Automated System for Opinion Detection of Breathing Problem Discussions in Medical Forum Using Deep Neural Network

Somenath Nag Choudhury[1] and **Asif Ekbal**[2]
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Bihar
India, 801106
[1]somenath.nc@gmail.com, [2]asif@iitp.ac.in

Abstract

Chest X-ray radiology majorly focuses on diseases like consolidation, pneumothorax, pleural effusion, lung collapse, etc., causing breathing and circulation problems. A tendency to share such problems in the forums for an answer without revealing personal demographics is also very common. However, we have observed more visitors than authors, which leads to a very poor average reply per discussion (3 to 12 only), and also many left with no or late replies in the forums. To alleviate the process, and ease of acquiring the best replies from multiple discussions, we propose a supervised learning framework by automatic scrapping and annotation of breathing problem-related group discussions from the patient.info[1]forum and determine the associated sentiment of the most voted respondent post using Bi-LSTM. We assume the most voted reply is the most factual and experienced. We mainly scrapped and determined the sentiment of bronchiectasis, asthma, pneumonia, and respiratory disease-related posts. After filtering and augmentation, a total of 1,748 posts were used for training our Stacked Bi-LSTM model and achieved an overall accuracy of 90%.

1 Introduction

Opinionated feedback (Mäntylä et al., 2018) specifically in medical interactions has the remarkable capability to affect public sentiments towards better healthcare. People use medical forums or medical blogs to easily access health-related information and to get mental support from people in similar situations. It also fascinates the practitioners, medical experts, and medical representatives for their better personal, societal, and business enhancements (Chen, 2013). The Original Poster (OP) posts with confusion, stress, or anxiety; then, the sentiment changes by the replies of the Respondent Posters (RP) and ultimately stops. The sentiment expressed has two major predictors: time and author and different aspects of them (van Uden-Kraan et al., 2008).

These long illustrated online honest confessions with detailed briefing are better than in-person conversations (Davcheva et al., 2019) and create Computational Health Mines (Bobicev and Sokolova, 2018). This facilitates the intervention of the mining experts (Mansingh et al., 2009), who can help in greatly reducing healthcare costs, by more than $300 per year (Pramanik et al., 2020). The government initiatives are also evident regarding the sharing of information among patients (Wang et al., 2019), opinion against vaccination (D'Andrea et al., 2019), adhering to agency-level complaints (Bastani et al., 2019) etc. Close to 47% of the online users (Yadav et al., 2018a) looking for mainly three types of support from these Online Health Communities (OHC). These are support by information, support by emotion, and support by companionship (Balakrishnan et al., 2021). But, quality assessments to improve the Quality of Life (QoL) of these supports, is difficult due to the contributions from diversified knowledge retainers (Kamalov et al., 2017). Also, facts and experience can play a pivotal role in such cases (Carrillo-de Albornoz et al., 2019).

These posts not only help in interacting with the most influencing participants known as Opinion Leaders (Bamakan et al., 2019), but also allow to form of small groups with the same opinion, called Echo Chambers (Cinelli et al., 2021). The basic category of sentiments, experience, facts, and opinions associated with this unstructured and noisy information differs by the use of functional keywords (Ali et al., 2013) and behavioral patterns (Balaji et al., 2021). This requires intelligent questionnaires (Opitz et al., 2014) and careful intervention of medical lexicons(Su and Peng, 2012), to alleviate the time and processing complexity.

Instead of having potential scopes in applica-

[1]https://patient.info/forums

tions like consumer satisfaction, relation extraction, change of role, regular and comparative opinions detection, and opinion polarization, there exist multiple challenges in maintaining and controlling the posts, engaging more authors than visitors, delay in reply, ability to inherit the conceptions and post-completion feedback(Davazdahemami and Delen, 2019; Sokolova and Bobicev, 2013; Liu, 2012; van Uden-Kraan et al., 2008).

We have closely observed the structure of the forum and found that under the section "Condition and Medicine Categories", there are a total of 30 categories. We have chosen the category "Chest and Lungs". There are a total of 11,754 members and 64,125 posts in 15 groups (as of 07.11.2023). Each of these groups is associated with some discussions, distributed over multiple pages. Each discussion is associated with a discussion title, original post, and replies, and by default sorted by oldest.

The original post mostly bears negative sentiments like depression, fear, anxiety, sorrow, helplessness, frustration, etc. An OP expects a most-voted comment to be of a person who has recovered from a similar situation in a guided way. Thus, a positive comment must be associated with self-experience and suggestions from an RP. But, OP may ignore a comment, sharing mere sympathy or a comment from a person yet to recover. Thus, for each post statement (P) by the Original Poster (OP), satisfying the minimum number of replies, we have to automatically fetch the maximum voted comment (Q) and determine the opinion expressed in Q as positive, negative, or neutral based on information support, self-experience, and emotional support.

We have contributed to this paper in the following ways: 1) Filter-based automatic scrapping of posts to consider posts with maximum votes only. 2) Annotating based on three key components, self-experience, suggestion, and disease recovery. 3) Stacked BiLSTM model for opinion classification.

2 Related Work

The landscape of opinions associated with health (physical or mental), medicine (good or bad), treatment (working or not), and medication (effective or not) encourages verse range of initiatives.

Automated System Generation. An automated medical assistant can help in the constant analysis, monitoring, and recommendation against patient emotions and sentiments thus reducing human efforts and time. The Semantic Knowledge-Based Graph Network (Sem-KGN), a textual entailment approach with domain knowledge and entity relation extraction was used for automated question answering achieved 56.17% accuracy better than BioBERT (Yadav et al., 2020). A sentiment-aware recommendation system developed by (Aipe et al., 2019) with a CNN-LSTM-CNN-based ensemble model for suggestive treatment reported to improve the baseline by 9.13%. An automated scrapping system was implemented by (Baskaran and Ramanujam, 2018) with a well-studied structure and representation of the posts in the medical forum achieved 100% accuracy.

Surgery and Treatment. The sentiment associated with pre and post-quality of the socio-personal life of the patient is a major concern for surgery and treatment. A sentence-level feature-based learning (Ali et al., 2013), applying medical questionnaires about the endurance (Opitz et al., 2014), determining the narrative differences (Balakrishnan et al., 2021), or thematic analysis (Sinha et al., 2018) found to be useful techniques for such opinion detection. Feature-based approach with SVM classifier reaches 64.2% for opinionated and non-opinionated posts and sentiment embedding approach gives 91.9% accuracy for patient emotion.

Topic Modelling. Automatic detection of emergent themes of sharing experience and advice, or feeling motivational from the forum post can be done using question-answering (Buchanan and Coulson, 2007), “Social Support Behaviour Code”, (Perrone et al., 2015), Unique coding categories similar to content retrieval from online support groups for the mentally challenged developed by Perron (van Uden-Kraan et al., 2008) (Chen, 2012) etc. The result reveals sharing personal experiences (30%-71%) was the most common followed by providing information (26%-70%) and support (23%-40%).

Drug Reviews. The impact analysis on the market of an operational drug is of utmost importance for a quality healthcare ecosystem. Attempts are made to determine the output of such reviews for a medical condition and medication either separately (Yadav et al., 2018a) or in a multitasking environment (Yadav et al., 2018b). Sentiments like exists, recovery, and deteriorates are used for medical conditions and effective, ineffective, and serious adverse effects are for medications. An improvement was also witnessed by introducing medical context (Yadav et al., 2018c). Adverse Drug Reaction (ADR) was observed to be identified with an accuracy of

59.74 % and 77.33%. A hybrid approach of lexicon and learning-based was adopted, reflecting a significant accuracy of 96%. (Saad et al., 2021)

Emotion Classification. Emotions associated with the posts also cover a diversified range of "basic" emotions like joy, sadness, anger, encouragement, gratitude, etc. In the pursuit of capturing, a 6-class (encouragement, facts, confusion, gratitude, facts +encouragement, and uncertain) classification was proposed by (Bobicev and Sokolova, 2014). They used a newly created lexicon (HealthAffect) and two learning algorithms NB and KNN, and achieved an F-1 score of 51.8%. The ontology category-based approach (Bobicev and Sokolova, 2018) achieved 80% and above accuracy score for each category (eg: Intakes 95.9%, confusion 94.2%, symptoms 88.4%, etc.). An approach for categorization of different active users and six(6) sentiment classes was also done to realize the accuracy of 45% (Sokolova and Bobicev, 2015).

3 Proposed Methodology

3.1 The Problem Statistics

As per Table 1, we found the members as mostly visitors for each 15 groups. For eg. the group "Respiratory Symptoms and Disorder" has a total of 711 discussions with 5,031 replies leading to an average reply per discussion of 7.07, instead of having 1,739 members. This necessitates an automated approach for filter-based scraping method to consider a discussion with the highest vote, and then analyze the sentiments.

3.2 Problem Statement

To generalize the conception, we can consider Group G (eg. "Asthma") as a collection of some topics of discussion, {TD1, TD2,..., TDK}. Each TDi is associated with some number of replies (NRi). For a topic of discussion TDi if its NRi>= T, where T is the minimum reply criteria, we need to consider the TDi for further processing. Now assuming, TDi={R1,R2,...Rp}, Ri is the i-th reply and p>=T, each Ri is associated with some number of votes NVi. We need to next find the Rj, where Rj= Reply{Vj=max{NV1,NV2,..,NVp}}. We need to then process this Rj and determine its sentiment. We have considered Q as positive if it contains both self-experience and support information, optionally with emotional support. Q is neutral if either support information or self-experience exists but not both. The negative label is used if only accompanied by emotional support but no information or experience, and also the RP is uncertain about the outcome because himself under treatment. The data annotation scheme is described in Figure 1.

Disease	Avg.Reply	Members
Asthma	5.72	744
Breast Pain	3.64	982
Bronchiectasis	10.63	684
Chest Pain	4.36	1,295
COPD	11.44	1,537
COVID-19	11.66	570
Costochondritis	9.99	875
Fungal	8.44	975
Pneumothorax	7.72	231
Pulmonary Embolism	8.44	479
Pulmonary Fibrosis	8.45	335
Respiratory Symptoms	7.07	1,739
Sarcoidosis	12.94	447
Smoking	6.00	340
Steroids	8.45	521

Table 1: Avg. reply per discussion statistics for "Lungs and Chest" category

Statement (P)	Comment (Q)	Label of Q	Label Reasoning
Went to doctor two weeks ago. We did a sputum test. It came back Mrsa. Anyone have this result. No antibiotic nothing. Is this common with bronchiectasis.	Hi there I contracted mrsa when I was in hospital. You can catch it just by touching someone who is a carrier it's highly contagious and is treated with antibiotics. My advice is go the doctors as it can sometimes be serious if you have BX or other health issues. Good luck, Rach	Positive	Support by companionship (self-experience) + Support by information (suggestion) + No Emotional Support
How many of you have bouts of seeing reddish or blood tinged sputum? I know it is not uncommon in BX patients but it flips me out every time it happens! . If you do experience it, does it signal infection to you or could it just be irritation in the airways?	Oh gosh. Hemoptysis is my primary symptom (not blood-tinged sputum--just blood), usually several times a month. (I've had it a couple decades now.) And it tends to occur separately from active infection, so my pulmonologist treats it as caused by inflammation and tissue damage. I get corticosteroids and antibiotics if it's particularly heavy or frequent.	Positive	Support by companionship (self-experience) + Support by information (suggestion) + Support by Emotion
If I do not eat all morning I am ok. Then try to eat a LITTLE bit and the BIG coughing starts. I have tried eliminating wheat, dairy etc. But still cough. Getting VERY frustrated.!!!	I have the same trouble. Eating usually precipitates my coughing. It's embarrassing if I'm around other people! So strange, it's not like I'm aspirating my bagel. For me, it's just another example of how our bodily functions are more connected than we know.	Neutral	Support by companionship (self-experience) + No Support by information (suggestion)
I'm a 22 year old female. I first noticed this several months ago when I started kick-boxing. I've never been a very active person, and then I thought it was only when I was breathing out, so I didn't worry about it much. I figured wheezing while exercising for an overweight person was expected. (cropped)	Even a little bit of wheezing isn't normal if you are overweight. You need to speak to your doctor about this. Do you get allergies?	Neutral	No Support by companionship (self-experience) + Support by information (suggestion)
hi i am a 26 year old male who has had severe air hunger for going on a entire year its very bad i gasp and gulp air every minute of every single day im gasping yawning sighing and gulping once a minute at the very least... (cropped)	Hi are you in the uk or america? you seem really stressed but dont seem to be getting anywhere with it which is not doing you any good. id suggest you need to see a Specialist i apologise about the typing my ipad has amind of its own tonight!	Negative	No Support by companionship (self-experience) + No Support by information (suggestion) + Support by Emotion Only
Hi all! So after having multiple spontaneous pneumothorax on my right and left lung I had VAT pleurodesis on both of my lungs. My right lung was done last December, everything was ok until August, in which I had a slight pneumothorax. (cropped)	I'm in the same boat. So far I waited 5 weeks the lung got better but now here I am again with the bubbling noise. I do not want that surgery. Not sure what my future is going to be.	Negative	Not recovered yet

Figure 1: Data Annotation Scheme with reasoning

Input: URL of a Group G
Output: A CSV file containing 1. Original Post (P) with T>=2, 2. Maximum voted Respondent Post (Q), 3. Sentiment expressed in Q as positive, negative, or neutral.

3.3 Dataset Description

We have initially fetched (a total of 1,774) all the discussions, from the four discussion groups Asthma, Bronchiectasis, Pneumothorax, and Respiratory for filtering. The maximum voted replies (a total of 874) were then filtered and no-reply discussions were eliminated. The disease-wise dataset

Disease Group	A	B	C
Asthma	294	216	432
Bronchiectasis	656	292	584
Pneumothorax	118	85	170
Respiratory	706	281	562
Total	1,774	874	1748

Table 2: Disease-wise Data in use [A=Total Discussion, B= Total Filtered Discussion, C= Total Augmented Discussion]

Disease Class	A	B
Positive	307	614
Neutral	448	896
Negative	119	238
Total	874	1748

Table 3: Class Distribution [A= Before Augmentation, B= After Augmentation]

statistics and their class distribution along with augmentation are reflected in Table 2 and Table 3 respectively.

3.4 Methodology

Our methodology can be simplified using the following steps. Step 1: Identification of the URL and class structure of the specific group (Eg. Asthma) and page elements (Eg. Discussion Titles, Comments, no.of pages, etc.). Step 2: Iterate each page (containing some number of discussions) of the group until no next page is available. For each page for each discussion, we have fetched the discussion title, discussion page link, and no. of comments in a .csv file. Step 3: After generating all such group-wise .csv files, for each such file, for each link, we have fetched the statement and comments only if no.of comments is more than two (T>=2). Step 4: For each such link satisfying the criteria, we performed the pre-processing and augmentations followed by deploying the BiLSTM model.

3.5 Implementation Specifications

To help in implementing the methodology, we have used the following specifications.

Step 1: The Puppeteer JS framework was considered for the automation purpose and a node.js backend was used to serve the automation logic. A new window was opened with a specific url using the puppeteer.launch() command.

Step 2: Then the entire XHTML data was being parsed line-wise, to find the required content. First, we needed to find the div.className associated with the cookies agreement button. If present, we initiate a button.click() event to handle the change and proceed with the rest of the work. After the entire Document was loaded in the browser, we counted the total number of pages available for that particular group (G),using the submit.reply-pagination tag. Then we iterate over each page and find the post title, href link associated with each thread, and their corresponding reply count for that particular thread using the cardb__block div element. The cardb__block was evaluated using getAttribute method for the 'href' value and textContent. This href URL, number of comments, and post title was further stored in a <DiseaseGroup>.csv file. After completion of each page, the "Next" button span.className was found and a button.click() event was triggered. This process continues until no Next Page button is available.

Step 3: For each such <DiseaseGroup>.csv created for a particular disease-group data, all the discussion page links were read, and for each link a new window was opened with maximum voted reply as a query parameter. Using the .h1.post__title tag query selector, we scraped out the entire thread title, thread post content using .input.moderation-conent tag and comment for each post using similar .input.moderation-conent inside of .post__content.break-word tag. After storing the data using an array, the array was written back to another csv file using a csvWriter.writeRecords() module.

Step 4: The experiments were done on a 3-layer Stacked Bi-LSTM network. The embedding dimensions used are input_dim as 10,000, and output_dim fixed as 100, along with pre-trained weights like Gensim. The input_length parameter referred to the maximum length of text sequences used as input. Each of the 3-layer Bi-LSTM layers consists of 128 units, with an interfacing 60% dropout layer added for regularisation to prevent over-fitting. In the end, a dense Layer was used as the output layer for all classification tasks with a softmax activation on 3-classes of data. Finally, the model was compiled using Categorical Cross-entropy Loss for multi-class classification. Adam optimizer was used as the optimization technique. The model was trained for 100 epochs, with a learning rate of 0.0001. Early Stopping and Model Checkpoints were also used as callback functions.

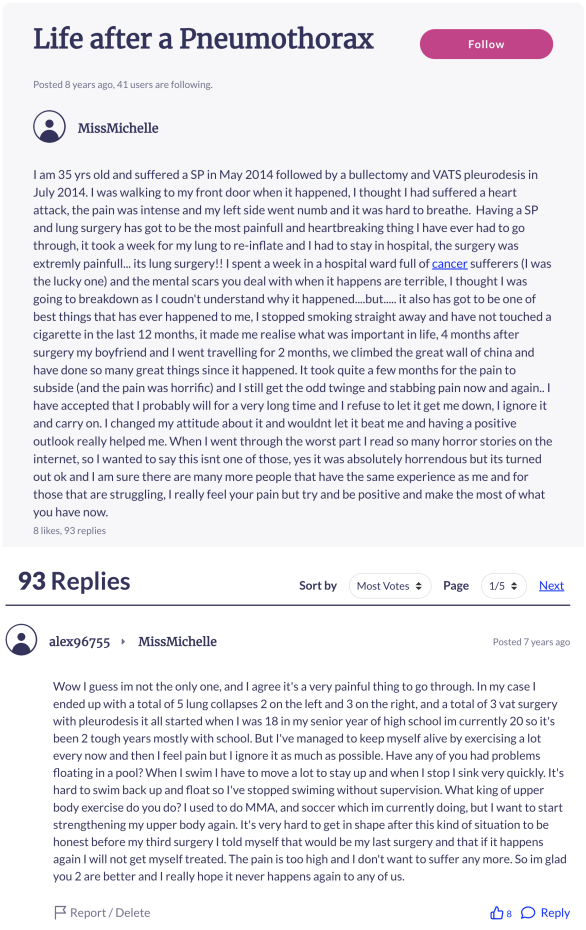

Figure 2: Criteria Based Data Scrapping Sample

4 System and Result Analysis

4.1 System Architecture

The trained Bi-LSTM model was deployed further using a service called Gradio for the interface part and HuggingFace spaces was used for deployment backend runtime.The Gradio Interface serverd as a frontend, where the user would be providing the link of the disease group (G) they want to evaluate the posts for. Once the link is submitted, the Gradio Frontend sends the link to a node backend server via an POST API request, which is captured by the service. Further processing of the link is handled by the node server. Once all the statements (P) & most voted comments (Q) are collected, the entire information is sent back to the gradio frontend as a json object. This json object is then parsed by the Bi-LSTM model and corresponding output labels are appended to each entry. After processing all the entries, the resultant csv is provided to the user for download and end use.

Statement (P)	Comment (Q)	Label of Q	Label Reasoning
I've been prescribed one but not shown how to use it or what normal is. At present, following the written instructions my highest reading is 200 I am 63 female and 5 ft 2 ins	Hi lynne if you contact your surgery you asthma nurse will show you the correct way to measure your peak flow every persons different I use mine to keep a record . It is very usefull to keep track of your asthma , when you use it I always do it 3 times I find you get a true indication that way if you can't get an appointment you could ring asthma uk Im sure they would put you right hope this is some help.	Positive	Support by companionship (self-experience) + Support by information (suggestion) + No Emotional Support
Has anybody used saline in a nebuliser to loosen phlegm to cough it up ?	Good question, I sometimes dream my lungs not with a nebuliser riven though I have one. I sterilize water in a pot for 5 minutes then I add a few drops of ucoliptys oil and then put a towel over my head and breath in the steam for 15 minutes, not to close because it could hurt your lungs. I do this for 15 minutes 2 times a day if I am sick. i thought about adding salt to the water but I have never heard of any good medical advice that it would be helpful. I hope someone comes along that has some practise and more knowledge about it.	Neutral	Support by companionship (self-experience) + No Support by information (suggestion)
I need your help on a problem that I haven't checked with the doctor yet hoping that it's not serious. For last a few years I have a jelly like transparent, sticky mucus in my chest. I never felt any pain thereby, nor does it make me cough, but I do cough or spit that mucus out adverternly. That substance keeps on developing in my lungs, or some air passage. In cold, I rarely feel minute weezing, or congestion, but I can walk 10 miles. Can some one please help me understand what that is, is it normal?,	I have had it for four years. As a singer, my voice started cracking and after doing all the usual like cutting our dairy, got referred to a consultant who said he could find nothing wrong but there was rather more mucus than normal. He couldn't come up with anything that could be causing it nor how to sort it. So now I can't sing, and sometimes makes me cough a lot. I take Symbicort inhaler twice a day .	Negative	Not recovered yet

Figure 3: Output class verification for Asthma

4.2 Results

	precision	recall	f1-score	support
0	0.97	0.88	0.93	259
1	0.91	0.91	0.91	255
2	0.81	0.89	0.85	226
accuracy			0.90	740
macro avg	0.90	0.90	0.90	740
weighted avg	0.90	0.90	0.90	740

Figure 4: Classification Report

The generated output CSV file is manually verified for three cases. Case 1: Fig 2 depicts a sample from "Asthma" and corresponding manual verification confirms the 100% accuracy of fetching target discussions with minimum replies and associated most voted comments. Case 2: Fig 3 represents a sample class label verification of the most voted comment for "Asthma". Case 3: Fig 4 represents verification of the performance of the classification model and witnessed an overall accuracy of 90%. While dealing with NLP tasks on textual data, a True Positive (TP) outcome is highly important, since several decisions in the medical domain is highly sensitive to information. The confusion matrix in Fig. 5 shows an overview of training samples being classified into multiple classes. From the figure, it is evident that most of the data is classified in the correct classes, enhancing the Accuracy. Furthermore, with the increase in epochs, the training and validation accuracy also increased as in Fig. 7. It is also observed that the loss vs accuracy curve in Fig. 6 indicates improvement of the model with epochs.

5 Conclusion

Using this automated system we can successfully scrape, and determine the sentiment of the maximum voted reply against a discussion with satisfactory accuracy in minimum system require-

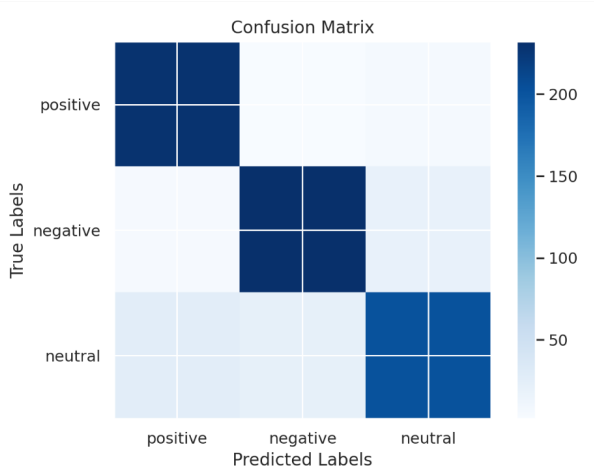

Figure 5: Opinion Confusion Matrix

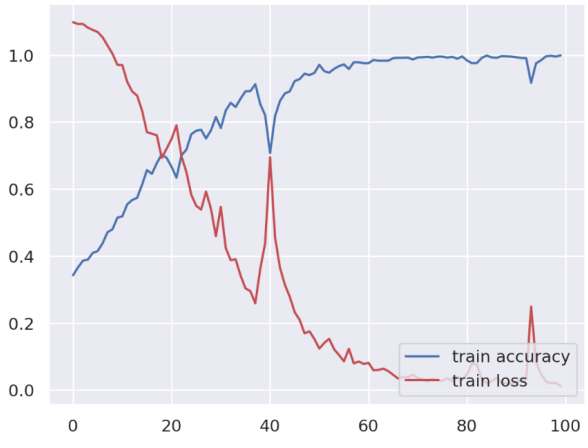

Figure 6: Training Accuracy vs Training Loss

ments. Since a reply post can be made either directly against the original post or reply to another respondent, it will help both the OP and others (visitors and members) to automatically track the most agreed-upon comment and its opinion.

Limitations

Complete automation for all the groups of all the categories is yet to be implemented. An error analysis of nested structuring of the replies during target comment selection needs to be verified. Also, a complete GUI with user query-based filtering for scrapping can be added. The filtered data set was found to be very small for more accurate predictions.

Ethics Statement

We perform our experiments on a dataset created by automatic scrapping and annotating. If any training examples are associated with some slurs, abuses,

Figure 7: Training Accuracy vs Validation Accuracy

and other derogatory terms, it was considered as public opinion only. We urge the forum community to use our application and we are fully committed to providing a more scalable user-friendly complete GUI-based application in the future.

References

Alan Aipe, Mukuntha Narayanan Sundararaman, and Asif Ekbal. 2019. Sentiment-aware recommendation system for healthcare using social media. *arXiv preprint arXiv:1909.08686*.

Tanveer Ali, Marina Sokolova, David Schramm, and Diana Inkpen. 2013. Opinion learning from medical forums. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 18–24.

TK Balaji, Chandra Sekhara Rao Annavarapu, and Annushree Bablani. 2021. Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, 40:100395.

Athira Balakrishnan, Sumam Mary Idicula, and Josette Jones. 2021. Deep learning based analysis of sentiment dynamics in online cancer community forums: An experience. *Health Informatics Journal*, 27(2):14604582211007537.

Seyed Mojtaba Hosseini Bamakan, Ildar Nurgaliev, and Qiang Qu. 2019. Opinion leader detection: A methodological review. *Expert Systems with Applications*, 115:200–222.

Umamageswari Baskaran and Kalpana Ramanujam. 2018. Automated scraping of structured data records from health discussion forums using semantic analysis. *Informatics in Medicine Unlocked*, 10:149–158.

Kaveh Bastani, Hamed Namavari, and Jeffrey Shaffer. 2019. Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. *Expert Systems with Applications*, 127:256–271.

Victoria Bobicev and Marina Sokolova. 2014. Sentiment analysis in health related forums. In *Microelectronics and Computer Science*, pages 213–216.

Victoria Bobicev and Marina Sokolova. 2018. Thumbs up and down: Sentiment analysis of medical online forums. Association for Computational Linguistics.

Heather Buchanan and Neil S Coulson. 2007. Accessing dental anxiety online support groups: An exploratory qualitative study of motives and experiences. *Patient education and counseling*, 66(3):263–269.

Jorge Carrillo-de Albornoz, Ahmet Aker, Emina Kurtic, and Laura Plaza. 2019. Beyond opinion classification: Extracting facts, opinions and experiences from health forums. *PloS one*, 14(1):e0209961.

Annie Chen. 2013. Patient experience in online support forums: Modeling interpersonal interactions and medication use. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 16–22.

Annie T Chen. 2012. Exploring online support spaces: using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups. *Patient education and counseling*, 87(2):250–257.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Eleonora D'Andrea, Pietro Ducange, Alessio Bechini, Alessandro Renda, and Francesco Marcelloni. 2019. Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, 116:209–226.

Behrooz Davazdahemami and Dursun Delen. 2019. The confounding role of common diabetes medications in developing acute renal failure: A data mining approach with emphasis on drug-drug interactions. *Expert Systems with Applications*, 123:168–177.

Elena Davcheva, Martin Adam, and Alexander Benlian. 2019. User dynamics in mental health forums - a sentiment analysis perspective. In *Wirtschaftsinformatik*.

MV Kamalov, VY Dobrynin, YE Balykina, and RS Martynov. 2017. Analysis of user activities on popular medical forums. In *Journal of Physics: Conference Series*, volume 913, page 012007. IOP Publishing.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Gunjan Mansingh, Kweku-Muata Osei-Bryson, and Han Reichgelt. 2009. Issues in knowledge access, retrieval and sharing–case studies in a caribbean health sector. *Expert Systems with Applications*, 36(2):2853–2863.

Mika V Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32.

Thomas Opitz, Jérôme Azé, Sandra Bringay, Cyrille Joutard, Christian Lavergne, and Caroline Mollevi. 2014. Breast cancer and quality of life: medical information extraction from health forums. In *MIE: Medical Informatics Europe*, pages 1070–1074.

Marie E Perrone, David Carmody, Louis H Philipson, and Siri Atma W Greeley. 2015. An online monogenic diabetes discussion group: supporting families and fueling new research. *Translational Research*, 166(5):425–431.

Md Ileas Pramanik, Raymond YK Lau, Md Abul Kalam Azad, Md Sakir Hossain, Md Kamal Hossain Chowdhury, and BK Karmaker. 2020. Healthcare informatics and analytics in big data. *Expert Systems with Applications*, 152:113388.

Eysha Saad, Sadia Din, Ramish Jamil, Furqan Rustam, Arif Mehmood, Imran Ashraf, and Gyu Sang Choi. 2021. Determining the efficiency of drugs under special conditions from users' reviews on healthcare web forums. *IEEE Access*.

Ashnish Sinha, Tom Porter, and Andrew Wilson. 2018. The use of online health forums by patients with chronic cough: qualitative study. *Journal of medical Internet research*, 20(1):e19.

Marina Sokolova and Victoria Bobicev. 2013. What sentiments can be found in medical forums? RANLP.

Marina Sokolova and Victoria Bobicev. 2015. Learning relationship between authors' activity and sentiments: A case study of online medical forums. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 604–610.

Chuan-Jun Su and Chun Wei Peng. 2012. Multiagent ontology-based web 2.0 platform for medical rehabilitation. *Expert Systems with Applications*, 39(12):10311–10323.

Cornelia F van Uden-Kraan, Constance HC Drossaert, Erik Taal, CEI Lebrun, KW Drossaers-Bakker, WM Smit, ER Seydel, and Mart AFJ van de Laar. 2008. Coping with somatic illnesses in online support groups: do the feared disadvantages actually occur? *Computers in human behavior*, 24(2):309–324.

Sophia Y Wang, Tina Hernandez-Boussard, Robert T Chang, and Suzann Pershing. 2019. Understanding patient attitudes toward multifocal intraocular lenses in online medical forums through sentiment analysis. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 1378–1382. IOS Press.

Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2018a. Medical sentiment analysis

using social media: Towards building a patient assisted system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shweta Yadav, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya, and Amit Sheth. 2018b. Multi-task learning framework for mining crowd intelligence towards clinical treatment.

Shweta Yadav, Vishal Pallagani, and Amit Sheth. 2020. Medical knowledge-enriched textual entailment framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1795–1801, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shweta Yadav, Joy Sain, Amit Sheth, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2018c. Leveraging medical sentiment to understand patients health on social media. *arXiv preprint arXiv:1807.11172*.

```
const puppeteer = require("puppeteer");
const ObjectsToCsv = require("objects-to-csv");
let browser;

const csv = require("csv-parser");
const fs = require("fs");
let results = [];
const createCsvWriter = require("csv-writer").createObjectCsvWriter;

exports.collectGroupLinks = async ({link}) => {
  browser = await puppeteer.launch({ headless: false }); // default is true

  const page = await browser.newPage();

  // Navigate to the website
  await page.goto(`${link}`);
    // "https://patient.info/forums/discuss/browse/respiratory-symptoms-and-disorders-2014"
//   );

  const buttonExists = await page.evaluate(() => {
    const button = document.querySelector(".css-11i4sx8");
    return button !== null;
  });

  console.log(`Button exists: ${buttonExists}`);

  await page.click(".css-11i4sx8");

  const selectTags = await page.$$(".submit.reply-pagination");

  let totalOptionCount = 0;
  for (let i = 0; i < selectTags.length; i++) {
    const optionTags = await selectTags[i].$$("option");
    totalOptionCount += optionTags.length;
  }

  console.log(`Total number of page tags: ${totalOptionCount}`);
  var d = [];
  for (let k = 0; k < totalOptionCount; k++) {
    // page count
    const elementExists = await page.evaluate(() => {
      return !!document.querySelector(".cardb__block");
    });

    if (elementExists) {
      const elements = await page.$$(".cardb__block");
      console.log(
        "Found",
```

```
      elements.length,
      'elements with class "cardb__block": Page ',
      k + 1
    );

    for (let i = 0; i < elements.length; i++) {
      //card block count
      const title = await elements[i].$eval(
        "h3.post__title",
        (div) => div.textContent
      );
      const hrefValue = await elements[i].$eval("h3.post__title a", (a) =>
        a.getAttribute("href")
      );
      const actions = await elements[i].$$("div.actions");
      const secondActions = actions[1];
      const content = await secondActions.$eval(
        "span",
        (span) => span.textContent
      );
      const data = {
        problem: title,
        link: hrefValue,
        comments: content,
      };
      d.push(data);
    }

    const nav_btn = await page.$$(".link__text");
    if (nav_btn) {
      const [nextSpan] = await page.$x("//span[contains(text(), 'Next')]");
      if (nextSpan) {
        await nextSpan.click();

        console.log("clicked");
      } else break;
      await page.waitForNavigation();
    }
  } else {
    console.log('Element with class "cardb__block" does not exist');
  }
}

const csv = new ObjectsToCsv(d);
await csv.toDisk("./groupLinks.csv");
console.log("Exported");
// Close the browser
await browser.close();
};
```

```
exports.collectLinkData = async () => {
  const csvWriter = createCsvWriter({
    path: "group_findings_with_results_most_voted.csv",
    header: [
      { id: "#", title: "#" },
      { id: "link", title: "Link" },
      { id: "noOfCom", title: "noOfCom" },
      { id: "title", title: "Title" },
      { id: "statement", title: "Statement" },
      { id: "comments", title: "Comments" },
      { id: "upvotes", title: "Upvotes" },
    ],
  });

  fs.createReadStream("groupLinks.csv")
    .pipe(csv())
    .on("data", (data) => results.push(data))
    .on("end", async () => {
      console.log("Imported");
      console.log(results.length + " items found");
      var d = [];
      for (let i = 0; i < results.length; i++) {
        if (results[i].comments > 1) {
          const browser = await puppeteer.launch({ headless: false }); // default is true
          const page = await browser.newPage();

          await page.goto(
            "https://patient.info" +
              results[i].link +
              "?order=mostvotes#topic-replies",
            { waitUntil: "domcontentloaded" }
          );

          const buttonExists = await page.evaluate(() => {
            const button = document.querySelector(".css-11i4sx8");

            return button !== null;
          });
          if (buttonExists) {
            console.log(`Button exists: ${buttonExists}`);

            page.click(".css-11i4sx8");
          }
          const text = await page.evaluate(() => {
            const h1 = document.querySelector("h1.u-h1.post__title");
            return h1?.innerText;
          });
          const upvotes = await page.evaluate(() => {
```

```
          const span = document.querySelector("span.post__count");
          return span?.innerText;
        });
        const value = await page?.$eval(
          "input.moderation-conent",
          (input) => input?.value
        ); // Post of Original Poster

        const elementValue = await page.evaluate(() => {
          const parentDiv = document.querySelector(
            ".post__content.break-word"
          );
          const inputField = parentDiv.querySelector(
            "input.moderation-conent"
          );
          return inputField?.value;
        });

        const data = [
          {
            "#": i + 1,
            link: results[i].link,
            noOfCom: results[i].comments,
            title: text,
            statement: value,
            comments: elementValue,
            upvotes: upvotes,
          },
        ];

        await csvWriter
          .writeRecords(data)
          .then(() => console.log("CSV file updated successfully"))
          .catch((err) => console.error(err));

        console.log("done");
        await browser.close();
      }
    }
  });
};


const { collectGroupLinks, collectLinkData } = require("./collectData")

exports.startCollection = async(req,res)=>{
   const {link} = req.body
  await collectGroupLinks({link})
  await collectLinkData()
```

```
  res.send({code:200,msg:"Complete"})
}

const express = require("express");
const { startCollection } = require("../controllers/startCollect");
const dataRouter = express.Router();

dataRouter.post("/collect-from-url", startCollection);

module.exports = dataRouter;

const express = require('express')
const cors = require('cors')
const { readdirSync } = require("fs");
const dotenv = require('dotenv').config();
const path = require("path")
const app = express();
app.use(express.json());

app.use(cors({ origin: true }));
app.use(function (req, res, next) {
   res.header("Access-Control-Allow-Origin", "*");
   res.header("Access-Control-Allow-Headers", "Origin, X-Requested-With, Content-Type, Accept");
   next();
});


app.use("/", require("./routes/collectData"))
app.listen(5001, () => {
   console.log(`server is running on port 5001..`);
});
```