Prescribing Doctors: Topic Modelling Approaches in the AI-led Diagnosis Pipeline

Introduction

This project seeks to evaluate the appropriateness and accuracy of topic modeling approaches in classifying reports of symptoms by patients and primary caregivers without formal training or background in the medical field. Over the past year, there has been an increasing interest in the healthcare industry to create LLM-based diagnostic tools as an alternative to receiving patients in person at hospitals and clinics, especially with respect to symptoms and diseases that do not pose an immediate threat to the patient or others. AI-led diagnostic tools can help conserve resources in an industry that is already overburdened by demand and process inefficiencies.

However, given the current regulatory environment in most countries, it seems unlikely that any AI-led diagnosis effort can function without final approval (and concomitant acceptance of legal liability for malpractice and misdiagnosis) from a licensed healthcare professional in a particular jurisdiction. While this project is very far removed from any attempt to create an LLM-based diagnostic tool, it seeks to understand how topic modeling can aid in the assignment of professionals in relevant specialties to a particular patient report and AI-recommended prescription that they can then approve or modify based on independent analysis.

The Corpus

The corpus consists of posts from two different subreddits: r/AskDocs and r/DiagnoseMe. Both are popular subreddits on the online platform Reddit where users can seek medical advice and ask for help with their health-related concerns. r?AskDocs has a vetting process to “flag” users who are licensed medical professionals, although those who are unlicensed or unvetted can still post their responses. r/DiagnoseMe seems to encourage users to submit pictures as supplements to self-descriptions of symptoms, and allows participation from non-professionals who may have had similar experiences. Users often post medical inquiries, including symptoms, medical history, and any relevant information, for medical professionals on the subreddit who provide insights and suggestions. In many cases, users also post on behalf of their family members or about medical issues faced by those other than themselves. Most users of these subreddits do not seem to possess specialized knowledge in healthcare, which makes the corpus an optimal stand-in for the kinds of reports a future diagnostic tool may receive from lay users.

Pre-processing

From these two subreddits, the top 1000 posts were extracted and the ones consisting solely of images were removed. Top posts are those that have generated the most engagement, as measured through upvotes and the number of comments. The resulting dataset consisted of approximately a thousand two hundred texts. The pre-processing of the corpus consisted only of lemmatization, while stopwords and lower frequency words were removed at a later point as recommended for BERTopic.

Objectives

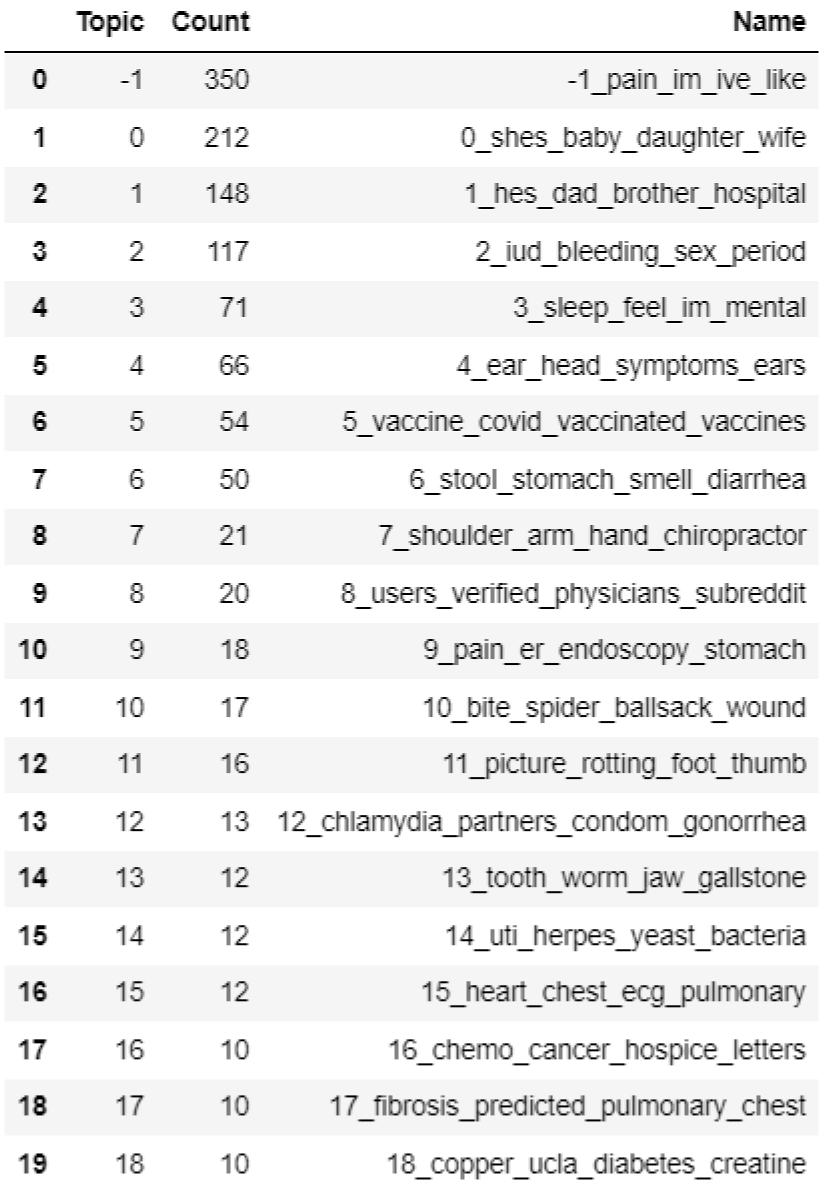
The ideal outcome for this exercise would be a classification of the symptoms report that maps clearly onto either a specialty area in healthcare or a specific set of organs.

Methods

BERTopic is a natural language processing (NLP) technique and library that is based on the BERT (Bidirectional Encoder Representations from Transformers) model and is designed to perform topic modeling on collections of textual data using a two-step process that involves embedding and clustering. While alternatives to BERTopic were explored, none that were currently available open-source seemed more specifically suited to texts that straddled medical and non-medical contexts, and BERTopic consistently produced more accurate results than Top2Vec and LDA in preliminary investigations. Due to constraints built into PRAW, it seemed impossible to extract more than 1000 posts per subreddit. While topic modelling both subreddits separately, the results were much less accurate. After embeddings were generated and documents clustered, stopwords were removed using CountVectorizer, and ClassTfidfTransformer was used to reduce the impact of high-frequency words in the corpus.

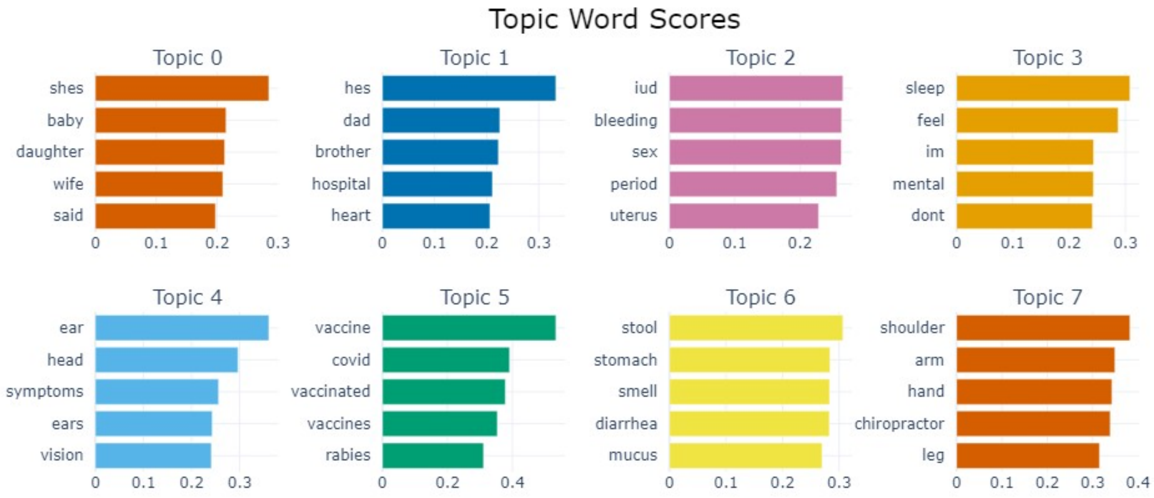
Topic Modelling User-reported Symptoms

The results of the project were fairly encouraging, though the number of topics (18) seemed quite limited. Topic clusters seemed to group accurately based on specific medical topics and body parts easily identifiable with a particular medical specialty, and the impact of extraneous details.



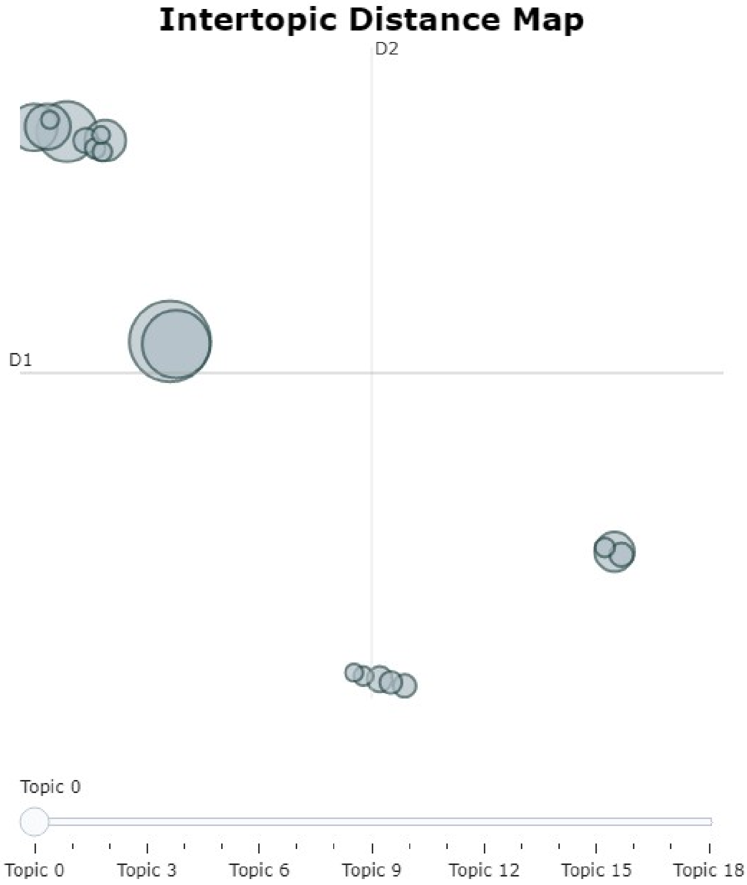
	Topic	Count	Name
0	-1	350	-1_pain_im_ive_like
1	0	212	0_shes_baby_daughter_wife
2	1	148	1_hes_dad_brother_hospital
3	2	117	2_iud_bleeding_sex_period
4	3	71	3_sleep_feel_im_mental
5	4	66	4_ear_head_symptoms_ears
6	5	54	5_vaccine_covid_vaccinated_vaccines
7	6	50	6_stool_stomach_smell_diarrhea
8	7	21	7_shoulder_arm_hand_chiropractor
9	8	20	8_users_verified_physicians_subreddit
10	9	18	9_pain_er_endoscopy_stomach
11	10	17	10_bite_spider_ballsack_wound
12	11	16	11_picture_rotting_foot_thumb
13	12	13	12_chlamydia_partners_condom_gonorrhea
14	13	12	13_tooth_worm_jaw_gallstone
15	14	12	14_uti_herpes_yeast_bacteria
16	15	12	15_heart_chest_ecg_pulmonary
17	16	10	16_chemo_cancer_hospice_letters
18	17	10	17_fibrosis_predicted_pulmonary_chest
19	18	10	18_copper_ucla_diabetes_creatine

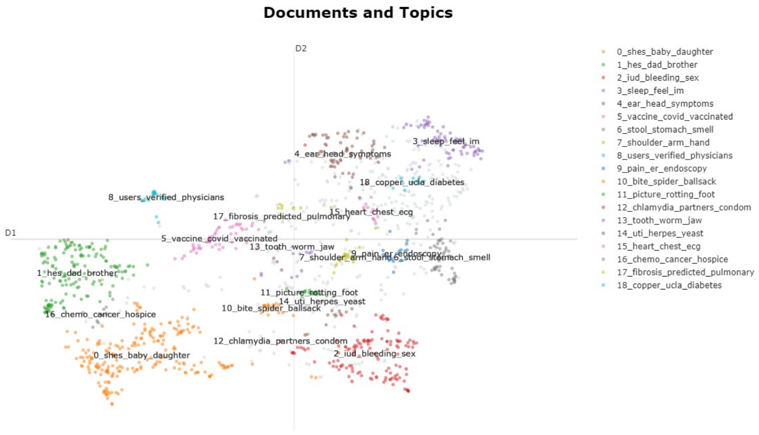
Most of these are easily mappable to specialties (and specialists) such as gynaecology and obstetrics, cardiology, sleep medicine, ENT, orthopedics, urology etc, except for the topics representing COVID-related issues and subreddit housekeeping threads. Also noteworthy is the occurrence of “pain” in the negatives list along with misspelled words that have been excluded. While for the purposes of this project, pain is irrelevant, this information is likely to need special consideration in a hypothetical AI-led diagnostic tool. Descriptions of pain and their classifications will form an integral part of recommendations for further action, and guaging the need to visit an in-person facility and the minimization of panic for the ultimate user, who may or may not be the patient themself.

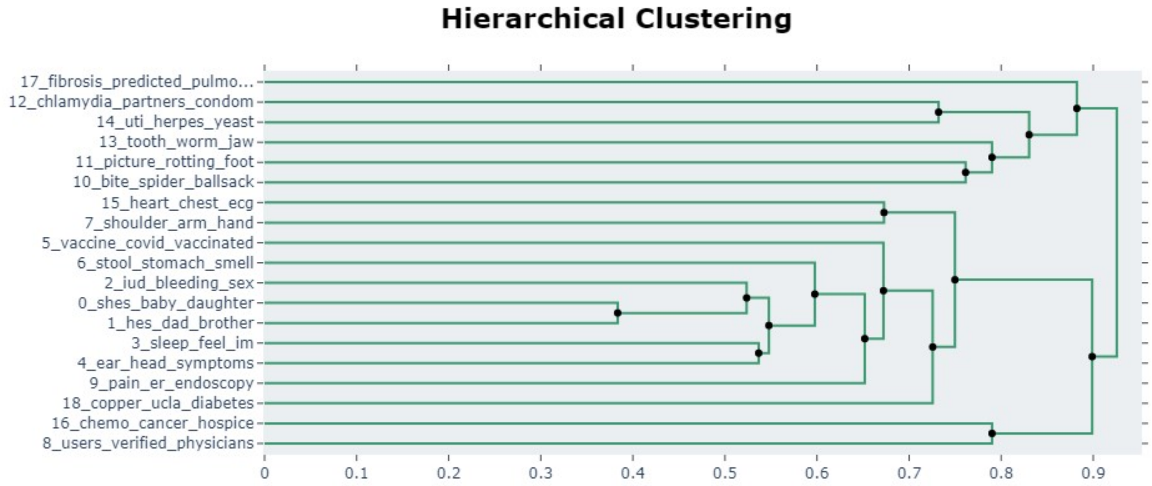


Intertopic Distances

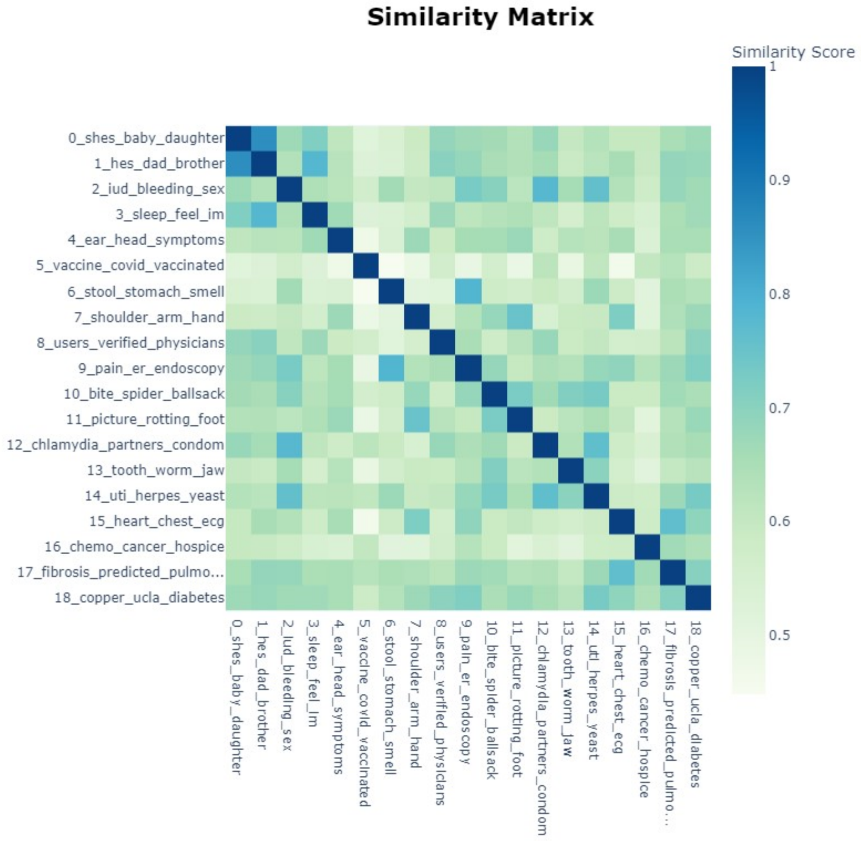
Intertopic distances were also fairly unsurprising, with dental issues being clubbed with gastroenterology, and various limb-realted topics clustering with each other. However, COVID topics seemed also to be closer to the first group that contained gynecological concerns, which is also proximate to topics that represent Ear, Nose & Throat. The two topics on the X-axis are those that contain information about relationships with the presumable patient (baby, daughter, brother etc.)


Intertopic Distance Map
D2
D1
Topic 0
Topic 0
Topic 3
Topic 6
Topic 9
Topic 12
Topic 15
Topic 18


Documents and Topics
D2
D1
3_sleep_feel_im
4_ear_head_symptoms
18_copper_ucla_diabetes
8_users_verified_physicians
17_fibrosis_predicted_pulmonary
15_heart_chest_ecg
5_vaccine_covid_vaccinated
13_tooth_worm_jaw
9_pain_er_endoscopy
7_shoulder_arm_hand
6_stool_stomach_smell
1_hes_dad_brother
11_picture_rotting_foot
14_uti_herpes_yeast
10_bite_spider_ballsack
16_chemo_cancer_hospice
12_chlamydia_partners_condom
2_iud_bleeding_sex
0_shes_baby_daughter
0_shes_baby_daughter
1_hes_dad_brother
2_iud_bleeding_sex
3_sleep_feel_im
4_ear_head_symptoms
5_vaccine_covid_vaccinated
6_stool_stomach_smell
7_shoulder_arm_hand
8_users_verified_physicians
9_pain_er_endoscopy
10_bite_spider_ballsack
11_picture_rotting_foot
12_chlamydia_partners_condom
13_tooth_worm_jaw
14_uti_herpes_yeast
15_heart_chest_ecg
16_chemo_cancer_hospice
17_fibrosis_predicted_pulmonary
18_copper_ucla_diabetes



Hierarchical Clustering



Similarity Matrix

Conclusions

Overall, topic modeling seems to be a particularly accurate approach to classifying symptoms reports by laypersons. It's likely to play a key role in identifying human corroborators and reviewers for the diagnostic work of more complex algorithms that produce prescriptions based on human input.

Future Directions

While this project explored one possible topic modeling approach with BERTopic, the need exists for a model finetuned on bigger corpora of both patient and professional reports of symptoms. There is also a need to develop models which evaluate expressions of pain in patient reports to gauge severity and make recommendations for action timelines.

References

"R/AskDocs." *Reddit*, www.reddit.com/r/AskDocs/.

"R/DiagnoseMe." *Reddit*, www.reddit.com/r/DiagnoseMe/.

Gilson, Aidan, et al. "How does CHATGPT perform on the United States Medical Licensing Examination? the implications of large language models for medical education and knowledge assessment." JMIR Medical Education 9.1 (2023): e45312.

Johnson, Douglas, et al. "Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-gpt model." (2023).

Biswas, Som. "ChatGPT and the future of medical writing." Radiology 307.2 (2023): e223312