BERTopic Analysis of AI in Healthcare Discussions on Reddit(words)

April 26, 2024

Srilalitha Lakshmi Anusha Chebolu, Spring 2024

1 Introduction

The integration of Artificial Intelligence (AI) into healthcare represents one of the most significant paradigm shifts in modern medicine. AI's potential to mimic human cognitive functions has led to transformative changes across various healthcare domains, from diagnostic processes to treatment and patient management. Public forums and social media platforms, such as Reddit, serve as vital spaces where individuals discuss and evaluate the implications of AI in healthcare. This study aims to explore the public conversations on AI in healthcare through Reddit platform. By applying BERTopic modeling, a method that leverages natural language processing to extract predominant topics from large text corpora.

The existing literature provides a foundation for understanding AI's applications in healthcare and the general public's attitude towards technological advancements. For instance, Davenport and Kalakota et al. (2019) discuss the operational and ethical dimensions of AI in healthcare settings, noting the technology's potential to enhance diagnostics and personalized medicine. Similarly, Jiang et al. (2017) explores the broad spectrum of AI applications and emphasizes the need for studies that examine public engagement with AI technologies. Building on these studies, this research will delve into discussions within the Reddit community, offering a granular analysis of how AI in healthcare is perceived and debated online.

2 Research Question

What topics are most frequently discussed in Reddit comments about AI in healthcare?

3 Method

3.1 Data

The data for this study is sourced from 38 Reddit posts that dive into discussions about AI in healthcare. These posts were chosen from 10 subreddits: r/ChatGPT, r/singularity, r/medicine, r/technology, r/MachineLearning, r/ArtificialIntelligence, r/healthcare, r/AskReddit, r/datascience, and r/OpenAI, which are recognized for their vibrant dialogue on the perceptions surrounding AI's role in healthcare. For example, one notable post from r/singularity titled 'Rapid AI adoption in healthcare will come from patients, not doctors,' has over 50 comments, and another post from r/ChatGPT, 'AI in healthcare: What do you think?' features 66 comments. The compilation consists of 1,131 comments from these posts. These comments are extracted using the Python Reddit API Wrapper (PRAW) through an authorized instance to efficiently gather first-level comments from each selected post while excluding nested replies to maintain focus on direct responses to the original posts. The posts were manually chosen from the past year (between April 2023 and April 2024) to capture the most recent viewpoints.

3.1.1 Prepocessing

Each comment was assessed by filtering out contributions from bots based on common indicators such as usernames containing 'bot', 'auto', or 'moderator' to make sure the comments are made by a human. Additionally, comments flagged as 'deleted' or 'removed' were excluded to maintain the quality and integrity of the dataset. The dataset, structured in a DataFrame, includes columns for Id, username, the month and year of the comment, and the comment text itself. Below is the sample of 5 records from the dataset.

ID	Username	MonthYear	Comments
1	DependentBonus768	April 2024	r/ChatGPT AI has so much potential to improve healthcare - from helping doctors diagnose to enabling precision care for patients which couldn't have been possible earlier. But AI can also go wrong if ethics aren't made a priority from the start.Open for further discussion on the same!!
2	NoAdvertising1842	March 2024	With using AI algorithm to stratify accumulated medical data in past 50 years to drive precision in diagnosis and match with patient current history. So ai can assign the best treatment plan than whoever MD is there. This may apply to all medical non-interventional cases at least for now
3	Sweenybeans	March 2024	AI is extremely limited at the moment. I think this will be a joke. IBM tried DR. Watson a while ago and it failed.
4	Emmad_1	February 2024	With the kind of money in healthcare fees and expenses, it would be ludicrous to think AI models to replace healthcare workers isn't being made.
5	Eizonix	January 2024	AI's potential in healthcare truly excites me as well—its capacity to transform care is vast. Recently, AI has made strides in personalized medicine, particularly in predicting patient outcomes and treatment responses. What's revolutionary is AI's role in analyzing large datasets to uncover patterns that might not be evident to human researchers. This is not only speeding up the drug discovery process but also helping in tailoring treatments to individual genetic profiles. Another area where AI shows promise is in improving diagnostic accuracy, reducing errors in image-based diagnoses like X-rays and MRIs. AI algorithms can now detect nuances that escape even seasoned radiologists, leading to earlier and more accurate diagnoses. I believe we're just scratching the surface of what AI can achieve in improving both the efficiency and the quality of healthcare.

Table 1: Sample records from the dataset

3.2 Analysis

3.2.1 Topic Modeling with BERTopic

Topic modeling is an unsupervised machine learning technique that's capable of scanning a set of documents(here comments), detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents. The analysis was done by the utilization of BERTopic model, a probabilistic technique for topic modeling that posits comments are a mixture of topics and that topics are a mixture of words.

BERTopic starts by transforming the text data into embeddings using a pre-trained BERT model, which captures the contextual relationships between words in comments. These embeddings are then reduced in dimensionality through UMAP (Uniform Manifold Approximation and Projection) to enhance clustering quality. Following dimension reduction, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is applied to group similar text embeddings into clusters, each representing a potential topic. The model also adjusts clusters based on class probabilities to improve the granularity of topics.

The process began with the Sentence Transformer model "all-MiniLM-L6-v2" to generate embeddings for the textual data, transforming the raw comments into high-dimensional vector space representations. These embeddings capture the language used in the discussions, essential for the subsequent clustering and topic extraction phases. The dimensionality of the embeddings was reduced using the UMAP algorithm, configured to retain five principal components with cosine metric to preserve local and global data structures. Clustering of these reduced embeddings was performed using KMeans, set to identify 15 distinct clusters, representing the discussion topics. For vectorization, a CountVectorizer is employed with settings to eliminate common English stop words, ensuring that only meaningful terms contributed to the

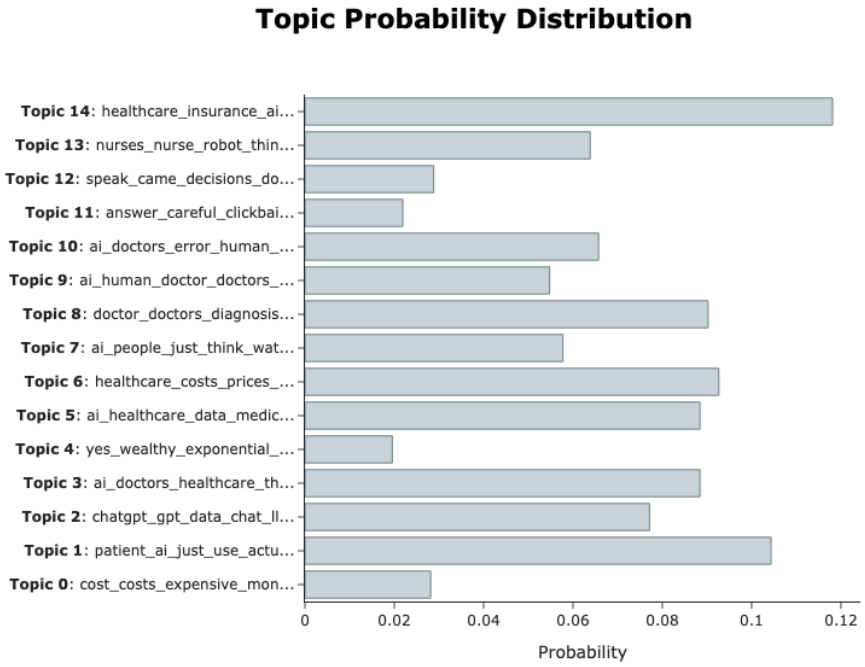

Figure 1: Topic distribution for comment ID = 100

topic modeling. The BERTopic model was then applied, integrating these components with representation model featuring 'KeyBERTInspired' for keyword extraction, 'PartOfSpeech' to refine the context, and 'MaximalMarginalRelevance' for improving topic distinctiveness.

Topic	Count	Name
0	69	0_cost_costs_expensive_money
1	60	1_patient_ai_just_use
2	75	2_chatgpt_gpt_data_chat
3	92	3_ai_doctors_healthcare_think
4	54	4_yes_wealthy_exponential_agreed
5	116	5_ai_healthcare_data_medical
6	78	6_healthcare_costs_prices_cost
7	98	7_ai_people_just_think
8	112	8_doctor_doctors_diagnosis_symptoms
9	101	9_ai_human_doctor_doctors
10	58	10_ai_doctors_error_human — sue ai
11	35	11_answer_careful_clickbait_gave
12	20	12_speak_came_decisions_doctors
13	45	13_nurses_nurse_robot_think
14	85	14_healthcare_insurance_ai_companies

Table 2: Topics Identified in AI Healthcare Discussions on Reddit

4 Results

The probabilities of the comment across 15 topics has been calculated on comment ID =100 which states "I don't think you've met the average patient. Most old people (those that need healthcare the most) can barely use their phone. They are not going to self diagnose and self treat at home." using the 'approximate_distribution' method. The comment is categorized into topic number 14 which has highest probability 0.118. The visualization of the topic distribution(fig 2) depicts this classification, emphasizing the comment's alignment with the identified topic.

The Hierarchical Clustering(fig. 3) visualizes the clustering of topics based on the similarity of their constituent words. Each branch represents a group of terms that are semantically closer to each other, and the length of the branches indicates the level of similarity or dissimilarity between clusters. Larger distances between branches imply less similarity. The tight clustering of certain terms suggests strong relationships and common discourse themes within the discussions.

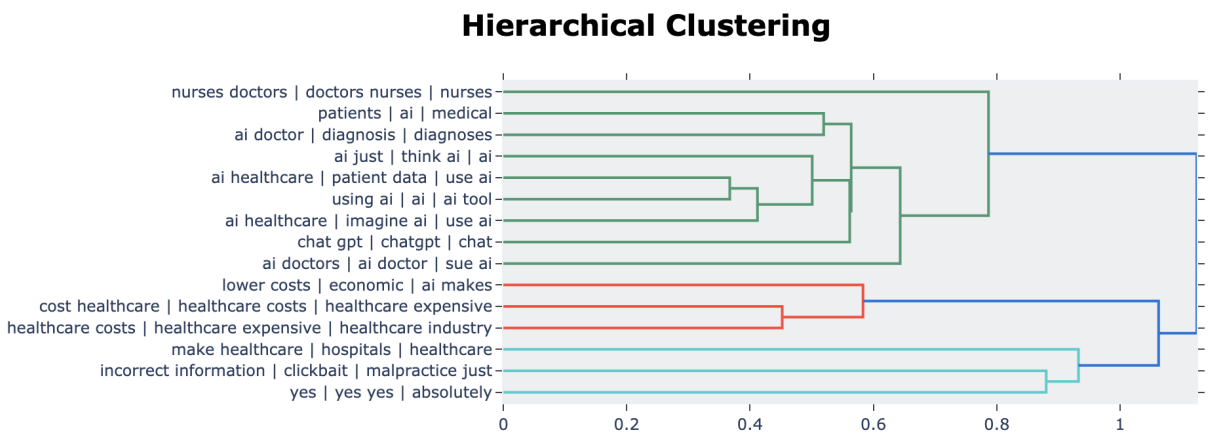

Figure 2: Hierarchical Clustering

The Intertopic Distance Map(fig. 4) provides a visual representation of the different topic clusters generated from the model and how they are situated in relation to each other. The size of each circle likely corresponds to the weight of each topic within the dataset, with larger circles indicating topics with more associated comments. The proximity between any two circles suggests topic similarity; topics that are closer together are more likely to share common terms and concepts. The map's axes, D1 and D2, are principal components that result from dimensionality reduction, usually performed to help visualize high-dimensional data.

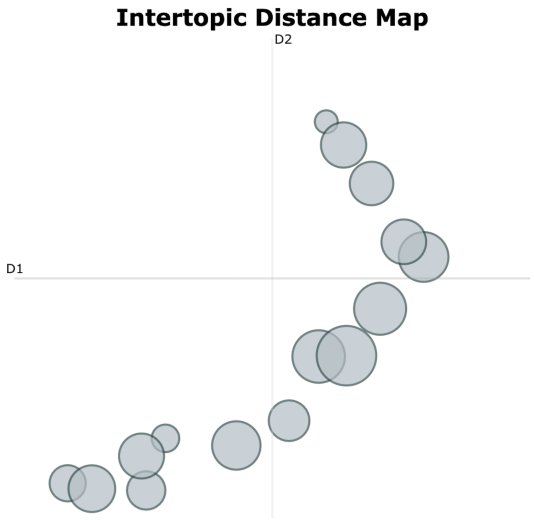

Figure 3: Mapping Topic Similarity

The bar chart(fig 5) presents the word significance within selected topics. Topics 2, 3, 5, 6, 7, 8, 9, and 14 are depicted, each with a distinct set of terms. Here, Topic 6 emphasizes financial aspects: "healthcare," "costs," "prices," and "profit," indicating economic discussions, while Topic 14 suggests a focus on "healthcare," "insurance," and "companies," reflecting conversations around the business side of healthcare AI. The varying lengths of bars across topics illustrate the different emphases given to each term within the discussions, showcasing the diverse angles from which public approach the subject of AI in healthcare.

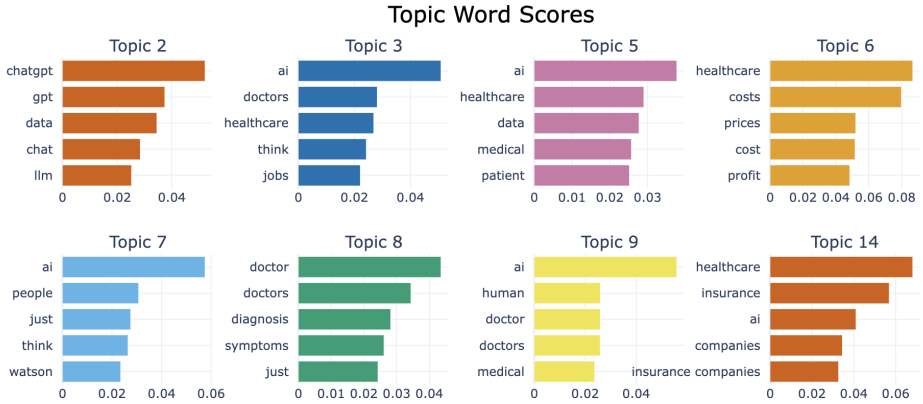

Figure 4: Topic word scores

5 Conclusion and Limitations

The research conducted offers valuable insights into public sentiment surrounding Apple's Project Titan, as reflected in Reddit discussions. Through the application of LDA topic modeling, key topics have been identified that underpin the discourse around Apple's innovation, market strategy, and product development. The sentiment analysis revealed that both before and after Apple's announcement, there was a notable presence of negative sentiment with fluctuations in positivity that ultimately trended downwards, reflecting a shift from cautious optimism to disappointment and concern among the community.

While the study provides comprehensive insights, it is not without limitations. The analysis relies on data from Reddit, which represents only a fraction of global online discourse and may not fully capture the diversity of global perspectives. Additionally, the sentiment analysis tool used, VADER, is optimized for English-language text, which may not accurately reflect sentiment in multilingual discussions.

6 References

1. Davenport, Thomas, and Ravi Kalakota. "The potential for artificial intelligence in healthcare." Future healthcare journal 6, no. 2 (2019): 94.

2. Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. "Artificial intelligence in healthcare: past, present and future." Stroke and vascular neurology 2, no. 4 (2017).