures the underlying patterns of the data. The 'Hyperopt' library was chosen for this task, renowned for its efficacy in determining optimal hyperparameters. The principal aim was to secure a high coherence score, a metric that gauges the clarity and quality of topics produced by the LDA model. This score was computed using a tool from the 'gensim' library.

Several hyperparameters play pivotal roles in the model's performance. 'num topics' dictates how many distinct topics the model should identify in the dataset. The parameters 'alpha' and 'eta' influence how topics are distributed across tweets and how words are distributed across topics, respectively. The 'passes' parameter determines how many times the model reviews the entire dataset. To navigate through the potential combinations of these hyperparameters, the Tree-structured Parzen Estimator (TPE) method was employed. This method intelligently selects the next set of hyperparameters based on the performance of previous sets.

The provided graph offers a visual representation of the coherence scores across different hyperparameter combinations. A higher point on the graph signifies a more coherent and interpretable topic structure. The highest peak in the graph represents the most optimal set of hyperparameters, which, in this case, turned out to be 'num topics' of 14, 'alpha' of 0.5, 'eta' as 'symmetric', and 'passes' of 5. This configuration ensured that the model unveiled 14 well-defined topics from the dataset, as suggested by the peak coherence score.

3.6 LDA Model Implementation and Analysis

Latent Dirichlet Allocation (LDA) is distinguished by its probabilistic nature. Unlike deterministic methods, LDA ensures that a tweet, even when aligned with multiple topics, is primarily linked with the one that represents it best. This attribute becomes particularly significant when delving into the multifaceted content of platforms like Twitter. Here, users often blend topics, sentiments, and discussions, making traditional keyword-based analyses potentially inadequate.

Several hyperparameters play pivotal roles in the model's performance. num topics dictates how many distinct topics the model should identify in the dataset. The parameters alpha and eta influence how topics are distributed across tweets and how words are distributed across topics, respectively. The passes parameter determines how many times the model reviews the entire dataset. To navigate through the potential combinations of these hyperparameters, the Tree-structured Parzen Estimator (TPE) method was employed. This method intelligently selects the next set of hyperparameters based on the performance of previous sets.

For this study, the LDA model was meticulously tailored to resonate with the unique characteristics of the dataset. Training parameters were judiciously chosen: a chunk size of 1,000 ensured that each training batch had a rich variety of data; 5 training passes over the corpus guaranteed thorough learning and model convergence. By allowing the model to run through 2000 iterations, its capacity to discern subtle patterns and nuances in the data was significantly bolstered. The eta was set to 'symmetric' and alpha was set to '0.5'. Additionally, the 'eval every' parameter, set to 1, enabled the model's performance to be periodically assessed during its training phase. Such periodic evaluations are pivotal in detecting and rectifying any potential divergences early in the training process.

As a result of this diligent training process, the LDA model adeptly assigned each tweet to one of eight discernible topics. These weren't mere clusters of words but coherent thematic groupings that provided a more structured and meaningful perspective on the Twitter dataset. Each topic was emblematic of a certain narrative or sentiment, characterized by a specific set of frequently co-occurring words.

To visualise the comprehensibility and accessibility of the results, the study employed pyLDAvis. This tool transformed the abstract topic distributions into a tangible and interactive visual format. Through pyLDAvis, one could discern the relative importance of each topic, explore the pivotal terms defining them, and gauge the inter-topic distances. Such a visualization not only demystifies the topic modeling results but also offers stakeholders a user-friendly way to glean insights, facilitating informed decision-making and further research directions.

3.7 Assigning Topic Labels

In the course of utilizing the LDA model, a variety of distinct topics were identified, each stemming from the words present in the dataset. A pivotal aspect of this research phase was to adeptly label these topics, ensuring that each label resonated with the essence of its corresponding topic. To achieve this, a detailed examination was conducted on the words that consistently appeared within each topic. These words, by virtue of their recurrence, acted as crucial pointers, shedding light on the overarching theme or central message of the topic.

To initiate this labeling process, the study began by pinpointing the most predominant words that characterized each topic. These words, by their sheer prominence, provided invaluable insights, serving as a window into the core ideas and themes that each topic encapsulated. However, the task extended beyond just listing and identifying these words. An integral part of the process was deciphering the connections between these words, understanding their mutual relationships, and discerning the collective story or narrative they wove when considered together.

For example, in a topic where terms such as ”doctor”, “early”, “detection”, and ”consult” stood out, the interconnected significance of these terms was brought to the fore. When viewed collectively, these terms painted a vivid picture emphasizing the crucial role and importance of early medical intervention and timely consultation. In light of this understanding, the topic was aptly labeled as ”Healthcare Consultation”.

This meticulous and methodical approach ensured that each topic was paired with a label that was both descriptive and truly reflective of its content. This careful labeling process played a pivotal role in making the insights derived from the LDA model lucid, precise, and easy for readers to grasp.

3.8 Model Evaluation

The vast, interconnected landscape of cancer-related discussions on Twitter presented a complex challenge. To navigate and make sense of this intricate web, the research leveraged the capabilities of the LDA model, renowned for its precision and depth of analysis. However, the mere selection of a model wasn't enough. It was paramount to ensure the model's adaptability and reliability for this specific task. Hence, an exhaustive evaluation procedure was put into

motion. This procedure was holistic, not just relying on numerical metrics but also incorporating observational and interpretative insights. This multifaceted approach was vital in ensuring that the model was not just theoretically sound but also practically effective. In this section, the specifics of this evaluation are detailed. It delves into the rationale behind each assessment measure, explicates the foundational theories that anchored these choices, walks through the hands-on application of these measures, and finally, sheds light on the broader significance and ramifications of the results obtained.

3.8.1 Perplexity

Perplexity, firmly anchored in the domain of information theory, plays a crucial role when assessing models operating within probabilistic frameworks. Essentially, it offers a lens to gauge the model's degree of uncertainty when tasked with predicting subsequent words in a given corpus. Breaking it down, perplexity essentially evaluates how taken aback or surprised the model is upon encountering unfamiliar data. A more intuitive interpretation would be to view it as a measure of the model's predictive accuracy: a lower perplexity score implies the model is better equipped and less startled when making predictions on unseen samples.

The evaluation of the model revealed a perplexity score of -4.0599922248359395. This relatively low score is indicative of the model's proficiency in understanding and reflecting the patterns and word distributions present in the dataset. Such a score conveys that when the model encounters fresh data, especially if this data resonates with the patterns inherent to the original dataset, it tends to offer predictions with high accuracy. Beyond its mathematical significance, this score underscores the practical strength of the model, highlighting its potential to perform reliably in real-world scenarios and its ability to adapt to similar data distributions effectively.

3.8.2 Jaccard Similarity

One of the hallmarks of a reliable model is its ability to maintain consistency across different datasets. Consistency ensures that the model's findings are not just mere artifacts of a particular dataset but are representative of the underlying patterns within the data. To assess this crucial aspect of the LDA model, the Jaccard similarity metric was chosen. Grounded in set theory principles, the Jaccard similarity offers a method to quantify the similarity between two sets. It does this by dividing the size of the intersection of the sets by the size of their union, yielding a percentage score that reflects the degree of overlap.

To operationalize this measure in the study's context, the dataset was bifurcated randomly, creating two distinct subsets. The LDA model was then trained on each subset independently. After training, the dominant words characterizing the topics from both training runs were compared. This comparison yielded an average Jaccard similarity score of 0.17841289448616593. On its own, this figure might not appear particularly impressive. However, when placed within the intricate realm of topic modeling, it gains significance. In topic modeling, even minor shifts in data can lead to changes in word distributions. Thus, achieving such a score suggests that the model is both stable and reliable, adeptly capturing consistent topic structures irrespective of the specific data subset it's trained on.

3.8.3 Manual Evaluation

Beyond the confines of strict numerical evaluations, there's a vast domain where human insight and expertise become invaluable. While quantitative metrics provide a foundational understanding of a model's performance, they don't always capture the nuances and context inherent to human language and understanding. Recognizing this gap, and the paramount importance of ensuring that the model's outputs resonate with the human understanding of the subject, a comprehensive manual evaluation was instituted.

Each topic generated by the LDA model underwent rigorous scrutiny. Thee results were closely inspected including the dominant words within these topics, ensuring they formed a cohesive and logical narrative. The focus wasn't just on the internal coherence of the topics but also on their alignment with the wider discourse on cancer. Moreover, a significant emphasis was placed on interpretability. By interpretability, the study refers to the intuitive clarity with which the topics and the words associated with them mirror the real-world discussions and concerns about cancer. It's about ensuring that anyone familiar with the domain can look at the topics and instantly recognize their relevance and meaning.

This dual-pronged evaluation, combining mathematical rigor with human intuition, ensured that the insights extracted from the model were grounded in both analytical precision and real-world relevance. It fortified the belief that the results are not just numbers on a page but are tangible insights that can inform and guide real-world discussions and actions concerning cancer.

Chapter 4

Results

The vast scope of cancer-related discussions on Twitter from 2009 to 2022 paints a vivid picture of the evolving dialogue shaped by medical breakthroughs, society's views, and personal stories. As more people flocked to social media platforms, these digital spaces turned into hubs for sharing feelings, latest research updates, and talks about various cancer-related topics. The data we gathered from these online chats shows us the changing nature of how people talk and think about cancer over these years.

In this chapter, we dive into the results from our deep analysis of these Twitter conversations about cancer. Using powerful data analysis tools, especially the Latent Dirichlet Allocation (LDA), we managed to sift through huge amounts of raw data to identify clear topics. These topics give us a better structure to understand the main points of the cancer-related discussions on Twitter and highlight the main feelings, worries, and stories that have been most talked about during this period.

4.1 Topics Identified

One of the primary outcomes of the LDA analysis was the identification of distinct topics that emerged from the dataset. Each topic serves as a thematic cluster, representing a specific narrative or sentiment associated with cancer. Below, we explore each of these topics in detail:

Topic	Top N Words	Rationale for the Label

Emotions & Events	'emotions', 'events', 'feelings', 'occurrences', 'responses'	This topic seems to discuss the emotional reactions or sentiments people express in response to certain events or occurrences.
Research & Articles	'research', 'articles', 'studies', 'publications', 'findings'	The prominent words point towards scholarly works and the dissemination of academic or investigative findings.
Prostate & Colon Cancer	'prostate', 'colon', 'cancer', 'treatment', 'diagnosis'	The focus of this topic is clearly on two specific types of cancers.
Ovarian Cancer Awareness	'ovarian', 'cancer', 'awareness', 'prevention', 'campaign'	This topic distinctly revolves around raising awareness about ovarian cancer.
Medical Check-ups & Awareness	'medical', 'check-ups', 'awareness', 'health', 'screening'	The emphasis here is on the importance of regular health screenings.
Grieving & Health Importance	'grieving', 'health', 'importance', 'loss', 'support'	This topic delves into the emotional and psychological aspects of dealing with loss.
Lung Cancer Focus	'lung', 'cancer', 'focus', 'research', 'treatment'	The primary emphasis in this topic is on lung cancer.
Family Health Concerns	'family', 'health', 'concerns', 'well-being', 'care'	The topic encapsulates discussions and concerns related to the health of family members.
Emotions & Medical Outcomes	'emotions', 'medical', 'outcomes', 'feelings', 'recovery'	This topic highlights the intricate relationship between emotional states and their potential ramifications on medical results.
Healthy Habits	'healthy', 'habits', 'lifestyle', 'nutrition', 'exercise'	The topic emphasizes the significance of adopting a health-centric lifestyle.
Breast Cancer Awareness	'breast', 'cancer', 'awareness', 'campaign', 'prevention'	This topic is centered around spreading awareness about breast cancer.
Social Support & Health	'social', 'support', 'health', 'community', 'well-being'	The topic underscores the indispensable role of social support systems in promoting good health.
Healthcare Consultation	'healthcare', 'consultation', 'advice','professional', 'guidance'	This topic leans towards the professional aspect of healthcare.

Table 4.1: Summary of Topics, Words, and Rationales

4.2 Topic Distribution

The complexities embedded in cancer-related dialogues on Twitter are illuminated not just by the nature of the topics themselves, but also by the frequency of their occurrence. Observing the distribution of these topics offers valuable insights, shedding light on which subjects deeply

resonate with the online community and capture their attention. Moreover, this distribution acts as a barometer, indicating which areas have achieved significant traction and which might still be underrepresented, suggesting areas where further emphasis or awareness campaigns could be beneficial. This nuanced understanding can guide stakeholders in tailoring their communication strategies, ensuring that pivotal discussions gain the prominence they deserve.

4.2.1 Distribution of Topics in Tweets

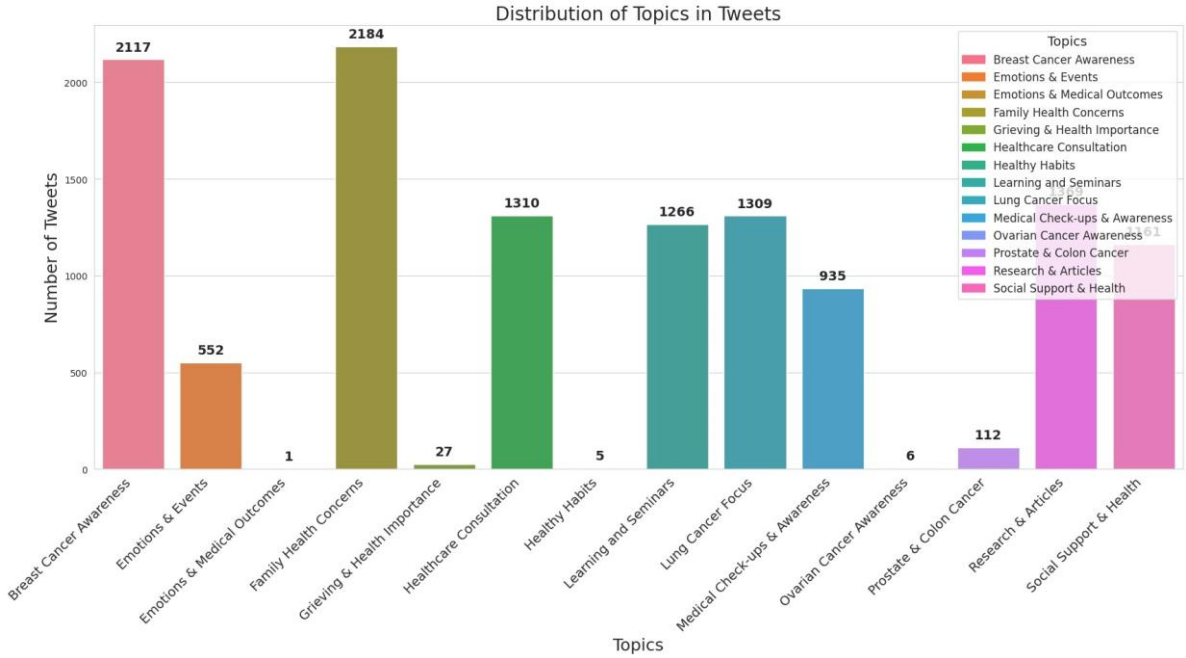

Figure 4.1: Distribution of Topics in Tweets
(Graph generated by python code)

Cancer-related conversations on Twitter from 2009 to 2022 offer a rich tapestry of insights, emotions, and shared knowledge. The bar graph illustrating the distribution of topics in tweets paints a vivid picture of what resonates most within the online community.

At the forefront, we observe the topic ”Family Health Concerns” garnering the highest number of tweets, standing at 2184. This significant number underscores the collective concern of families when it comes to cancer. It’s not just the individuals who are affected or diagnosed, but the entire family unit feels the impact. The vast discussions on this topic suggest that cancer is often viewed through a familial lens, emphasizing collective well-being, shared experiences, and mutual support.

Close on its heels is ”Breast Cancer Awareness” with 2117 tweets. The widespread discussions around breast cancer could be attributed to the numerous awareness campaigns, fundraisers, and research breakthroughs over the years. The global push for early detection, coupled with celebrity endorsements and real-life narratives, has likely propelled this topic to its prominent position.

Equally notable is the prominence of ”Healthcare Consultation” and ”Lung Cancer Focus,” each with 1310 and 1309 tweets respectively. These figures underline the significance of professional healthcare guidance. People are actively seeking expert advice, consultations, and clarity in their cancer journeys. The parallel emphasis on lung cancer, a type often associated with high

mortality rates, accentuates the urgent need for research, advanced treatments, and patient stories.

”Research & Articles” with 1369 tweets and ”Learning and Seminars” with 1266 tweets emphasize the community’s thirst for knowledge. The continuous evolution of cancer research means new findings, treatments, and methodologies emerge regularly. The discussions around these topics indicate an informed and engaged community, keen on staying updated and leveraging knowledge for better health outcomes.

A related but distinct topic, ”Medical Check-ups & Awareness,” has 935 tweets. Regular health check-ups, screenings, and a broader theme of health consciousness seem to resonate with many. The emphasis suggests that early detection and proactive health management are top priorities for many Twitter users. ”Social Support & Health” with 1161 tweets brings to light the importance of community in the cancer journey. The discussions revolve around the invaluable role of support groups, shared experiences, and the mental well-being derived from communal interactions.

However, not all topics have vast numbers. ”Prostate & Colon Cancer” sees 112 tweets, while ”Grieving & Health Importance” has 27. These numbers, though comparatively lower, emphasize niche areas of discussion. Grieving, in particular, touches upon the emotional aftermath of cancer, be it the loss of a loved one or coping with a diagnosis. Surprisingly, topics like ”Ovarian Cancer Awareness,” ”Healthy Habits,” and ”Emotions & Medical Outcomes” have very few tweets, 6, 5, and 1 respectively. These numbers could indicate potential gaps in awareness or perhaps reflect areas that have not been as widely publicized or discussed in the broader cancer narrative on Twitter.

In summary, the distribution of topics offers a panoramic view of cancer-related discussions on Twitter. From broad themes of family health and breast cancer awareness to niche areas like grieving and specific cancer types, the graph encapsulates the multifaceted nature of the discourse. The varied numbers across topics suggest areas of focus, potential interventions, and regions where awareness campaigns might be beneficial. As the digital age progresses, such insights become instrumental in shaping health communication strategies, research directions, and public health interventions.

4.2.2 Topic distribution over time

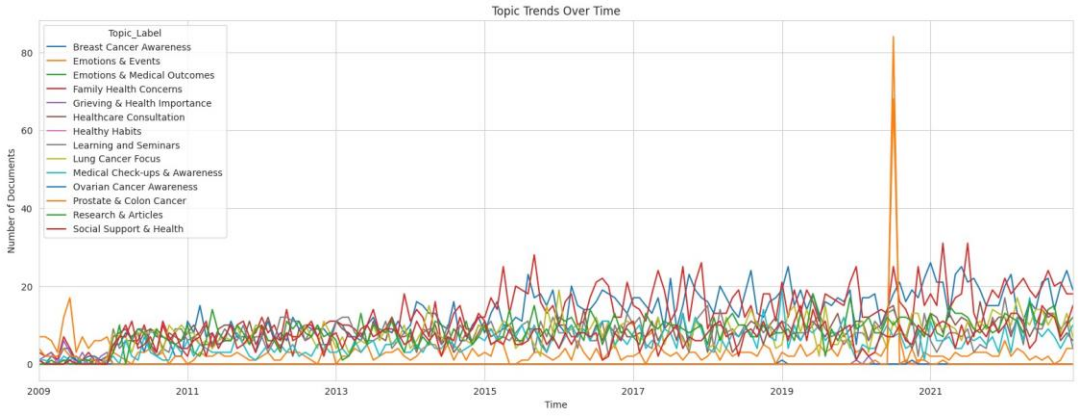

Figure 4.2: Distribution of Topics in Tweets

(Graph generated by python code)

The graph showcasing the distribution of cancer-related topics on Twitter over the span from 2009 to 2022 provides a comprehensive overview of the shifting dynamics of public engagement concerning this crucial health matter. Each topic's trajectory tells its unique story, revealing the collective consciousness of the masses in their interactions with and understanding of cancer.

One of the most striking features of the graph is the consistent and gradual increase in most topics. This steady rise indicates an escalating awareness and engagement with the multifaceted challenges and concerns surrounding cancer. As we delve deeper into individual topics, nuanced patterns emerge.

The topic of "Breast Cancer Awareness" has been gaining consistent momentum. The increasing trajectory might be indicative of the potential rise in breast cancer cases globally. This upward trend also reflects the persistent efforts by organizations, activists, and survivors in raising awareness, advocating for early detection, and mobilizing support for research.

The year 2020 stands out prominently with a pronounced spike in the "Emotions & Events" topic. As the world grappled with the unforeseen challenges of the COVID-19 pandemic, emotional responses and personal narratives became more pronounced. The pandemic's intersection with cancer discussions might have intensified the emotional outpourings, especially given the vulnerabilities and health concerns during that period.

"Family Health Concerns" presents another intriguing trend. After 2013, there's a noticeable uptick in discussions centered on the health implications for families dealing with cancer. This suggests a broader societal shift towards understanding health as a collective concern, emphasizing the interconnectedness of family members' well-being.

Moreover, the consistent rise in the "Medical Check-ups & Awareness" topic underscores a growing societal emphasis on preventative health measures. The increasing curve suggests that the public is becoming more proactive, understanding the significance of early detection and regular medical consultations.

Interestingly, while all topics have their unique trajectories, the overarching trend is undeniable: discussions about cancer on Twitter have been amplifying year after year. This overarching growth in cancer-related discourse underscores the platform's importance as a primary venue for health education, mutual support, and advocacy.

In conclusion, the graph serves as a testament to the evolving and expanding nature of cancer-related discussions on digital platforms. The increasing volume of discussions around breast cancer, family health implications, and the importance of medical check-ups signal the pressing health concerns of our times. With cancer discourse on an upward trajectory, continued analysis in this domain is not just beneficial but essential, especially as we anticipate more robust engagement in the coming years.

Chapter 5

Discussion and Analysis

The study's results, encapsulated in intricate graphs and detailed topic descriptions, offer a profound glimpse into the collective psyche of the masses as they engage with cancer-related discussions on Twitter. This chapter delves deeper into the implications of these findings, drawing connections with larger societal trends, medical advancements, and the role of digital platforms in shaping public discourse.

5.1 Emphasis on Emotional Well-being

The rise in the "Emotions & Events" topic, especially the pronounced spike in 2020, underscores the profound emotional toll that health crises, including cancer, can have on individuals and communities. In an era defined by the rapid dissemination of information, Twitter has emerged as a platform where people share their vulnerabilities, seek support, and offer solace. The intertwining of emotions with events, as seen in 2020 due to the COVID-19 pandemic, highlights the need for healthcare professionals and policymakers to consider emotional wellbeing as a crucial component of overall health.

Supporting this observation, a study by (Nemes and Kiss, 2021) titled "Information Extraction and Named Entity Recognition Supported Social Media Sentiment Analysis during the COVID19 Pandemic" delves into the sentiments of people on Twitter during the pandemic. The sentiment analysis, based on the BERT model, provides a deeper understanding of people's emotional state during this period. The data from their analyses further support the emotional categories and offer a comprehensive understanding that can be a starting point for other disciplines such as linguistics or psychology. This study reinforces the idea that platforms like Twitter serve as a mirror to society's emotional pulse, especially during health crises.

5.2 Growing Awareness and Advocacy

The consistent rise in topics such as "Breast Cancer Awareness" and "Medical Check-ups & Awareness" on Twitter suggests that advocacy efforts, both from institutional bodies and grassroots movements, are resonating with the public. The prominence of breast cancer discussions, potentially linked to an uptick in cases, underscores the urgency and significance of these dialogues. As awareness expands, the prospects for early detection, enhanced treatment options, and improved patient outcomes also increase.

Grassroots campaigns like Breast Cancer Now's "Wear it pink" initiative have been pivotal in

amplifying awareness (Breast Cancer Now, n.d.). This campaign, which galvanizes individuals to don pink and fundraise, has fostered a sense of community among those impacted by breast

cancer. Concurrently, institutional endeavors, such as Cancer Research UK's "Cancer Awareness in the Workplace" program, emphasize the importance of structured efforts in promoting early detection (Cancer Research UK, n.d.). By offering training to employees, this initiative bridges the gap between awareness and actionable insights.

Moreover, educational resources, such as articles from the SIU School of Medicine, play a crucial role in disseminating accurate information about breast cancer (SIU School of Medicine, n.d.). Such articles ensure that the public is well-informed, aiding in dispelling myths and encouraging early detection. Collectively, these multifaceted efforts are shaping a proactive and informed approach towards breast cancer, as reflected in the evolving discourse on Twitter.

5.3 The Role of Digital Platforms in Healthcare

The consistent growth in cancer-related discussions on Twitter over the years underscores the platform's evolving role in healthcare. As healthcare brands increasingly turn to Twitter to guide patients to relevant resources, the platform has emerged as a pivotal tool in steering patients towards necessary healthcare information (*Twitter for Healthcare Brands*, n.d.). Hospitals and healthcare providers are not just limited to sharing the latest research findings but are also leveraging Twitter to improve patient care and foster community engagement (*Tweeting and Treating: How Hospitals Use Twitter to Improve Care*, n.d.).

Furthermore, the rise of Twitter usage among specific medical specialties, such as hematology and oncology, signifies its growing influence in these areas, facilitating the sharing of knowledge and best practices (*Twitter's Role in Hematology and Oncology*, n.d.). Beyond individual specialties, the broader healthcare industry is recognizing Twitter's potential, especially in tracking and predicting disease outbreaks, which showcases its indispensable role in public health (*Twitter's New Role in the Healthcare Industry*, n.d.).

Moreover, as healthcare organizations increasingly use Twitter to disseminate medical information, it's evident that the platform is bridging the gap between medical professionals and the general public, ensuring that accurate and timely information reaches those who need it (*Exploring How Healthcare Organizations Use Twitter*, n.d.). The collaborative nature of Twitter also fosters connections among researchers, physicians, and patients, making it an invaluable tool for advancing biomedical research and promoting health advocacy (*Social Medicine: Twitter in Healthcare*, n.d.).

5.4 Future Implications

The increasing prominence of cancer-related discussions on Twitter signifies a broader shift in how individuals seek and share health information. As the digital age progresses, more individuals are turning to platforms like Twitter to voice their experiences, seek advice, and find community. This evolving landscape presents both challenges and opportunities for the healthcare sector.

For researchers, the vast and diverse dataset available on Twitter can serve as a goldmine

of insights. Patterns in discourse can reveal emerging trends, patient concerns, or even early indicators of public health crises. As more individuals join the platform, the richness and depth

of these conversations are only expected to grow, providing a more comprehensive view of public sentiment and knowledge regarding cancer.

Healthcare professionals stand to benefit from this wealth of real-time feedback. By actively engaging with or monitoring these discussions, they can address misconceptions, provide clarity on complex topics, and offer support in ways that were previously unfeasible. This direct line to patients and the public can also inform patient care, ensuring that it is more aligned with patient needs and concerns.

5.5 Summary

Twitter's role in shaping the global discourse on cancer is undeniable. The platform serves as a mirror, reflecting the multifaceted concerns, hopes, and challenges faced by those affected by the disease. Each tweet, whether it's a personal narrative, a question, or a shared piece of information, adds to a complex mosaic that provides invaluable insights into societal perceptions of cancer.

In this digital era, platforms like Twitter are not just communication tools; they are powerful instruments that can drive change, inform research, and foster community. The conversations happening on these platforms provide a pulse on public sentiment, offering a real-time, unfiltered view into the collective psyche of society.

This study, by delving deep into these conversations, has not only highlighted the current discourse but has also underscored the potential of such platforms in shaping the future of healthcare. As we continue to grapple with the challenges posed by cancer, it is clear that platforms like Twitter will play an increasingly pivotal role in guiding our approach, ensuring it is both informed and empathetic.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The exploration of cancer-related conversations on Twitter from 2009 to 2022 has illuminated the collective consciousness surrounding cancer. The platform has become a crucial medium for patients, healthcare practitioners, and the general populace to voice their concerns, share experiences, and disseminate information.

Several notable trends emerged from the analysis. The rising discussions around ”Breast Cancer Awareness” may be a reflection of increased cases or a testament to successful awareness campaigns. The growth in ”Family Health Concerns” post-2013 reveals the shift in health discourse that emphasizes the holistic well-being of families. The marked spike in ”Emotions & Events” in 2020, possibly linked to the COVID-19 pandemic, showcases the profound influence of global occurrences on personal health narratives.

Overall, the ascending trajectory in cancer discussions underscores the enhanced public interest and the pivotal role of digital platforms in health discourse.

6.2 Future Work

The analysis of cancer-related discussions on Twitter from 2009 to 2022 has offered a window into the public’s evolving dialogue surrounding this pivotal health concern. As we delve into our findings, numerous avenues for more in-depth exploration emerge, promising richer insights and a broader understanding of the cancer discourse. Here are some potential directions:

1. Sentiment Assessment in Depth: Utilizing advanced Natural Language Processing (NLP) techniques can allow for a nuanced sentiment analysis, uncovering the intricacies of public sentiment tied to each topic. This deeper understanding can offer stakeholders pivotal insights into public perception.

2. Regional and Demographic Breakdown: Analyzing discussions by geographical regions or demographic groups might reveal unique perspectives on cancer, driven by variables such as healthcare systems, cultural norms, or socio-economic factors.

3. Short-Term Temporal Analysis: A granular look into monthly or weekly patterns can illustrate the immediate impacts of global events, medical breakthroughs, or awareness campaigns, guiding stakeholders in their communication strategies.

4. Integration with Other Data Sources: Merging Twitter data with other resources like news outlets, academic research, or patient forums can paint a more comprehensive picture, indicating how different platforms influence the broader cancer narrative.

5. Predictive Modeling and Trend Forecasting: Using the data at hand, we can design predictive models to forecast emerging trends in cancer discussions, assisting organizations and policymakers in proactive planning.

6. Exploring Different Social Media Platforms: Platforms like Facebook or Reddit might offer more detailed narratives, queries, or discussions. Analyzing these platforms can provide a complementary view, enriching the broader understanding of public sentiments about cancer.

7. Effectiveness of Awareness Campaigns: Evaluating the impact of awareness drives by correlating them with patterns in Twitter discussions can offer feedback on their effectiveness, shaping future campaign strategies.

In summation, the road ahead is abundant with opportunities for richer analyses and broader explorations. As the digital domain continues to influence public discourse, deriving insights from it is pivotal for real-world health strategies and interventions.

Chapter 7

Reflection

Commencing this study endeavour was both a challenging and enlightening journey. When I first set out to analyse cancer-related Twitter discussions, I recognised the platform's enormous potential as a real-time pulse of public emotion. The journey's early phases were dominated by a key challenge: finding the appropriate datasets.

Finding pertinent datasets was a mission in itself. Although there is an abundance of data in the digital world, finding relevant datasets proved to be a difficult effort. Kaggle is a repository for many datasets, but sorting through them to locate the ones that would be good matches for the project was like looking for a needle in a haystack. A number of datasets were appealing at first sight, but upon closer examination, they lacked the necessary granularity or relevance. This phase made me understand that the beginning of a project, even if it's just searching for data, can set the tone for everything that comes after.

When I finally dived into the tweets, it was like opening a book filled with stories. There were so many different discussions about cancer. Some people shared their personal journeys, while others spread messages about awareness or support. It was touching to see how Twitter became a space for people to connect, share, and find comfort. This project quickly became more than just a task for me; it was a chance to see the real emotions and experiences people had when dealing with cancer.

From a work perspective, the project was a great learning curve. I had to figure out how to sort topics in tweets, clean up the data to make it useful, and ensure I was on the right track. Every challenge taught me something new. There were times I had to try different approaches or rethink my methods. These experiences showed me that research isn't just about getting to the end result; it's about the journey and the lessons you pick up along the way.

Looking back, I realize how much I've grown through this project. Not only did I gain new skills in handling data and understanding topics, but I also learned something bigger from this project. I saw how using technology, like Twitter, can show us what people feel and think. And if we use what we learn in the right way, we can help and make a difference in their lives. This project made me see how powerful technology can be when we use it to understand and help people.

References

Blankenship, E. B., Goff, M. E., Yin, J., Tse, Z., Fu, K. W., Liang, H., Saroha, N. and Fung, I. C. (2015), 'Sentiment, contents, and retweets: A study of two vaccine-related twitter datasets', *The Yale journal of biology and medicine* 88(1), 21.

Blei, D. M. (2012), 'Probabilistic topic models', *Communications of the ACM* 55, 77.
URL: *https://www.eecis.udel.edu/shatkay/Course/papers/UIntrotoTopicModelsBlei20115.pdf*

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of Machine Learning Research* 3, 993–1022.
URL: *https://jmlr.org/papers/volume3/blei03a/blei03a.pdf*

Breast Cancer Now, n. (n.d.), 'Get involved'. Accessed: yyyy-mm-dd.
URL: *https://breastcancernow.org/get-involved*

Cancer Research UK, n. (n.d.), 'Cancer awareness in the workplace'. Accessed: yyyy-mm-dd. URL: *https://www.cancerresearchuk.org/health-professional/awareness-andprevention/cancer-awareness-in-the-workplace*

Chew, C. and Eysenbach, G. (2020), 'Pandemic information seeking: Early findings from the COVID-19 pandemic', *Journal of Medical Internet Research* 22(6), e19707.

Egger, R. and Yu, J. (2022), 'A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts', *Frontiers in Sociology* 7.

Exploring How Healthcare Organizations Use Twitter (n.d.).
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6511983/*

George, L. E. and Birla, L. (2018), 'A study of topic modeling methods', *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)* .

Global Cancer Observatory (2020), 'Global cancer observatory'.
URL: *https://gco.iarc.fr/*

Habbat, N., Anoun, H. and Hassouni, L. (2021), 'Topic modeling and sentiment analysis with lda and nmf on moroccan tweets', *Lecture notes in networks and systems* pp. 147–161.

Hagg, E., Dahinten, V. S. and Currie, L. M. (2018), 'The emerging use of social media for health-related purposes in low and middle-income countries: A scoping review', *International Journal of Medical Informatics* 115, 92–105.

Hidayatullah, A. F., Aditya, S. K., Karimah and Gardini, S. T. (2019), 'Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (lda)', *IOP Conference Series: Materials Science and Engineering* 482, 012033.

Islam, T. (2019), 'Yoga-veganism: Correlation mining of twitter health data.', *arXiv (Cornell University)* .

Korda, H. and Itani, Z. (2011), 'Harnessing social media for health promotion and behavior change', *Health Promotion Practice* 14, 15–23.

Liu, Z., Wang, Y. and Wang, W. (2015), Bayesian parameter estimation in lda, *in* 'International Conference on Computer Information Systems and Industrial Applications'. License: CC BY-NC 4.0.

McClellan, C., Ali, M. M., Mutter, R., Kroutil, L. and Landwehr, J. (2017), 'Using social media to monitor mental health discussions - evidence from twitter', *Journal of the American Medical Informatics Association* 24(3), 496–502.
URL: *https://academic.oup.com/jamia/article-pdf/24/3/496/13063197/ocw133.pdf*

Mohr, J. W. and Bogdanov, P. (2013), 'Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?', *Sociological Methods & Research* 42(3), 293– 319.

Montenegro, C., Ligutom, C., Orio, J. V. and Ramacho, D. A. M. (2018), 'Using latent dirichlet allocation for topic modeling and document clustering of dumaguete city twitter dataset', *Proceedings of the 2018 International Conference on Computing and Data Engineering* .

Mulunda, C. K., Wagacha, P. W. and Muchemi, L. (2018), 'Review of trends in topic modeling techniques, tools, inference algorithms and applications', *2018 5th International Conference on Soft Computing Machine Intelligence (ISCMI)* .

Negara, E. S., Triadi, D. and Andryani, R. (2019), 'Topic modelling twitter data with latent dirichlet allocation method', *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)* .

Nemes, L. and Kiss, A. (2021), 'Information extraction and named entity recognition supported social media sentiment analysis during the covid-19 pandemic', *Applied Sciences* 11(22), 11017.

Ostrowski, D. A. (2015), 'Using latent dirichlet allocation for topic modelling in twitter'.
URL: *https://ieeexplore.ieee.org/abstract/document/7050858*

Pershad, Y., Hangge, P., Albadawi, H. and Oklu, R. (2018), 'Social medicine: Twitter in healthcare', *Journal of Clinical Medicine* 7, 121.

Pratama, T. R., Richasdy, D. and Purbolaksono, M. D. (2022), 'Telkom university opinion topic modeling on twitter using latent dirichlet allocation during covid-19 pandemic', *JURNAL MEDIA INFORMATIKA BUDIDARMA* 6(4), 1816–1825.

Rahman, S., Hossain, S. S., Arman, M. S., Rawshan, L., Toma, T. R., Rafiq, F. B. and Badruzzaman, K. B. M. (2020), 'Assessing the effectiveness of topic modeling algorithms in discovering generic label with description'.
URL: *https://www.semanticscholar.org/paper/Assessing-the-Effectiveness-of-TopicModeling-in-Rahman-Hossain/449b4dd2e36560adeec4fc86543c2d80f15cb424?utm*$_s$*ource = direct*$_l$*ink*

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A. and Boyd-Graber, J. (2015), 'Beyond lda: Exploring supervised topic modeling for depression-related language in twitter', *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* .

Roberts, M. E., Stewart, B. M. and Tingley, D. (2018), 'Building the bridge: Topic modeling for comparative research', *Sociological Science* 5, 785–812.

Sasaki, K., Yoshikawa, T. and Furuhashi, T. (2014), 'Online topic model for twitter considering dynamics of user interests and topic trends', *CiteSeer X (The Pennsylvania State University)* .

Saxton, M. (2018), 'A gentle introduction to topic modeling using python', *Theological Librarianship* 11, 18–27.
URL: *https://api.semanticscholar.org/CorpusID:172077473*

SIU School of Medicine, n. (n.d.), 'Raise your breast cancer awareness'. Accessed: yyyy-mmdd.
URL: *https://www.siumed.edu/blog/raise-your-breast-cancer-awareness*

Social media use among healthcare professionals (2022), *Human Resource Management International Digest* 30(4), 23–25.
URL: *https://www.emerald.com/insight/content/doi/10.1108/HRMID-03-20220045/full/html*

Social Medicine: Twitter in Healthcare (n.d.).
URL: *https://www.jmir.org/2011/2/e33/*

Srinivasan, V. and K, C. (2021), 'Data aggregation of tweets and topic modelling based on the twitter dataset', *2021 the 3rd International Conference on Big Data Engineering and Technology (BDET)* .

Stor´opoli, J., Kang, H. and Pereira, V. (2020), 'Topic modeling: How and why to use in management research', *RAUSP Management Journal* 55(4), 529–545.

Sugawara, Y., Narimatsu, H., Hozawa, A., Shao, L., Otani, K. and Fukao, A. (2012), 'Cancer patients on twitter: a novel patient community on social media', *BMC Research Notes* 5, 699.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3599295/*

Tweeting and Treating: How Hospitals Use Twitter to Improve Care (n.d.).
URL: *https://www.ragan.com/tweeting-and-treating-how-hospitals-use-twitter-toimprove-care/*

Twitter for Healthcare Brands (n.d.).
URL: *https://sproutsocial.com/insights/twitter-for-healthcare-brands/*

Twitter's New Role in the Healthcare Industry (n.d.).
URL: *https://www.hhmglobal.com/knowledge-bank/news/twitters-new-role-in-thehealthcare-industry*

Twitter's Role in Hematology and Oncology (n.d.).

URL: *https://www.cancertherapyadvisor.com/home/cancer-topics/generaloncology/twitters-role-in-hematology-and-oncology/*

Vos, S. C., Sutton, J., Gibson, C. B. and Butts, C. T. (2019), 'Celebrity cancer on twitter: Mapping a novel opportunity for cancer prevention', *Cancer Control : Journal of the Moffitt Cancer Center* 26, 1073274819825826.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6396054/*

World Health Organization (2022), 'World health organization,2022'.
URL: *https://www.who.int/news-room/fact-sheets/detail/cancer*

Yang, S. and Zhang, H.-Y. (2018), 'Text mining of twitter data using a latent dirichlet allocation topic model and sentiment analysis'.

ZDNET (n.d.).
URL: *https://www.zdnet.com/article/we-will-spend-420-million-years-on-social-media-in-2021/*

Appendix A

An Appendix Chapter

A.1 Project Code

- The project's code can be accessed via the provided link.

  https://gitlab.com/msc-datascience-and-advanced-computing/analysing-twitter-conversationsusing-topic-modelling.git

Appendix B An Appendix

Chapter

B.1 Project Specification Form

MSc Project

Specification Form

Version 03

Please complete this form in printed or typed text (Times New Roman size 11). A copy of the approved document will have to be included as appendix in the actual dissertation to help establish to which extend the project has been successful.

Section A

1. Project identification

Proposed dissertation title (maximum 15 words)
Analysing Twitter Conversations On Cancer Using Topic Modeling

2. Student details

Name (*in full*)
Shruti S Pradhan
e-mail address mh825102@student.reading.ac.uk

Term time address : 17, Charwoodhouse, Coley Park. RG16QR
Contact telephone number: +44 7436365623

3. Supervisor *Name and contact details of the staff member*

Name: Dr. Lily Sun
e-mail address: lily.sun@reading.ac.uk

4. **The supervisor** *specification*

The person identified in Section A4 hereby approves the dissertation

Name: *Dr. Lily Sun*

Signature:

Date: 15-23-06

Fill in section A5 and A6, if the project has industrial input.

5. Company Partner *Name of organisation involved in the project*	

6. *Details of contact person* *in the organisation involved in the project*	Title *(e.g. Mr/Mrs/Dr)*	Name	
	Address		
	Tel. No.	Fax No.	e-mail address

MSc Project Specification Form

Section B – Overall Programme

1. Background and Literature review

Please describe in the space provided the background to the project and write a short literature review highlighting relevant developments on the topic of the dissertation

The rise of social media platforms has transformed the way information is shared and consumed in today's digital era. Among these platforms, Twitter stands out for its unique ability to facilitate real-time communication, enabling users to share and access information instantaneously (Social media use among healthcare professionals, 2022). This rapid exchange of information has been particularly impactful in the realm of healthcare, where timely and accurate dissemination of knowledge can be crucial. Healthcare professionals, research institutions, and patients alike have leveraged Twitter to discuss medical advancements, share personal health journeys, and raise awareness about various health conditions (Wu and Feng, 2021). In fact, during global health emergencies, such as the COVID-19 pandemic, Twitter played a pivotal role in the dissemination of guidelines, research findings, and public health advisories (Hagg, Dahinten and Currie, 2018). Such widespread discussions on health topics underscore Twitter's significance as a vital tool for public health communication (Korda and Itani, 2011).

However, the vast and dynamic nature of Twitter data presents challenges. With millions of tweets generated daily, manually sifting through this immense volume of text to extract meaningful health-related information is a daunting task. This is where the power of computational techniques, particularly topic modeling, comes into the picture (Saxton, 2018). Topic modeling is designed to analyze large volumes of unstructured text and categorize them into distinct topics or themes. In the healthcare domain, topic modeling offers a structured and organized summary of the most prevalent health discussions, providing insights into trending health topics, emerging medical research, and public sentiments towards various health issues (George and Birla, 2018).

Research has shown the efficacy of topic modeling in analyzing large datasets, such as those from Twitter, to understand public health concerns and patterns
(Rahman et al., 2020). For instance, by categorizing extensive health-related tweets into distinct themes, researchers can gain insights into the most discussed medical conditions, the efficacy of health campaigns, or even track the spread of health misinformation \citep{yang_2018_text}. Such analysis is invaluable for healthcare professionals, policymakers, and researchers as it provides a pulse on public sentiment and can inform strategies for public health communication (McClellan et al., 2017).

Furthermore, the adaptability of topic modeling is evident in its diverse applications in healthcare research \citep{mulunda_2018_review}. Whether it's understanding patient sentiments, gauging the effectiveness of health campaigns, or identifying areas of misinformation, topic modeling provides a robust framework for drawing insights from textual data (Storópoli, Kang, and Pereira, 2020). Its ability to seamlessly integrate qualitative and quantitative analyses offers a holistic view, bridging the gap between largescale data analysis and detailed interpretations. However, as with any computational technique, the application of topic modeling in healthcare comes with its set of challenges. The quality of insights derived largely depends on the quality of data inputted. Given the informal nature of tweets, which often include slang, abbreviations, and emojis, preprocessing and cleaning the data become paramount for effective topic modeling. Moreover, the dynamic nature of language, especially on platforms like Twitter, means that topic modeling algorithms need to be constantly refined to stay relevant and accurate (Mohr and Bogdanov, 2013).

In conclusion, the combination of Twitter's influential role in healthcare discussions and the power of topic modeling presents a promising avenue for healthcare research (Roberts, Stewart, and Tingley, 2018). By harnessing the potential of these tools, stakeholders can not only stay updated with the ever-evolving health discourse but also inform and shape strategies that are in tune with the needs and sentiments of the public.

MSc Project Specification Form

Section B – Overall Programme (continued)

2. Research question, justification and objectives

Please describe in the space provided the research question to be answered, justify why the topic is important at the present time, and describe the specific objectives of research against which your achievements will be measured.

The primary goal of our research is to undertake an in-depth analysis of discussions surrounding cancer on Twitter, specifically focusing on the UK audience. We aim to probe the diverse aspects of these dialogues, capturing a wide spectrum of subjects. This exploration will provide insights into the behaviors and common public understanding of cancer. Additionally, it will shed light on the personal stories of those affected by the ailment. An integral part of our investigation will be to understand the impact of existing support networks and communities on Twitter. We aim to evaluate how these platforms assist individuals and their families in managing the trials brought about by cancer. Through this comprehensive approach, we hope to deepen our understanding of the digital conversations around cancer in the UK's Twitter space.

To realize this aim, we've set the following objectives:

- Data Collection: Our foremost objective is to amass a wide-ranging set of tweets from the UK that touch upon cancer. This collection will encompass both medically oriented information and personal anecdotes.
- Keyword Utilization: We plan to employ an exhaustive list of keywords to ensure that our research encompasses the full breadth of discussions. This will include regional and local terminologies and phrases to capture the unique UK-centric discourse.
- Data Refinement: Post collection, we will process the data to remove any unrelated content and distractions. This refinement will ensure that we have a concentrated dataset ready for in-depth scrutiny.
- Model Parameterization: An essential objective will be to ascertain the best parameters for our analytical model, ensuring the results are both accurate and insightful.
- Topic Modeling: We will leverage topic modeling techniques to extract primary themes from the amassed tweets. Our focus will be on subjects like lifestyle habits associated with cancer, emotional sentiments, and dialogues about support mechanisms like helplines.
- Methodology Evaluation: Lastly, a key objective will be to meticulously evaluate the methodologies we employ, ensuring they are both rigorous and precise in capturing the crux of the dialogues.

By adhering to these objectives, we aim to offer a nuanced understanding of the UK's discourse on cancer via Twitter. This insight will encompass both the factual aspects of the disease and the personal narratives, along with an examination of the available support structures.

MSc Project Specification Form

Section B – Overall Programme (continued)

3. Methodology

Please describe the methodology that you will use to achieve the objectives stated in Section B2.

The methodology planned for this study is structured to ensure a meticulous analysis of cancer-related discussions on Twitter.

The initial phase involves thorough data pre-processing. The dataset will be refined by removing special characters, numbers, and URLs. Following this, the tweets will be tokenized, breaking them into individual words or tokens. Stop words, which are common words that typically don't contribute significant meaning in textual analysis, will be removed. Furthermore, lemmatization will be applied, converting words to their base or root form, ensuring consistent analysis across the dataset.

The study's central component will be topic modeling. Various topic modeling techniques will be explored to identify the most suitable one for the dataset. Techniques under consideration include Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA). The decision to select a particular technique will be guided by the dataset's characteristics and the research objectives.

Model optimization will be a crucial step in the methodology. The coherence score, which measures the quality of the topics generated, will play a vital role. This score will guide the determination of the optimal number of topics, ensuring they are coherent and contextually relevant.

Visualization tools will then be employed to provide a graphical representation of the topics. The exact tools will be contingent on the chosen topic modeling technique, ensuring compatibility and clarity in visualization.

Once the topics are identified, a manual review will be undertaken to assign descriptive labels to each topic. This topic labeling process will ensure that the topics are easily interpretable and contextually meaningful.

Concluding the methodology will be the evaluation phase. This will involve a qualitative review of the generated topics to ensure their relevance and accuracy in capturing the essence of cancer discussions on Twitter. Apart from the coherence score, the evaluation will also consider metrics like perplexity, which measures the model's predictive accuracy. Moreover, a hands-on qualitative assessment will be done to ensure that the topics align with the contextual meanings and narratives present in the dataset.

In essence, the proposed methodology seeks to use a combination of techniques and tools to provide profound insight into cancer-related discussions on Twitter, capturing both the depth and breadth of the discourse.

MSc Project Specification Form

Section B – Overall Programme (continued)

4. References

Please provide a list of references made in Sections B1, B2 and B3. The formatting of the references must comply with the Style Guide for Technical Reports and Academic Papers.

Hagg, E., Dahinten, V.S. and Currie, L.M. (2018). The emerging use of social media for health-related purposes in low and middle-income countries: A scoping review. *International Journal of Medical Informatics*, 115, pp.92–105. doi:https://doi.org/10.1016/j.ijmedinf.2018.04.010.

Korda, H. and Itani, Z. (2011). Harnessing Social Media for Health Promotion and Behavior Change. *Health Promotion Practice*, 14(1), pp.15–23. doi:https://doi.org/10.1177/1524839911405850.

McClellan, C., Ali, M.M., Mutter, R., Kroutil, L. and Landwehr, J., 2017. Using social media to monitor mental health discussions - evidence from Twitter. Journal of the American Medical Informatics Association, 24(3), pp.496-502.

Mohr, J.W. and Bogdanov, P., 2013. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? Sociological Methods & Research, 42(3), pp.293-319. SAGE Publications Sage CA: Los Angeles, CA.

Mulunda, C.K., Wagacha, P.W. and Muchemi, L. (2018). Review of Trends in Topic Modeling Techniques, Tools, Inference Algorithms and Applications. *2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. doi:https://doi.org/10.1109/iscmi.2018.8703231.

Rahman, S., Hossain, S.S., Arman, M.S., Rawshan, L., Toma, T.R., Rafiq, F.B. and Badruzzaman, K.B.M., 2020. Assessing the Effectiveness of Topic Modeling Algorithms in Discovering Generic Label with Description. In: Arai, K., Kapoor, S. and Bhatia, R. (eds.) Springer International Publishing, pp. 224–236. doi:10.1007/978-3-030-39442-4_18

Roberts, M.E., Stewart, B.M. and Tingley, D., 2018. Building the Bridge: Topic Modeling for Comparative Research. Sociological Science, 5, pp.785-812. SAGE Publications Sage CA: Los Angeles, CA.

Saxton, M.D. (2018). A Gentle Introduction to Topic Modeling Using Python. *Theological Librarianship*, 11(1), pp.18–27. doi:https://doi.org/10.31046/tl.v11i1.506.

Social media use among healthcare professionals. (2022). *Human Resource Management International Digest*. doi:https://doi.org/10.1108/hrmid-03-2022-0045.

Storópoli, J., Kang, H. and Pereira, V., 2020. Topic Modeling: How and Why to Use in Management Research. RAUSP Management Journal, 55(4), pp.529-545. Emerald Publishing Limited.

Wu, P. and Feng, R. (2021). Social Media and Health: Emerging Trends and Future

Directions for Research on Young Adults. *International Journal of Environmental Research and Public Health*, 18(15), p.8141. doi:https://doi.org/10.3390/ijerph18158141.

Yang, S. and Zhang, H.-Y. (2018). Text Mining of Twitter Data Using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis. doi:https://doi.org/10.5281/zenodo.1317350.

MSc Project Specification Form

Section C – Social, legal and ethical issues

Describe Social, legal and ethical issues that apply to your project (If your project requires a questionnaire/interview for conducting research and/or collecting data, you will need to apply for an ethical approval). *Does your project require ethical approval?*	

MSc Project Specification Form

Section D – Work plan

Task No	Task Describe each task to be undertaken and, where appropriate, indicate how it will be carried out. Please consider tasks between the submission of this specification until the submission of the dissertation in August.	Effort *(person weeks)*	Indicate deliverables and decision points (*e.g. data analysis completed, designs completed, software implemented, hardware procured, system configured,..*)
1.	**Literature Review**	**2 weeks**	
2.	**Data Acquisition and Exploration**	**1 week**	
3.	**Data Preprocessing**	**2 weeks**	
4.	**Model Exploration**	**3 weeks**	
5.	**Topic Generation and Labeling**	**3 weeks**	
6.	**Evaluation and insights**	**2 week**	
7.	**Report Drafting**	**4 week**	

MSc Project Specification Form

Section E - Time Plan

For each task identified in Section D, shade the months during which you will be engaged and mark deliverables and decision points.

Task No	February	March	April	May	June	July	August	September
1.			2 weeks					
2.			1 week					
3.			2 weeks					
4.				3 weeks				
5.					3 weeks			
6.					2 weeks			
7.						4 weeks		

MSc Project Specification Form

Section F – Costing

Justification of the budget

Item description	Cost (£) *incl. VAT*

Total	