M.Com: A Call for Enhanced Medical Complaint Analysis Model for Indian Healthcare Sectors

Abstract. In the aftermath of the global COVID-19 global pandemic, healthcare systems face significant challenges in restoring service quality. This highlights the need to address key sectors like delays in treatment, physician inattentive nature, lack of insurance and hygiene issue. The microblogging nature of social media airs these concerns through online posts and blogs (text). A single medical concern might have multiple aspects (e.g. negligence, hygiene, billing, etc.), and in densely populated countries like India, effective mechanisms to identify and address medical complaints are paramount for restoring patient care. Past research primarily identified complaints but lacked datasets to understand underlying issues. Motivated by this research gap, we introduce a novel and comprehensive first-of-its-kind unimodal medical complaint dataset known as *M.Com*. This dataset comprises nine distinct aspect categories, providing robust solutions through a multilabel classification task, showcasing a nuanced comprehension of medical complaints across over 7000 sample instances. Utilizing cutting-edge Large Language Models (LLMs) like Mistral, Gemini, and Zephyr, we conducted preliminary experiments on medical aspect classification and complaint identification tasks on our dataset, exploring both fine-tuning and zero-shot learning settings. Our research findings demonstrate significant performance disparities between Bert-configured models and LLMs in both aspect classification and complaint identification tasks, achieving an accuracy of 89% in a few-shot setting. This unveils new avenues for exploration without the need for extensive training settings, making notable contributions to the broader research community.

1 Introduction

Complaints in healthcare sectors are frequently viewed through a critical lens, reflecting on the quality of patient care. Complaints due to negligence have risen by 30-40% over the past five years, but less than 10% of doctors are held accountable [1]. Patients and their families represent a valuable repository of insights that can significantly contribute to elevating the standards of care. medical institutions frequently encounter a spectrum of complaints ranging from negligence, hygiene issues, incorrect treatment, and unwarranted charges to diminishing patient satisfaction and waning public trust. Amidst this expansive landscape of user feedback, distinguishing between coherent and rational reviews underscores the dissonance between expectations and actual experiences [26]. Most complaints originate from patients, who engage with various medicolegal entities like medical boards or compensation authorities for diverse reasons. These may range from seeking financial compensation to requesting a review of a doctor's fitness to practice, or simply desiring an apology or clarification [4, 3]. A recent study by Denecke et al. [8] conducted a preliminary investigation to capture medical sentiment from both clinical narratives and medical social media sources.

Previous research in complaint mining predominantly centered on unimodal review complaints [15, 27]. The microblogging nature of social media provides an analytical way to scrutinise and understand public psychology [10]. Several shared task challenges [20, 21] have been conducted to explore social media texts for capturing user opinions within medical contexts. Understanding the fundamental reasons why people resort to human surveys, which are challenging to conduct on a large scale, poses a significant challenge.

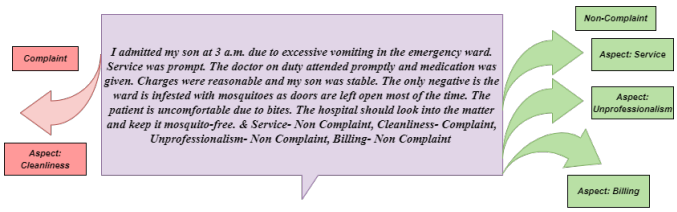

Figure 1. An instance of Aspect Based Medical Complaint identification task

For instance, as depicted in Figure 1, a patient has provided feedback on their experience in the hospital's emergency ward. While the feedback is predominantly positive, the patient does not explicitly address concerns regarding cleanliness, despite mentioning issues such as mosquitoes biting due to open doors. However, such implicit aspects of complaints shed light on areas for improvement, facilitating a detailed analysis to enhance patient healthcare initiatives. ***Motivation***: Numerous datasets have been investigated regarding various medical sentiments. Yet, thus far, there are no datasets exclusively dedicated to identifying and alerting medical institutions about service-related complaints with multilabel aspects, as illustrated in Figure 1. This study presents *M.Com*, a large-scale generalized dataset sourced from online social media blogs, encompassing nine distinct aspects related to medical service contexts. Furthermore, in order to assess the capabilities of generative AI and LLMs, we conducted experiments utilizing various LLMs and transformed the task into a multilabel classification task.

This study aims to answer two major questions:

1. Does the proposed dataset exhibit generalizability for deep learning methodologies and large language models?

2. How is society deriving benefit from this research endeavour?

What does the paper add?

- Introducing a unique medical complaint dataset known as *M.Com*, which comprises over 7000 medical grievances reported across social media platforms. This dataset encompasses nine distinct aspects for analysis and evaluation.
- Moreover, we conducted fine-tuning on BERT Based models and well-known Large Language Models like Zephyr, Mistral and Deep Seek under various conditions to carry out aspect-based complaint identification tasks. This paper outlines the prospective advancements of Large Language Models (LLMs) to enhance healthcare service initiatives and provide assistance to a broad spectrum of individuals in need.

2 Literature Review

Prior research in the medical sector has endeavoured to identify various types of concerns within complaints, emphasizing the significance of our dataset. Moreover, these studies have explored the multilabel classification task to address nuanced aspects of complaints in healthcare settings. In recent years, there has been a substantial increase in research focused on discerning and aggregating subjective or non-factual textual expressions, which encapsulate individuals' opinions, sentiments, or emotions, particularly within medical blog texts.

- **Medical Concern Related Task:** In 2015, a review was presented on the emerging natural language processing technique to analysis social media blogs and post for real-world clinical tasks [12]. Glen et al. demonstrated the feasibility of using social media data to detect those at risk for suicide [5]. In 2020, O.Baclic et al. explored how data analysis from diverse sources bolsters public health functions, improving surveillance, challenges, prevention strategies, and health promotion with expert insights [2]. Demener et al. provided a review of recent developments in clinical and consumer-generated text processing [7] by highlighting comprehensive discussions on disease modelling, predictive analytics using clinical texts, social media text analysis for healthcare quality assessment, trends in online interventions, and consumer health question answering. Furthermore, the rapid expansion of medical context datasets has leaned on social media sources, including medical weblog-based datasets [9], medical entity-based datasets [30], and sentiment-aware medical health mining datasets [34].
- **Complaint Identification:** The complaint's nature is implicit, lacking explicit blame attribution [28]. In the field of pragmatics, the work by [25] pioneered the classification of complaints into five distinct types: a) Beyond Reproach, b) Explicit Complaint, c) Statement of Disapproval, d) Warning, and e) Allegation. However, Trosborg et al.'s foundational study [32] identified four primary levels of complaint severity: a) Implicit Reproach Absent, b) Disapproval, c) Accusation, and d) Blame. Gradually, Transformers-based compliant categorization based on their severity criteria was developed by [15]. In other domains, such as finance and e-commerce, complaint-based datasets such [6, 31] were developed.
- **Multi-Label Classification:** Multi-label text classification presents a classic machine learning challenge, with various datasets available across diverse domains. Notably, the Reuters dataset [19] features news articles categorized into distinct topics, while the MIMIC-III dataset [16] comprises extensive medical records annotated with multiple ICD-9 codes. Furthermore, datasets like SemEval 2018 Task 1 [24] encompassing emotion subclasses in tweets , in e-commerce 30PT dataset with 38 unique attributes [18], and TREC-IS [22] containing categories for disaster-related tweets contribute to multi-label classification research.
  However, within the medical domain, there is a notable absence of specific or relevant datasets tailored to our research initiative focusing on medical complaints.

3 CORPUS INFORMATION

To the best of our understanding, the task of creating a medical complaint-mining dataset for medical sector working initiatives domains is significant. In this study, we created a unique *M.Com* dataset encompassing medical multiple-aspect Complaints. This dataset spans nine diverse domains, *billings, unavailability, inefficiency, unprofessionalism, pharmacy, cleanliness, service, behaviour and negligence* with a total of 7669 samples. Table 1 represents the sample instances in *M.Com* corpus. The detailed dataset creation procedure is represented in Fig 4.

3.0.1 Selection of Reviews:

To compile a comprehensive dataset, we embarked on a meticulous journey of data collection from diverse online platforms. Leveraging tools such as BeautifulSoup and Selenium, we scoured various social media platforms, including Twitter, Reddit, and numerous healthcare-centric forums such as MouthShut, Voxya, Indian Consumer Forum, and India Complaints. This exhaustive approach ensured the acquisition of a diverse range of data sources, yielding a high-quality dataset for our analyses. Table 2 displays several significant keywords employed during the dataset collection process.

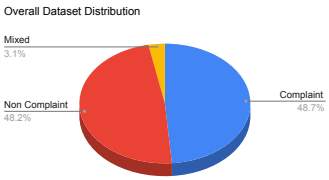

Figure 2. Overall Compalaint and Non Complaint Distribution in *M.Com* Corpus

3.0.2 Sanitization of the Corpus:

Upon gathering the requisite data, the next crucial step involved preprocessing to ensure its cleanliness and relevance. Given the potential presence of irrelevant content, it was imperative to validate the data. This entailed the removal of extraneous reviews that did not contribute to the research objectives. Furthermore, the scraped data often required cleaning to enhance its usability. Employing various strategies, including removing NaN values, deduplicating tweets, consolidating data from multiple sources, and eliminating HTML symbols and special characters, we meticulously refined the dataset. Additionally, any missing values were addressed to ensure the integrity and completeness of the data. This rigorous preprocessing phase laid the groundwork for subsequent analyses, ensuring that our dataset was primed for effective binary and multiclass classification tasks. Figure 2 depicts the complaint class statistics of our final *M.Com* corpus.

Table 1. Instances of *M.Com* Dataset with corresponding Aspects with Complaint and Non-Complaint labels; Highlighted sections in red and blue indicates Complaint and Non-complaint sections

Sample	Aspects
I admitted my son at 3 a.m. due to excessive vomiting in the emergency ward. Service was prompt. The doctor on duty attended promptly and medication was given. Charges were reasonable and my son was stable. The only negative is the ward is infested with mosquitoes as doors are left open most of the time. The patient is uncomfortable due to bites. The hospital should look into the matter and keep it mosquito-free.	Service - Non Complaint, Cleanliness - Complaint, Unprofessionalism - Non Complaint, Billing - Non Complaint
My mother admitted in SICU in apollo OMR hospital. I like the services provided by the ICU team doctors and nurses and other supporting staffs. Due to their service my mother recovered at the earliest.	Service - Non Complaint, Behavior - Non Complaint
Useless and unfit nursing staff..and non medical staff they have taken one whole day to prepare one discharge summary these nursing staff come to work only for time pass..these r not interested in working they r there only for money.	Service - Complaint, Inefficiency -Complaint, Unprofessionalism - Complaint

Table 2. Useful Key-terms during medical complaint sample collections

Generic Key Terms	side-effects, medical errors, patient safety, issues, doctor complaints, medication errors, wait times, misdiagnosis, healthcare, patient feedback, medical malpractice, unnecessary, treatments, health insurance problems, customer service, Pharmacy Complaint
Aspect Specific Key Terms	billing, negligence, behavior, service quality, time efficiency, resource shortages, accusations, disputes, cleanliness standards, online services, pharmacy operations, medicine expiry, medical negligence occurrence, improper diagnoses, medicine delivery, incorrect treatments, unjustified billing.

3.0.3 Selection of Aspects for Complaints

An integral phase in our research methodology involved carefully selecting the aspect categories for classification within our dataset. To achieve this, we embarked on a systematic approach that began with an initial exploration of 500 tweets. This exploratory phase aimed to identify various themes or classes prevalent in the corpus of tweets. Upon thorough examination, we uncovered a plethora of distinct categories, numbering approximately 40. However, to ensure practicality and simplicity in our classification scheme, we recognized the necessity of consolidating these categories into a more manageable set. To accomplish this task, we deliberated on the overarching themes encapsulated within each category, aiming to generalize their names to encompass broader concepts. Through this process of abstraction, we distilled the original 40 classes into a more concise set of 10 classes. This transformation involved synthesizing similar categories and refining their descriptors to capture the essence of each class in a more generalized manner. For instance, individual categories about specific aspects of complaints, such as service quality, cleanliness, and behavior, were amalgamated into broader categories reflecting overarching themes. By adopting this approach, we ensured that the resulting classes retained their relevance and comprehensibility while adhering to the constraints of using general categories. Figure 3 describes the aspect-wise sample statistics.

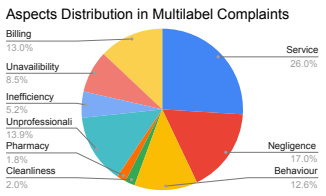

Figure 3. Statistics of various Aspects in *M.Com* dataset

3.0.4 Corpus Annotation

For annotation purposes, we have hired nine interns. We asked the firm to get each tweet annotated independently by three annotators. The annotators were a set of university graduates aged 20-25 who are well-versed in English and conversant with the review. The annotators were tasked with immersing themselves in medical concerns-related topics for a duration of two weeks, ensuring they acquired sufficient knowledge in this domain. To mitigate potential biases, the annotators were kept unaware of each other's identities and were instructed to carry out their annotations in isolation (Annotators were paid $0.5 per sample). Each review was labelled by three annotators from this set given by the firm. Throughout the annotation phase, the annotators adhered to the standard guidelines outlined in Table 3. We

Table 3. Standard guidelines for annotation task

S.No.	Annotation Guidelines
1	Annotations must be conducted without external influence, thoroughly studying the referenced samples.
2	Verify the text instance and find hidden aspects if present.
3	For each sample, a maximum of 5 and a minimum of one aspect can be chosen. Each aspect must have a complaint and non-complaint label.
4	Any bias, ambiguous symbols and erroneous labels observed during annotations must rectify.

first provided a set of 1,000 randomly sampled tweets (from the set of tweets obtained after duplicate removal, as described in Section 3.1) to the firm to be annotated independently by three annotators. After a week, the firm completed the annotations, and we cross-checked the annotations for 100 of these tweets to ensure correctness. Other than a few minor label corrections, the annotations looked good. We discussed with the annotators and clarified some of their doubts regarding the class descriptions. We also clarified that explanations (for a particular label) within a particular tweet can also consist of non-contiguous words. Ensuring consensus among multiple annotators is crucial when working on annotation tasks involving two or more annotators while creating a reliable annotated dataset. Fleiss' kappa score [11], a standard metric commonly used for this purpose, was employed to evaluate the agreement between the three annotators in our study. The computed Fleiss' kappa scores for the aspect category and aspect-level annotation tasks were 0.67, and 0.84, respectively. These scores indicate a high level of agreement among the annotators, suggesting that the annotations are of benchmark quality.

4 Methodology

This section extensively explores the problem formulation and establishes a clear overview of the research work. Following this, subse-

quent sections analyze the key components of the applied methods, offering insights into why its implementation is justified.

4.1 A. Problem Statement

We conceptualize the Healthcare binary and aspect-based complaint identification model as a generative function that transforms input reviews into fundamental feature pairs, encompassing aspect categories, target complaint-non-complaint labels, and causal expressions. The framework comprises two sequential stages: Aspect-based complaint classification (AC) and Complaint Identification (CI).

4.2 Phase-1 Aspect-based Complaint Classification (AC)

In the initial phase of our analysis, we conducted an aspect-based classification. After reviewing 500 customer feedback entries, we identified over 35 distinct classes. These were subsequently consolidated into 9 broader categories to streamline the analysis. Each review was then classified according to these refined aspects.

For aspect-based classification in natural language processing (NLP), the objective is to identify the relevant aspect or sentiment expressed in a given text. Mathematically, for a single text T_i sample given an input X and an aspect A, the prediction of the aspect for this sample can be represented as follows: For a single text T_i, with input X and aspect A, the prediction for that single sample can be represented as:

$$P(A|X) = f(T_i)$$

Where $f(T_i)$ denotes the function that maps the text T_i to its aspect A. For n samples, the prediction can be generalized as:

$$P(A_i|X_i) = f(T_i) \quad \text{for } i = 1, 2, \ldots, n$$

Here, T_i represents the i-th text sample, X_i denotes the input, A_i is the aspect for the i-th sample, and $f(T_i)$ is the function determining the aspect based on the i-th text.

In this study, we utilized and fine-tuned several language models, including BERT and DistilBERT. These models feature a layer that accepts 768 inputs and produces 9 outputs corresponding to the nine aspects, with the selection based on the highest probability. Furthermore, we experimented with Mistral Instruct, a generative large language model (LLM) that is capable of directly producing aspect predictions. This model utilizes zero-shot and few-shot learning techniques, allowing it to effectively predict labels even with minimal task-specific tuning. The integration of these models has yielded promising results in accurately classifying the text into predefined aspects.

Zero-shot Classification Prompt Instruction: *You are a text classifier for the healthcare domain, you are an expert in classify the text. Classify the text given to you in one or more most relevant Aspect classes, but not more than 5 classes, your response must have most relevant Aspect classes in a list using comma as a delimiter, do not give any explanation or any note, only return the python list containing relevant Aspect classes.*
Aspect Classes: ['Service', 'Negligence', 'Behaviour', 'Cleanliness', 'Pharmacy', 'Unprofessionalism', 'Inefficiency', 'Unavailability', 'Billing'] Find Aspect classes for below text reviews

Few-shot Classification Prompt: Instruction: *Choose one or more labels among the following possibilities with the highest probability. Only return the labels in a python list, nothing more.*
Aspect Classes: ['Service', 'Negligence', 'Behaviour', 'Cleanliness', 'Pharmacy', 'Unprofessionalism', 'Inefficiency', 'Unavailability', 'Billing'] Find Aspect classes for below text [Review]

4.3 Phase-2: Complaint Identification (CI)

After identifying relevant aspects, the subsequent task involves classifying each review as either a complaint or a non-compliant. For fine-tuning BERT, we have incorporated an additional linear layer ANN that accepts 768 inputs and produces 27 outputs, which we have reshaped into a 9x3 matrix. The first column indicates whether the aspect belongs to the review, the second column determines if it is a complaint, and the third column identifies if it falls into the non-complaint category. This configuration enables BERT to output results whenever a review is processed through the model. In the case of other large language models, such as the Mistral Model, the system generates aspect-based complaints or non-complaint classifications

We have utilized various models for binary classification, employing both fine-tuning and zero-shot/few-shot techniques. Specifically, we fine-tuned BERT-Base-uncased and DistilBERT for binary classification, adapting them to align with our specific dataset and task requirements more closely. We applied zero-shot and few-shot techniques for other models, including Zephyr 7B alpha, Google Gemini, Mistral 7B, Mistral Instruct, Bio Mistral, and Deep Seek. Among these models, the Mistral models, particularly Mistral 7B, have demonstrated superior performance.

5 Results and Experiments

In this section, we shall discuss a task that can be directly applied to the dataset – Complaint Identification. We also apply several state-of-the-art methods for bench-marking each of the tasks. In the standard Complaint Identification task, each data point (reviews in our case) has to be assigned to one class out of Complaint or Non-Complaint. Then evaluated the performance of various machine learning models across different prompts and conditions. The goal was to identify the most effective models for natural language understanding of healthcare-related review tasks. The experiment involved three categories of models: zero-shot, few-shots, and fine-tuning models. Our research aims to investigate the following research questions:

RQ1: How do BERT-based models perform compared to current LLMs in classification tasks using the *M.Com* dataset?
RQ2: How do models' performance metrics vary across zero-shot, few-shot, and fine-tuning scenarios for dataset classification?
R3: Discuss the broad societal applicability of *M.Com*

5.1 Baseline Introductions:

- Bert-Based Uncased [17]: Bert-Base Uncased, a bidirectional encoder, is a pre-trained autoencoder language model developed on English Wikipedia and BookCorpus. It employs a sequence length of 512 and comprises approximately 110 million parameters.
- DistilBert [29]: DistilBERT is a compact Transformer model derived from the BERT architecture, designed for efficiency. During its pre-training phase, knowledge distillation is applied to achieve a 40% reduction in model size compared to BERT.
- Mistral [14]: Mistral is an LLM designed for enhanced efficiency and performance, utilizing grouped-query attention (GQA) to accelerate inference. It also employs sliding window attention

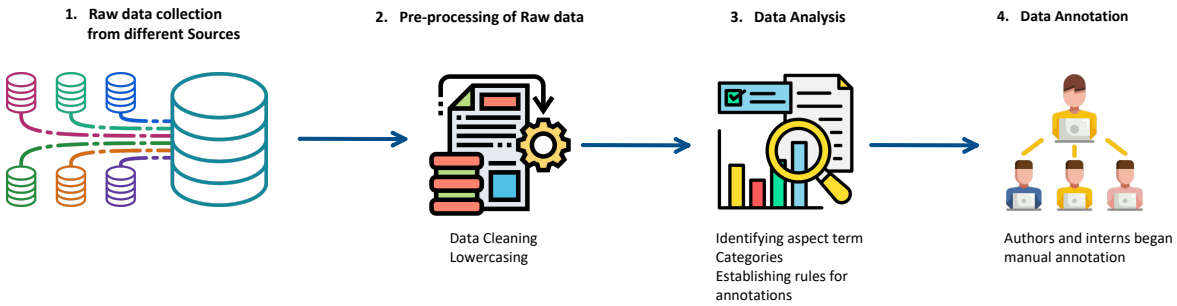

Figure 4. The intricate protocol of dataset creation phase

(SWA) to efficiently manage sequences of varying lengths, achieving this with reduced inference costs. This model comprises 7 billion parameters. We also tested the dataset on the different variants of Mistral, such as Bio Mistral and Mistral Instruct.

- Zephyr [33]: Zephyr-7B-, the second model in the Zephyr series, is a fine-tuned variant of the Mistral model, boasting 7 billion parameters, designed to function as a helpful assistant
- Deep Seek [13]: The DeepSeek Coder consists of code language models, individually trained on 2 trillion tokens, comprising 87% code and 13% natural language in English and Chinese.
- Google Gemini [23]: Gemini, developed by Google AI, is a multimodal large language model facilitating direct user interaction for tasks such as writing assistance, planning, and learning.

5.2 *Experimental Setup*

We assessed the performance of large language models (LLMs) in both zero-shot and few-shot settings. For the few-shot evaluation, we provided three illustrative examples, which included both complaint and non-complaint instances. These evaluations were conducted using a Google Colab notebook equipped with a 15 GB GPU, 12.7 GB of RAM, and 78.2 GB of disk storage. Additionally, we fine-tuned BERT and DistilBERT models on a comprehensive dataset comprising 7,000 samples. This approach allowed us to directly compare the effectiveness of fine-tuned models against baseline LLMs in handling structured text classification tasks. The fine-tuning process was tailored to optimize model performance while adhering to the computational constraints imposed by the available hardware resources.

Experimental Challenges: While generating the dataset, it was challenging to find non-complaint reviews on social media sites, as negative feedback predominates in the medical domain. Additionally, we encountered difficulties with multimodal data integration, as the images available often did not convey relevant information, leading us to focus on a unimodal approach. We also faced issues with reviews of excessive length, which are problematic for models like BERT that are not equipped to handle overly long texts. Consequently, such reviews had to be excluded. Another significant challenge was selecting the most pertinent aspects from the numerous possibilities, necessitating a focus on the most relevant aspects for our analysis.

5.3 *Analytical Discussion*

In this segment, we've responded to the aforementioned research inquiries, offering comprehensive analytical justifications for our experimental selections.

Answer to RQ1: How do BERT-based models perform compared to current LLMs in classification tasks using the MCom dataset? Upon fine-tuning two prominent models, DistilBERT and BERT-Base-uncased, with varying numbers of epochs, DistilBERT emerged as the superior choice in Aspect-Based Classification in Table 4 and Complaint Identification in Table 5. For complaint Identification, it demonstrated an accuracy of 82.95% and an F1 score of 82.95%. DistilBERT showcased a commendable performance. However, its efficacy paled in comparison to the remarkable results achieved by the Mistral Instruct LLM model in the few-shot learning paradigm.

Answer to RQ2: How do models' performance metrics vary across zero-shot, few-shot, and fine-tuning scenarios for dataset classification?
In the context of zero-shot learning, Google Gemini emerged as the top-performing model for complaint Identification, achieving an impressive accuracy of 79.11% and a F1 score of 81.72%, surpassing its counterparts. Transitioning to few-shot learning scenarios, our investigation encompassed four distinct models: Mistral Instruct, Mistral, Bio Mistral, and Deep Seek. Notably, Mistral Instruct exhibited exceptional performance, boasting an accuracy of 89.09% and an F1 score of 96.1%, establishing its supremacy among the examined models for complaint Identification and it also performed far better than all other LLMs in Aspect based classification, however fine tuned DistilBERT outperformed all other.

Answer to R3: Discuss the broad societal applicability of *M.Com*
The societal impact of *M.Com* is substantial, spanning multiple sectors. In healthcare, its skill in identifying medical concerns promises advancements in patient care and tailored treatment plans. In customer service and the medical field, its effectiveness in identifying complaints improves service quality and builds brand loyalty. Furthermore, *M.Com* supports social media monitoring, legal analysis, and public opinion tracking by providing actionable insights from large textual datasets. Its ability for multilabel aspect classification allows for detailed analysis, aiding informed decision-making and strategic planning across diverse contexts. Utilizing *M.Com*'s capabilities enhances decision-making, promotes patient care initiatives, and boosts efficiency across various applications.

Qualitative Analysis: In our study, we investigated the performance of several models under different conditions for complaint identification. Specifically, we evaluated three models, Google Gemini, Mistral 7B, and Zephyr 7B alpha, under zero-shot conditions. Additionally, we assessed five models—Mistral, Mistral Instruct, Bio Mistral, Deep Seek, and Zephyr 7B alpha—utilizing few-shot prompting techniques. Furthermore, using our dataset, we conducted fine-tuning experiments on two BERT-based models: DistilBERT and BERT-Base-uncased.

Table 4. Comparative Results of BERT-Base-uncased (Fine-tuned), DistilBERT (Fine-tuned), and Mistral Instruct (Few Shot) on Aspect-Based Classification (AC);All of the metrics values are in %, and we highlighted the maximum achieved scores in bold. The † denotes statistically significant findings where $p < 0.05$ at 5% significant level.

Aspects	Model	Accuracy	Precision	Recall	F1 Score
Service	**BERT-Base-uncased**	**89.16**†	90.21	93.23	90.99
	DistilBERT	88.42	**91.21**†	**93.43**†	**92.30**†
	Mistral Instruct	68.11	84.36	71.46	77.37
Negligence	**BERT-Base-uncased**	**67.73**†	62.16	**57.23**†	59.59
	DistilBERT	**65.11**†	64.00	56.36	59.94
	Mistral Instruct	63.75	**76.23**†	55.95	**64.53**†
Behaviour	**BERT-Base-uncased**	**71.42**†	**73.86**†	72.13	**72.98**†
	DistilBERT	71.18	66.67	**77.53**†	71.10
	Mistral Instruct	64.06	53.00	67.52	59.38
Cleanliness	**BERT-Base-uncased**	67.19	**60.23**†	62.26	61.22
	DistilBERT	**67.31**†	57.20	59.43	58.29
	Mistral Instruct	52.21	55.98	**79.16**†	**65.58**†
Pharmacy	**BERT-Base-uncased**	**73.20**†	61.44	**64.32**†	62.84
	DistilBERT	72.85	**63.21**†	63.45	63.32
	Mistral Instruct	51.63	59.00	72.91	**65.22**†
Unprofessionalism	**BERT-Base-uncased**	**71.42**†	73.52	**75.93**†	**74.91**†
	DistilBERT	70.19	**76.69**†	73.06	74.83
	Mistral Instruct	62.23	71.64	55.36	62.45
Inefficiency	**BERT-Base-uncased**	**67.34**†	**62.76**†	60.77	61.74
	DistilBERT	66.01	59.04	57.88	58.45
	Mistral Instruct	54.41	57.94	**80.50**†	**67.38**†
Unavailibility	**BERT-Base-uncased**	**69.98**†	62.65	61.90	62.27
	DistilBERT	68.17	59.20	61.43	60.29
	Mistral Instruct	62.66	**79.30**†	**68.16**†	**73.31**†
Billing	**BERT-Base-uncased**	83.16	74.55	76.87	75.69
	DistilBERT	**87.31**†	72.22	**81.25**†	76.47
	Mistral Instruct	67.40	**82.16**†	71.60	**76.52**†

Table 5. Comparative Results of BERT-Base-uncased, DistilBERT, and LLMs on Complaint Identification (CI). All of the metrics values are in %, and we highlighted the maximum achieved scores in bold. The † denotes statistically significant findings where $p < 0.05$ at 5% significant level.

Model	Condition	Accuracy	Precision	Recall	F1 Score
BERT-Base-uncased	Fine Tune	80.11	80.61	83.15	80.00
DistilBERT	Fine Tune	82.95	84.21	81.30	82.79
Google Gemini	Zero Shot	79.11	75.31	89.33	81.70
Zephyr 7B alpha	Zero Shot	74.18	67.77	96.52	79.60
	Few Shot	84.41	82.14	85.39	83.70
Mistral 7B	Zero Shot	60.30	57.14	96.27	71.70
	Few Shot	86.87	83.99	87.56	85.70
Mistral Instruct	Few Shot	**89.09**†	82.51	96.10	88.80
Bio Mistral	Few Shot	85.25	75.88	98.65	85.70
Deep Seek	Few Shot	85.14	83.59	83.31	83.40

On aspect-based classification, we employed Mistral Instruct under a few-shot learning condition. We further fine-tuned two state-of-the-art transformer models, namely BERT-Base-Uncased and DistilBERT, to enhance the performance of the classification task. These models were evaluated based on key metrics including Accuracy, Precision, Recall, and F1 score. In our comprehensive evaluation of various models using the *M.Com* dataset, we observed that while all models demonstrated satisfactory performance, they occasionally exhibited limitations. The detailed results of our evaluation are presented in Table 6. Mistral Instruct demonstrates a superior ability to identify complaint aspects within reviews compared to Deepseek and Zephyr. While Deepseek and Zephyr prioritize non-complaint aspects, Mistral Instruct effectively identifies phrases indicative of dissatisfaction, such as *"focusing on money"* and *" make money on dead body"* highlighted in red color.

Error Analysis In our analysis of misclassified reviews, we identified specific challenges that models faced in accurately classifying complaints. Notably, instances where reviews lacked explicit complaint indicators posed difficulties for the models. These cases required a deeper contextual understanding to determine whether the review truly represented a complaint.

A significant issue arose with the models' ability to interpret sarcastic reviews correctly. These sarcastic reviews' nuances and intended meanings posed challenges, leading to misinterpretations and incorrect classifications. Specifically, Gemini, Mistral, and Mistral Instruct exhibited similar shortcomings in detecting underlying sarcasm, resulting in misclassifications. The models misinterpreted phrases such as *"money minting press"* and *"have perfect recipe for making good balance sheet"* as non-complaints. However, in second review provided in Table 8, Gemini accurately identified the context of *"past trouble caused by the earlier hospital"* classifying the review correctly as a non-compliant, while Mistral and Mistral Instruct failed to do so.

Human Evaluation In our additional quality assessment phase mentioned in Table 7, we engaged human annotators to evaluate the performance of different LLMs, particularly those that previously ex-

Table 6. Qualitative Analysis between best performing models

Reviews	Mistral Instruct	Deepseek	Zephyr	Actual
I am going to discuss my experienced in batra hospital. Well it has good faculty and well experienced staff. They treat patient well. But I think so this can be done by every private hospitals. But they are only **focusing on money** even after death of patient. They still **make money on dead body** and can **earn even more after death of a person**, rest staff is fine.	Complaint	Non Complaint	Complaint	Complaint
I recently had a **disappointing experience** at the hospital's radiology department. It seems that their prioritization system is **not based on a first-come, first-served basis**.	Complaint	Non Complaint	Non Complaint	Complaint

Table 7. Performance Comparison of CI task between Human Evaluator and Popular LLMs; Sections highlighted in blue and red indicate Non-Complaint and Compliant

Reviews	Mistral Instruct	DistilBERT	BERT Base-uncased	Human Annotator
Highly **irresponsible, negligent doctors and staff**. Create panic situation to trick emotional relatives and friends to **fill their pocket**. Have made corrupt rules to **earn money**, Overly hyped.	Complaint	Complaint	Complaint	Complaint
Doctors are **very friendly** towards the patients.All the **staffs cordinate very well**.Hospital was **clean and tidy**.And all of them was so helpful in our need.	Non Complaint	Non Complaint	Non Complaint	Non Complaint

Table 8. Error-Analysis: Reviews wrongly classified by LLM Models

Reviews	Mistral Instruct	Mistral	Gemini	Actual Label
Its **money minting press** and they ***have perfect recipe for making good balance sheet*** you can have ***good case study of great business***. There lack of coordination among staff and departments you have to bridge this gap if you want to get treated. ***They help you to come closer with your patient both physically and emotionally because after all you are the only responsible for all his/her needs in hospital***. Surprises are part of your stay you will get used to soon.	Non Complaint	Non Complaint	Non Complaint	Complaint
I had frequent vomitings after consuming food. Because of this problem, I was unable to eat or digest anything. A friend referred me sri ramakrishna hospital. Dr.V.Arulselvan helped me on giving good treatment, now am good without the ***past trouble caused by the earlier hospital and the bad treatment they provided***.	Complaint	Complaint	Non Complaint	Non Complaint

hibited superior performance in the AC and CI tasks. Notably, in the analysis of the first sample within the dataset, Mistral-Instruct consistently identified the correct label, whereas BERT-Base Uncased misclassified it as non-complaint, aligning with its comparatively lower performance as indicated in Table 5. It is noteworthy that human annotators established the gold label annotations for the *M.Com* dataset. Remarkably, in this specific instance, Mistral-Instruct's predictions mirrored those of the human annotators, further underscoring its efficacy in accurately discerning labels.

6 Conclusion

In our research paper, we introduced *M.Com*, a comprehensive dataset sourced from online medical blogs, focused on aspect-based medical complaints, which hold significant societal relevance. *M.Com* encompasses samples that capture individuals' concerns within the healthcare sector, each annotated with multilabel aspect classifications. We conducted experiments utilizing various large language models, including DistilBERT, Mistral, and Zephyr, to reframe the classification task as a generative one. This dataset boasts generalization potential across different tasks. Furthermore, we presented benchmark results and positioned this tailored task as a proactive step towards enhancing patient care initiatives. For instance, models could potentially leverage meta-information such as customer names and their corresponding tweets to enhance predictive accuracy. To the best of our understanding, this dataset represents a pioneering corpus within the medical domain, poised to benefit both the research community and society at large, thereby laying the groundwork for future advancements in the field.

Ethical Statement This article incorporates diverse data inputs from various sources. The mentioned medical organisations or sector names are used solely for research community reference and not for commercial promotion. The authors refrain from expressing personal preferences in this paper.

7 Limitations & Future Endeveours

The current model is designed to process text data only; it cannot handle multimodal data such as Images, PDF files, or other video formats. There's a limitation regarding the specificity of recognized aspects. If an aspect is rare or was not included during the training phase, the model may fail to identify it, resulting in no aspect prediction. Additionally, the model has linguistic restrictions. Although it has been trained primarily on datasets comprising English and occasionally Hindi texts from Indian hospitals, its performance significantly deteriorates with other Indian languages such as Telugu, Tamil, Marathi, Gujarati, and Kannada. Moreover, the model is not equipped to handle foreign languages effectively. ***Future Work:*** We're advancing our multilabel classification approach by integrating causal explanations, which categorize review sentences to elaborate on their aspects, enabling more comprehensive analyses. Additionally, we're developing a summarization model to efficiently condense processed complaint content. Our goal is to incorporate diverse low-resource multilingual Indian languages and various data types, including images and documents, to enhance model adaptability and performance across different formats and linguistic contexts.

References

[1] Author(s). Title of the webpage. https://www.hindustantimes.com/health/the-wait-never-ends-complaints-of-medical-negligence-increase-but-justice-eludes-victims/story-eFDpT6vKYQSVhN0ovgCUBN.html. Accessed: April 10, 2024.

[2] O. Baclic, M. Tunis, K. Young, C. Doan, H. Swerdfeger, and J. Schonfeld. Artificial intelligence in public health: Challenges and opportunities for public health made possible by advances in natural language processing. *Canada Communicable Disease Report*, 46(6):161, 2020.

[3] S. Birkeland, R. Depont Christensen, N. Damsbo, and J. Kragstrup. Characteristics of complaints resulting in disciplinary actions against danish gps. *Scandinavian journal of primary health care*, 31(3):153–157, 2013.

[4] M. Bismark and E. A. Dauer. Motivations for medico-legal action: lessons from new zealand. *The Journal of legal medicine*, 27(1):55–70, 2006.

[5] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860, 2018.

[6] S. Das, A. Singh, R. Jain, S. Saha, and A. Maurya. Let the model make financial senses: A text2text generative approach for financial complaint identification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 58–69. Springer, 2023.

[7] D. Demner-Fushman and N. Elhadad. Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. *Yearbook of medical informatics*, 25(01):224–233, 2016.

[8] K. Denecke and Y. Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1): 17–27, 2015.

[9] K. Denecke and W. Nejdl. How valuable is medical social media data? content analysis of the medical web. *Information Sciences*, 179(12): 1870–1880, 2009.

[10] C. Dreisbach, T. A. Koleck, P. E. Bourne, and S. Bakken. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International journal of medical informatics*, 125:37–46, 2019.

[11] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[12] G. Gonzalez-Hernandez, A. Sarker, K. O'Connor, and G. Savova. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, 26 (01):214–227, 2017.

[13] J. Heinström. Fast surfing, broad scanning and deep diving: The influence of personality and study approach on students' information-seeking behavior. *Journal of documentation*, 61(2):228–247, 2005.

[14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.

[15] M. Jin and N. Aletras. Modeling the severity of complaints in social media. *arXiv preprint arXiv:2103.12428*, 2021.

[16] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1): 1–9, 2016.

[17] M. Khadhraoui, H. Bellaaj, M. B. Ammar, H. Hamam, and M. Jmaiel. Survey of bert-base models for scientific text classification: Covid-19 case study. *Applied Sciences*, 12(6):2891, 2022.

[18] A. Khandelwal, H. Mittal, S. S. Kulkarni, and D. Gupta. Large scale generative multimodal attribute extraction for e-commerce attributes. *arXiv preprint arXiv:2306.00379*, 2023.

[19] D. D. Lewis, Y. Yang, T. Russell-Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[20] D. E. Losada, F. Crestani, and J. Parapar. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 346–360. Springer, 2017.

[21] K. Loveys, P. Crutchley, E. Wyatt, and G. Coppersmith. Small but mighty: affective micropatterns for quantifying mental health from social media language. In *Proceedings of the fourth workshop on computational linguistics and clinical Psychology—From linguistic signal to clinical reality*, pages 85–95, 2017.

[22] R. McCreadie, C. Buntain, and I. Soboroff. Trec incident streams: Finding actionable information on social media. 2019.

[23] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, and M. N. Halgamuge. From google gemini to openai q*(q-star): A survey of reshaping the generative artificial intelligence (ai) research landscape. *arXiv preprint arXiv:2312.10868*, 2023.

[24] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.

[25] E. Olshtain and L. Weinbach. 10. complaints: A study of speech act behavior among native and non-native speakers of hebrew. In *The pragmatic perspective*, page 195. John Benjamins, 1987.

[26] D. Preotiuc-Pietro, M. Gaman, and N. Aletras. Automatically identifying complaints in social media. *arXiv preprint arXiv:1906.03890*, 2019.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1 (8):9, 2019.

[28] R. M. Reiter, R. H. Downing, and M. Iveson. Global expectations, local realities: All-inclusive hotel reviews and responses on tripadvisor. *Contrastive Pragmatics*, 1(aop):1–33, 2023.

[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[30] S. Scepanovic, E. Martin-Lopez, D. Quercia, and K. Baykaner. Extracting medical entities from social media. In *Proceedings of the ACM conference on health, inference, and learning*, pages 170–181, 2020.

[31] A. Singh, R. Bhatia, and S. Saha. Complaint and severity identification from online financial content. *IEEE Transactions on Computational Social Systems*, 2023.

[32] A. Trosborg. *Interlanguage pragmatics: Requests, complaints, and apologies*, volume 7. Walter de Gruyter, 2011.

[33] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf. Zephyr: Direct distillation of lm alignment, 2023.

[34] F.-C. Yang, A. J. Lee, and S.-C. Kuo. Mining health social media with sentiment analysis. *Journal of medical systems*, 40:1–8, 2016.