Discourse Mode Categorization of Bengali Social Media Health Text

Salim Sazzed
Department of Computer Science
Old Dominion University
Norfolk, VA, 23529
salim.sazzed@gmail.com

Abstract

The scarcity of annotated data is a major impediment to natural language processing (NLP) research in Bengali, a language that is considered low-resource. In particular, the health and medical domains suffer from a severe paucity of annotated data. Thus, this study aims to introduce BanglaSocialHealth, an annotated social media health corpus that provides sentence-level annotations of four distinct types of expression modes, namely narrative (NAR), informative (INF), suggestive (SUG), and inquiring (INQ) modes in Bengali. We provide details regarding the annotation procedures and report various statistics, such as the median and mean length of words in different sentence modes. Additionally, we apply classical machine learning (CML) classifiers and transformer-based language models to classify sentence modes. We find that most of the statistical properties are similar in different types of sentence modes. To determine the sentence mode, the transformer-based M-BERT model provides slightly better efficacy than the CML classifiers. Our developed corpus and analysis represent a much-needed contribution to Bengali NLP research in medical and health domains and have the potential to facilitate a range of downstream tasks, including question-answering, misinformation detection, and information retrieval.

1 Introduction

With the increasing popularity of social media, various types of online content generated by vast numbers of people have become available. The health and medicine-related data are no exception, accumulating at a high pace as more and more people are using social media for health-related queries and discussions (Andy et al., 2021; Ganti et al., 2022). In fact, nowadays, by possessing ample amounts of health-related information, social media has become one of the prominent data sources for health-related research. People all over the world use online health forums to acquire medical information. Besides, people share their experiences regarding diseases, symptoms, and related matters to help other patients. Due to the importance of medical and health text mining, the NLP community has organized a series of open challenges focusing on biomedical entity extraction and classification (Weissenbacher et al., 2019).

The importance of social support in online health forums has been discussed in many earlier studies (Wang et al., 2012; Yang et al., 2017). As individuals seek support and information regarding various health-related issues in health and well-being forums, it is imperative to analyze them for a better understanding of user needs and to provide required support (Andy et al., 2021; Moorhead et al., 2013). For example, Andy et al. (2021), in their study, classified COVID-related health text into the following four categories: i) Emotional Support Given, ii) Emotional Support Sought, iii) Informational Support Given, and iv) Informational Support Sought. Another health text classification task was performed by Ganti et al. (2022), where the authors classified health-related text into narratives and non-narrative categories. A study related to the identification of informative health posts was conducted by Olsen and Plank (2021).

Categorizing health-related text on social media into distinct discourse modes can be beneficial for a range of downstream natural language processing (NLP) tasks, as each mode has specific roles in health support and discussion. For instance, user-generated questions can provide insights into the outbreak of the disease over time (Wen and Rosé, 2012) and can facilitate the development of social support chatbots that cater to the needs of individuals seeking healthcare-related assistance (Wang et al., 2021). User narratives or experiences can reveal valuable information on disease symptoms and severity. In addition, it can help to find peers with similar experiences (Levonian et al., 2021). It is imperative to analyze the suggestions or information-

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 52–57

related discourse shared by peer users to detect the dissemination of disinformation and misinformation (Wang et al., 2019).

In English and a few other major languages, various health-related corpora are publicly available (Kolárik et al., 2008). However, although Bengali (also known as Bangla) is one of the most spoken languages in the world [1], such resources are typically not available (Sazzed, 2022). With the growing popularity of telemedicine and the availability of health and medical-related data written in Bengali, creating resources for developing an NLP-based health system in Bengali is a pressing necessity. To comprehend and automatically categorize health-related text, it is essential to have at least a moderate amount of annotated data.

Hence, in this study, we introduce a discourse mode annotated health corpus, the first of its kind, for the low-resource Bengali language. The dataset is created by retrieving publicly available health-related texts from a number of social media health forums. The retrieved text data are tokenized into sentence-level and annotated with four types of discourse modes: narrative (NAR), informative (INF), suggestive (SUG), and inquiring (INQ). The final corpus contains around 2000 sentences annotated by one of the four types of sentence modes. The details of the annotation procedure and various statistics of the sentence modes are provided. In addition, we present a baseline evaluation by employing multiple ML classifiers for the automatic categorization of the sentence modes. We observe that top classical ML classifiers and deep learning-based models demonstrate similar efficacy for the classification tasks.

1.1 Contributions

The main contributions of this study can be summarized as follows:

- To address the lack of annotated health-related text data in Bengali, we present a health corpus, BanglaSocialHealth, by collecting health data from various Bengali health forums.
- We manually annotate around 2000 sentences into four types of modes: narrative, informative, suggestive, and inquiring. The dataset is publicly available in the following link [2].
- We provide various statistics, such as frequency and attributes of discourse mode annotated sentences in the corpus.
- Finally, we provide the baseline evaluations of the classification tasks utilizing both classical machine learning classifiers and multi-lingual BERT.

[1] https://www.berlitz.com/blog/most-spoken-languages-world

[2] https://github.com/sazzadcsedu/BanglaHealthText.git

2 Creation of BanglaSocialHealth

2.1 Data Collection

We obtain health-related textual data from multiple Bengali Facebook pages where individuals actively engage in discussions related to health. These discussions involve inquiries about health issues, recommendations, and information sharing concerning symptoms and disease prevention. While the majority of posts consist of interactive discussions, such as questions, answers, and suggestions from individuals, we also encounter health-related articles authored by healthcare professionals that provide informative content. To maintain the anonymity of the users, we do not collect any user information; only user-written texts are extracted. Therefore, the dataset is anonymous. The data collection period spans from May 2022 to July 2022.

The posts are manually retrieved from the Facebook pages for annotation. We find the textual contents in the posts represent three different forms of languages: Bengali, English, and transliterated Bengali. Since we are only interested in Bengali text, we collected only the posts written primarily in Bengali. The excerpted texts are automatically segmented into sentence-level tokens based on the Bengali *dari* (i.e., 'l') delimiter, which is equivalent to the English full stop ('.') delimiter. The sentence-level tokens are then manually reviewed to ensure each represents a contextually meaningful single sentence. As social media data are noisy, it is not uncommon to have sentences with missing delimiters. Again, some sentences may end with different types of delimiters. The manual examination assures each instance represents a complete sentence. Any sentence written in English or in transliterated Bengali in a post is excluded from annotation.

2.2 Discourse Modes

The following four types of discourse modes are considered during the annotation process.

- **Narrative (NAR)**: This discourse mode is related to the written narratives. Narration is the use of a written or spoken commentary to convey a story, such as a particular event or scenario, to an audience [3]. For example, an individual may tell about the experience and suffering regarding a particular disease or symptoms.

- **Inquiring (INQ)**: This discourse mode pertains to sentences that embody user inquiries and requests for information and recommendations. For example, in a health forum, a user may ask questions or seek suggestions concerning disease/symptoms, or request information regarding other health-related concerns.

- **Informative (INF)**: This discourse mode comprises informative sentences, which primarily convey factual information. For instance, sentences that encompass information about disease attributes and preventive measures are categorized within this mode.

- **Suggestive (SUG)**:The suggestive sentence primarily encompasses suggestions, advice, or recommendations offered in response to an individual's request for guidance. For instance, when a user seeks recommendations for a specialized doctor, another user may respond by providing specific suggestions.

2.3 Data Annotation Guidelines

In order to assign the discourse mode at the sentence level, annotators are provided with the aforementioned definitions and corresponding examples as guidelines. Initially, two annotators label all the sentences, and a third annotator intervenes only in cases where there is a disagreement between the first two annotators. We observed an annotator agreement of 0.80, calculated using Cohen's kappa (Cohen, 1960), for the label assignment between the first two annotators.

2.4 Corpus Statistics and Examples

Table 1 shows the frequency and word-length distributions of various discourse modes in the corpus. As we can see that the corpus exhibits an imbalanced distribution across various discourse modes. The most prevalent mode is NAR, which accounts for approximately 42% of the 2000 sentences in the corpus, while the INQ mode has the lowest representation among the sentences.

[3] https://en.wikipedia.org/wiki/Narration

Mode	#Frequency (%)	Length (word) (median/mean/std.)
NAR	840 (42.00%)	9/11.47/9.07
INQ	296 (14.93%)	8/9.47/5.42
INF	425 (21.20%)	12/12.50/6.04
SUG	405 (20.36%)	10/11.35/6.80

Table 1: Statistics of various discourse modes in the annotated corpus

3 Classification

3.1 Classical ML Classifier

We employ four classical supervised ML classifiers: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Tree (GBT) for determining the discourse modes of sentences. For classical ML classifiers, we select word unigrams and bigrams features from the corpus and compute corresponding term frequency-inverse document frequency (tf-idf) scores that act as inputs for the classifiers. For all classifiers, the default parameter settings of the scikit-learn (Pedregosa et al., 2011) library are used with class weight set to *balanced* (when applicable).

3.2 Transformer-based Model

We fine-tune M-BERT (Devlin et al., 2018) language model for categorizing sentences into four classes (i.e., discourse modes): NAR, INQ, INF, and SUG. Since this is a classification task, we utilize the classification module of the M-BERT. The HuggingFace library (Wolf et al., 2019) is used to fine-tune M-BERT. Since the initial layers of M-BERT only learn very general features, we keep them unchanged. Only the last layer of the M-BERT is fine-tuned for our classification task. We tokenize and feed our input training data to fine-tune the M-BERT model; Afterward, the fine-tuned model is used for classifying the testing data. A mini-batch size of 8 and a learning rate of 4×10^{-5} are used. The validation and training split ratio is set to 80% and 20%. The model is optimized using the Adam optimizer (Kingma and Ba, 2014) based on cross-entropy loss. The model is trained for 3 epochs with early stopping criteria set.

Discourse Mode	Sample Sentences
NAR	কিন্তু ও এখন আর লিকুইড খাবার ছাড়া শক্ত কোনো খাবার ই খেতে পারে না, মুখে নিলেও গিলতে পারেনা একটু পরে ফেলে দেই, আর এখন তো কিছু মুখেও নিতে চায়না।
	English: But now he can't eat any solid food except liquid food, he can't swallow it after a while, and now he doesn't even want to take anything in his mouth.
INQ	কোন হার্টের রোগীর যদি হাত ভাঙার অপারেশন করতে হয় সেক্ষেত্রে কি অ্যানাস্থেসিয়া প্রয়োগ করা যায়?
	English: Can anesthesia be applied to a heart patient undergoing an arm fracture operation?
INF	ভিটামিন_বি_১২ অত্যন্ত প্রয়োজনীয় একটি পুষ্টি উপাদান, যা পানিতে দ্রবণীয় অনেক খাবারে পাওয়া যায়।
	English: Vitamin B12 is an essential nutrient found in many water-soluble foods.
SUG	পেটের মেদ কাটিয়ে উঠতে চাইলে প্রতিদিন প্রচুর পরিমাণে পানি পান করতে হবে।
	English: If you want to get rid of belly fat, you need to drink a lot of water every day.

Figure 1: Samples sentences representing various discourse modes

Classifier	Discourse Mode				
	NAR	**INQ**	**INF**	**SUG**	**Overall**
	P/R/F1	P/R/F1	P/R/F1	P/R/F1	P/R/F1
LR	0.77/0.86/0.81	0.89/0.7/0.78	0.74/0.77/0.76	0.79/0.71/0.75	0.80/0.76/0.78
SVM	0.71/0.91/0.80	0.93/0.60/0.73	0.80/0.70/0.75	0.83/0.66/0.73	0.82/0.72/0.76
RF	0.61/0.94/0.74	0.92/0.55/0.69	0.79/0.41/0.54	0.85/0.52/0.65	0.79/0.61/0.69
GBT	0.71/0.84/0.77	0.83/0.65/0.73	0.72/0.62/0.67	0.80/0.66/0.72	0.76/0.70/0.74
M-BERT	0.77/0.84/0.80	0.94/0.70/0.78	0.77/0.78/0.77	0.78/0.72/0.76	0.82/0.78/0.80

Table 2: Performance of various classifiers for discourse mode classification

Class	NAR	INQ	INF	SUG
NAR	724	19	64	34
INQ	72	202	8	14
INF	73	0	326	26
SUG	76	5	35	289

Table 3: Confusion matrix of LR classifier

4 Results and Discussion

We evaluate the performances of various classifiers based on 5-fold cross-validation and report F1 scores. Table 2 presents F1 scores and accuracies of CML classifiers and transformers-based M-BERT models for sentence mode identification. The results indicate that the LR classifier yields the best performance among the four traditional ML classifiers by achieving an F1 score of around 0.78. SVM performs similarly and achieves an F1 score of about 0.76. The tree-based methods show comparatively inferior performances.

We observe that the performance of CML classifiers is affected by the class distribution of the dataset. All classifiers yield better results for the narrative (NAR) mode as NAR mode represents the highest number of samples (42%) in the dataset. Although the transformer-based multilingual language model yields slightly better performance than the CML classifiers, the improvement is not significant compared to CML classifiers, which can be attributed to the limited amount of labeled data. With more labeled data incorporated, the improvement may be higher as transformer-based models have shown state-of-the-art performances for various NLP tasks across languages.

Table 3 portrays the confusion matrix of the LR classifier from a sample run. We observe that misclassification is affected by the distribution of classes in most cases. Since NAR contains the highest number of samples in the dataset, we notice false negative (FN) predictions of other modes mainly refer to NAR. Nevertheless, for the NAR class, the FN classifications are mostly predicted as INF, even though INF and SUG have a similar number of instances.

5 Summary and Future Work

Developing an effective framework for analyzing social media health data has substantial practical applications. However, such tools require annotated data which is hardly available in a low-resource language like Bengali. Therefore, we introduce a Bengali health corpus created from several Bengali social media health pages. We report detailed

annotation guidelines and procedures for the annotation. Moreover, we provide various statistics of four types of discourse modes in the annotated corpus. We make the corpus publicly available for the researchers. Future work will focus on enlarging the size of the annotated corpus. Besides, we will investigate how to leverage cross-lingual resources from other languages, such as English, to improve the performance of this classification task.

6 Ethical statement

Research is based on publicly available data on Facebook. No user personal information is included in the analysis, and no user identity is disclosed.

References

Anietie Andy, Brian Chu, Ramie Fathy, Barrington Bennett, Daniel Stokes, and Sharath Chandra Guntuku. 2021. Understanding social support expressed in a covid-19 online forum. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 19–27.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Achyutarama Ganti, Steven Wilson, Zexin Ma, Xinyan Zhao, and Rong Ma. 2022. Narrative detection and feature analysis in online health communities. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 57–65.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Corinna Kolárik, Roman Klinger, Christoph M Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. 2008. Chemical names: terminological resources and corpora annotation. In *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*.

Zachary Levonian, Marco Dow, Drew Erikson, Sourojit Ghosh, Hannah Miller Hillberg, Saumik Narayanan, Loren Terveen, and Svetlana Yarosh. 2021. Patterns of patient and caregiver mutual support connections in an online health community. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–46.

S Anne Moorhead, Diane E Hazlett, Laura Harrison, Jennifer K Carroll, Anthea Irwin, and Ciska Hoving. 2013. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research*, 15(4):e1933.

Benjamin Olsen and Barbara Plank. 2021. Finding the needle in a haystack: Extraction of informative covid-19 danish tweets. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 11–19.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Salim Sazzed. 2022. Banglabiomed: A biomedical named-entity annotated corpus for bangla (bengali). In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 323–329.

Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. 2021. Cass: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–31.

Yi-Chia Wang, Robert Kraut, and John M Levine. 2012. To stay or leave? the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 833–842.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael Paul, and Graciela Gonzalez. 2019. Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019. In *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, pages 21–30.

Miaomiao Wen and Carolyn Penstein Rosé. 2012. Understanding participant behavior trajectories in online health support groups using automatic extraction methods. In *Proceedings of the 2012 ACM International Conference on Supporting Group Work*, pages 179–188.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Diyi Yang, Robert Kraut, and John M Levine. 2017. Commitment of newcomers and old-timers to online health support communities. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 6363–6375.