Analyzing Gender and Age Differences in Symptom Descriptions from Reddit Posts in the 'AskDocs' Subreddit

Cajsa Schöld
Linköping University
Linköping, Sweden
cajsc235@student.liu.se

Abstract

This project investigates trends in symptom description from posts in the online forum 'AskDocs' for different gender and age groups by using natural language processing techniques. The symptoms are extracted from the posts by using Named Entity Recognition (NER) with the biomedical SpaCy model, `en_core_sci_sm`. The entites are linked to medical concepts in the Unified Medical Language System (UMLS) through ScispaCy. The analysis categorizes symptoms by gender and age, identifying the most frequently mentioned symptoms and the results are visualized using bar plots. The results indicates that there are more females than males using this forum to ask for medical advice and that the most active age group is people between 19 to 30 years old. The most common symptom is pain and we can see that pain is reported more by females than by males which aligns with existing research on gender differences in chronic pain. The project has some limitations, such as a limited and biased dataset and the use of a relatively small NLP model. Addressing these aspects could lead to a more robust analysis.

1 Introduction

The internet has become an essential resource for individuals seeking medical advice and a platform where people can share symptoms, ask questions, and access information(Wu et al., 2024). This project analyzes the posts from an online forum called 'AskDocs' on the platform Reddit. More specially the project investigates whether certain symptoms are reported more frequent by males or females and how different age groups uses this online forum when asking for medical help.

This is interesting because gaining more information and potentially reveling patterns about how different population groups seek medical help and describe symptoms can be used in medical research to improve healthcare communication. To be able to identify and address these biases could possibly lead to better diagnosis.

Therefor this project aims to analyze gender and age differences in the frequency and type of symptoms posted in the online forum by using the natural language processing techniques named entity recognition and concept linking.

2 Theory

In this section the theoretical background to the key concepts used in the project is presented.

2.1 Named Entity Recognition

Named entity recognition (NER) is a NLP technique used to identify entities in texts and classify them into categories. This project uses NER to extract the described symptoms from the posts. The entities in this project are therefor the medical symptoms and they are extracted by using the NLP software library SpaCy.

SpaCy is a robust NLP library that provides several different pre-trained models. In this project the SpaCy model `en_core_sci_sm` is used, which is adapted to process biomedical data and has been trained on scientific and medical texts(Explosion AI). It can for example identify medical entities such as symptoms, bacteria, medication etc. which makes it better suited for the data in this project than more general NLP models.

2.2 Concept Linking

After receiving the entities we want to link them to broader medical concepts in a knowledge database. Concept linking is similar to entity linking but the terms are linked to broader and more geneal concepts(Mohan and Li, 2019). The knowledge database used in this project is the UMLS.

2.3 Unified Medical Language System

The Unified Medical Language System, or UMLS, is a biomedical knowledge developed by the U.S.

National Library of Medicine. It aggregates several biomedical dictionaries into one unified framework. The key features of UMLS are CUIs, which are unique identifiers assigned to each concept, and the TUIs, which are the different categories of the concepts (e.g. symptom, disease)(National Library of Medicine). These features improve consistency within biomedical datasets. The Similarity is a measure of the semantic similarity between the pair of medical concepts. In the example below we can see that the pair 'painful' and 'pain' have a very high similarity value, while 'painful' together with 'hand pain' have a lower similarity.

This is an example of an output when using concept linking to UMLS on a text including the term 'painful':

- painful:
  - CUI: C0030193, Name: Pain, Type: Symptom (T184), Similarity: 0.97
  - CUI: C0239833, Name: Hand pain, Type: Symptom (T184), Similarity: 0.85
  - CUI: C0013456, Name: Earache, Type: Symptom (T184), Similarity: 0.85
  - CUI: C0239377, Name: Arm pain, Type: Symptom (T184), Similarity: 0.84
  - CUI: C0003862, Name: Arthralgia, Type: Symptom (T184), Similarity: 0.82

3 Data

The data used in this project is as previously mentioned collected from a discussion forum on Reddit called 'AskDocs'. This since it provides authentic human written text and represents the real concerns of people seeking for medical advice.

The data is collected by using the python Reddit API wrapper "PRAW". The 1000 latest posts from the forum was collected. The posts must contain the persons age and sex and this information (usually in the form 23F or M46) are used in the analyzis. The posts not containing information about sex and gender are filtered out, as well as posts not containing any symptoms. This gives that all 1000 posts can not be used and the final dataset size is of 329.

The preprocessing of the texts included converting all characters to lowercase, removing special characters and removing access blank spaces.

4 Method

4.1 Data Collection

The 1000 most recent posts were fetched using the `subreddit.new()` function. From each post the age and gender was extracted using regular expression patterns and the posts not including this information was filtered out and not included in the analysis. The texts are prepossessed as described in the section above.

4.2 Symptom Extraction

The symptoms are extracted by Named entity recognition with SpaCy, specifically the biomedical model `en_core_sci_sm`. The extracted entities are linked to the UMLS using the ScispaCy linker which is added to the model. Only the entities with the TUI 'T184', representing sign or symptom, are used in this analysis, and the name component of the symptom is extracted.

4.3 Grouping and Analysis

Using the age extracted, the posts where grouped into age ranges (0-18, 19-30, 31-45, 46-60, 61-75, 76+). The symptoms where aggregated for each group and the appearance of each symptom for each age range was calculated. The 10 most common symptoms was extracted and the result was visualized by plotting as bar charts, one where the frequency of each symptom where on the x-axis, and one plot where the frequency was normalized for each age group.

A similar approach was used to visualize the gender differences, where the 10 most common symptoms for each gender is extracted and plotted in a bar chart.

5 Results

Age Group	Count
0-18	36
19-30	189
31-45	79
46-60	20
61-75	5
76+	0
Total	329

Table 1: Number of approved samples for each age group.

The results show trends in how symptoms are described based on gender and age. We have a

Gender	Count
Female	204
Male	125
Total	329

Table 2: Number of approved samples for each gender.

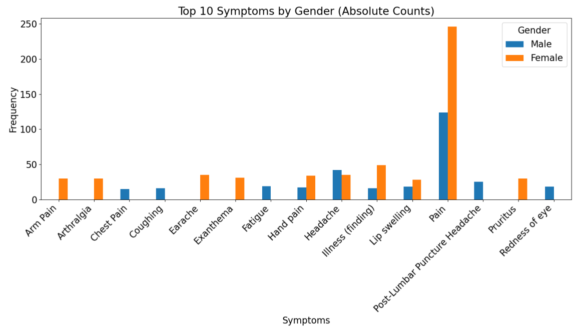

Figure 1: Comparison of the top 10 most frequently reported symptoms by gender in the ’AskDocs’ subreddit posts. The number of apperences of the symptoms are shown on the y-axis

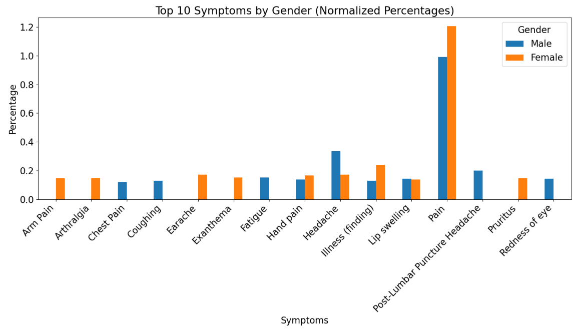

Figure 2: Comparison of the top 10 most frequently reported symptoms by gender in the ’AskDocs’ subreddit posts. The frequency is shown as a proportion of the total posts for each gender.

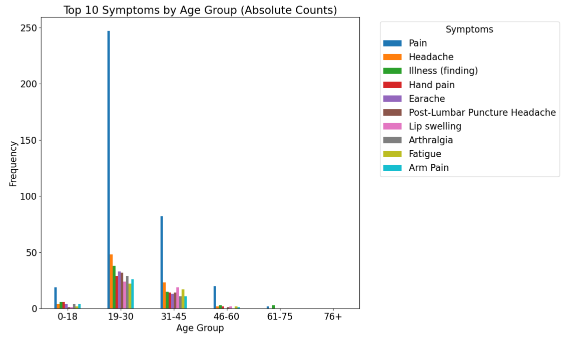

Figure 3: Distribution of the top 10 symptoms across different age groups. The bars represents the frequency of the symptoms.

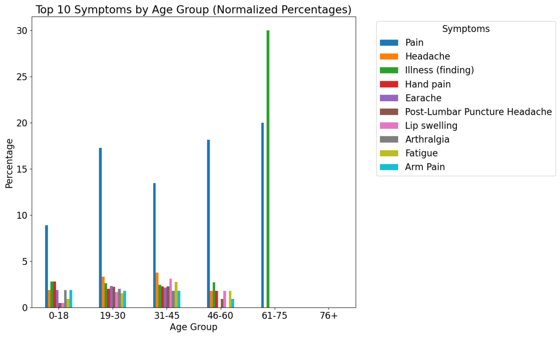

Figure 4: Normalized distribution of the top 10 symptom. Frequencies are expressed as percentages of the total symptoms reported within each age group.

larger part of female samples, and for the age range is 19-30 the largest group. Across all groups is ’pain’ the most common symptom. The normalized data illustrates proportional differences in symptom reporting between genders.

6 Discussion

From the results we can see that the age group 19-30 are the most frequent users of this Reddit forum when searching for medical advice. This aligns with Reddits overall user demographics, where the largest group of users (44 percent) are between 18-29 (Rush, n.d.). Regarding the different symptoms reported by the different age group we can see that pain is the most reported for almost all age groups, and regarding the rest of the symptoms we can not draw any conclusion due to a very limit dataset, especially for the older age groups.

We can also see that there are a larger number of females than males using this forum. This differs from Reddits user demographics, where the majority of 61.2 percent are males and 37.8 percent females(Rush, n.d.). A study from 2015 has shown that females use the internet more for health related questions, which can be an explanation to this. The study discuss that the potential reasons for this might be social structures which have made females more proactive when it comes to health, and also that females are often more targeted by health campaigns (Fox and Duggan, 2015).

From figure 1 we can see that the most common symptom is pain, and that it is reported more times by females then by males. This agrees with research in the area which shows that chronic pain is more common for females as for example migraine, musculoskeletal pain and irritable bowel syndrome. Additionally there are several chronic pain diagnosis that only occur in females, as for example endrometriosis and other gynecological-related conditions. (eClinicalMedicine, 2024)

6.1 Limitations

Using Reddits's API wrapper a maximum of 1000 posts can be collected, which limits the amount of data used in the analysis. Many of the posts also gets filtered out because they does not include the gender and age of the person, resulting in a even smaller dataset. Additionally, the small ScispaCy model is used in this project, which is efficient, but it may limit performance in recognizing and linking medical entities. Using a larger, more advanced model could improve the results.

Since the 1000 most recent posts are collected it result in that only posts from the last couple of days is analyzed and the data does not capture any broader patterns and is not representative over time. To make the data less biased it would be better to collect a random sample of the posts in the forum.

7 Conclusion

This project has investigated trends in symptom descriptions in the online forum for different demographic groups. In summary we can see that there are more females using this forum to ask for medical advice and that the symptom 'pain' is the most common for both genders but it is reported more frequently by females. This agrees with previous research stating that chronic pain is more common in females than males. Regarding the different age groups we can conclude that in the age group 19-30 is the most common group to ask for medical advice in this, which also aligns with Reddits user demographics. To conclude this project has investigated demographic trends in online health-seeking behavior using NLP techniques, and gaining more information like this can hopefully be used to improve healthcare communication.

References

eClinicalMedicine. 2024. Gendered pain: a call for recognition and health equity. *eClinicalMedicine*, 69.

Explosion AI. Scispacy: A natural language processing toolkit for biomedical and scientific text.

Susannah Fox and Maeve Duggan. 2015. Health information-seeking online: A field driven by women. *Journal of Medical Internet Research*, 17(6):e156.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *Automated Knowledge Base Construction (AKBC)*. Conference Paper.

National Library of Medicine. Unified medical language system (umls).

Recreation Rush. n.d. Reddit user statistics: A breakdown of reddit's user demographics. Accessed: 2025-01-17.

Bangan Wu, Qianqian Ben Liu, Xitong Guo, and Chen Yang. 2024. Investigating patients' adoption of online medical advice. *Decision Support Systems*, 176:114050.

A Example Appendix

This is an appendix.