Overview of the shared task on code-mixed information retrieval from social media data

Supriya Chanda[1,*,†], Sukomal Pal[1]

[1]Indian Institute of Technology (BHU) Varanasi, Uttar Pradesh, India

Abstract

The rise of multilingual communication on social media platforms such as Facebook, Twitter, and WhatsApp presents a compelling challenge for information retrieval in code-mixed contexts within natural language processing. This paper provides an overview of the Code-Mixed Information Retrieval Shared Task, which is part of the FIRE-2024 conference. The main focus of this experiment was the evaluation of how relevant documents code-mixed from a corpus of Bengali-English comments were to be given for a set of code-mixed queries. Six teams showed interest in participating in the shared task; two teams provided their runs. This article describes the models used by the competing teams and their performance evaluated on the Mean Average Precision (MAP), a significant metric used for information retrieval tasks.

Keywords

Code-Mixed, Bengali, English, Information Retrieval, Social Media

1. Introduction

The proliferation of multilingual and code-mixed content on digital platforms, especially in multilingual societies like India, brings challenging problems for Natural Language Processing (NLP) and Information Retrieval (IR). Code-mixing is the act of mixing two or more languages in a single discourse, a common linguistic phenomenon. Bengali-English and Hindi-English are typical examples in India. Traditional IR systems, mainly designed for monolingual datasets, face challenges when dealing with the complexities of code-mixed data. This calls for new approaches tailored to these hybrid linguistic environments. As online social networks continue to grow, many of its users communicate in native languages using foreign scripts. This is a norm in India, where people use the Roman script on social networks. The trend is mostly noticeable among migrants who form an online community to share relevant information and experiences.

These discussions usually contain code-mixed text, wherein users use informal, colloquial language often transliterated into Roman script. This lack of standardization makes it challenging to recognize and emphasize relevant answers from these discussions, especially when others are looking for the same information later. Our task is to create a means of identifying the most relevant answers to these code-mixed discussions. This will focus on Roman transliterated Bengali mixed with the English language.

The Bengali-English code mixing poses unique challenges for IR due to the inherent linguistic differences between the two languages. Bengali, being an inflectional language, has rich morphological variation, whereas English is a more rigidly structured language. These differences make standard IR tasks, such as tokenization, parsing, and language comprehension, challenging. Further complicating this task is the frequent use of Roman script for Bengali, which introduces transliteration issues, where non-standardized spellings and ambiguous language boundaries create additional hurdles for IR systems.

Despite the numerous advancements in multilingual NLP, research on IR for code-mixed languages still needs to be addressed. Much of the existing work has been on language identification, sentiment analysis, hate speech identification, and transliteration normalization. However, their application to IR

Forum for Information Retrieval Evaluation, December 12-15, 2024, India

*Corresponding author.

supriyachanda.rs.cse18@itbhu.ac.in (S. Chanda); spal.cse@itbhu.ac.in (S. Pal)

0000-0002-6344-8772 (S. Chanda)

in resource-scarce languages like Bengali needs an improvement. To bridge such gaps, linguistic insights can be integrated with machine learning approaches to handle the nuances that exist in code-mixed data. In recent years, we have explored various text processing tasks on code-mixed data like word-level language identification [1], sentiment analysis [2, 3, 4], hate speech identification [5] and sarcasm detection [6].

This paper outlines the overview of the **CMIR-2024: Code-Mixed Information Retrieval from Social Media Data** [1] shared task that focuses on developing IR systems for Bengali-English code-mixed data. The task focuses on contributing to more robust and inclusive IR systems that better serve multilingual digital communities by addressing the linguistic complexities of code-mixed text.

The participants would be provided with training and test dataset. This is an information retrieval task. Given a Query (Q), systems need to pinpoint the most relevant answers from these code-mixed documents. To our knowledge, this is the first shared task on information retrieval on Bengali-English Code-Mixed text.

This work discusses the various models submitted to the shared task and the results of the participating teams. The rest of the article is orchestrated as follows: Section 2 describes the shared task. Section 3 discusses about the dataset. Section 4 summarizes the systems and the methodologies used in each participating team for the shared task and highlights the features of each model. The analysis of the results and findings of the methodologies submitted by the participants are presented in Section 5. Concluding remarks are presented in Section 6.

2. Task Description

The task [2] deals with automatically determining the relevance of a query to a document within code-mixed data, mainly focusing on English and Roman transliterated Bengali. The idea is to classify whether a given document is relevant or not relevant to a query and rank the documents accordingly. It includes the handling of code-mixed text complexities, where the coexistence of elements from two languages, and informal non-standardized nature of language is dealt with. At the same time, this system should capture the correct semantic relationship between the query and the document.

We can define code-mixed IR (CMIR) like that when query terms and documents belong to different languages which may be using their native scripts or non-native ones. Here, both query and documents can contain multiple languages and scripts. If $q \in \left\langle L^{(i)}, S^{(j)} \right\rangle$ where $i \geq 2$ and $j \geq 1$. where L^i = union of i many languages and S^j = union of j many scripts. Similarly, the document pool thus becomes

$$D = \bigcup D_{L^{(i)},S^{(j)}}$$

where $L^{(i)} = \{l_1, l_2, \ldots, l_i\}$, $S^{(j)} = \{s_1, s_2, \ldots, s_j\}$ and
$D_{L^{(i)},S^{(j)}}$ = set of documents in language from $L^{(i)}$ written in script from $S^{(j)}$.

3. Dataset

It was challenging to find an appropriate code-mixed dataset on the web that matches our research objectives. Therefore, we created our own dataset by gathering data from social media platforms, namely Facebook [7]. We targeted groups and public pages with high engagement from Bengali-speaking users to ensure the inclusion of code-mixed Bengali-English language data. Bengali is the native language of people in both Bangladesh and the Indian state of West Bengal.

Through the data collection process, it was noticed that the majority of users post their questions in Facebook groups where replies are made through comments. In our dataset, queries are the original posts while the comments are documents containing which information needs to be extracted. This

[1] https://cmir-iitbhu.github.io/cmir/results.html
[2] https://cmir-iitbhu.github.io/cmir/

approach simply transformed the traditional information retrieval system by considering posts as a query and filtering the responses from comments.

The final dataset consists of 50 queries and 107,900 documents. We also tried different approaches to identify stopwords and measure their influence on information retrieval performance. Statistics of the dataset are as follows:.

Attributes	Values
Document and Query format	Text
Total number of documents in the corpora	107900
Total number of words	1363672
Total Number of unique words	84724
Total number of Bengali (BN) words	663363
Total Number of unique Bengali (BN) words	47510
Total number of English (EN) words	578480
Total Number of unique English (EN) words	25996
Total Number of Queries (Q)	50
Total Number of relevant documents (QRels)	802
Mean value of relevant documents per query	16.04

Table 1
Text collection statistics

4. Methodology

In total, six teams registered for the CMIR-2024: Code-Mixed Information Retrieval shared task. However, in this, only two; Team BITS and TextTitans were able to deliver their system outputs.

The Team **BITS team** examined numerous techniques, including more classic machine learning models as well as more advanced architectures built on top of the transformer pre-trained architecture. Sentence-BERT was front-and-center for semantic representation with Graph Neural Networks added in to capture relational information from the data. It then combined these methods together for the purpose of increasing retrieval of relevant information within the code-mixed text.

In contrast, the **TextTitans** team developed a novel methodology centered around the GPT-3.5 Turbo model. Their approach utilized a sequential engineering strategy to leverage the generative power of GPT-3.5 Turbo to handle code-mixed queries and improve retrieval accuracy. The fine-tuning of this model and the integration of the engineering steps tailored to the specific challenges of code-mixed IR were the aims of the team to address the linguistic complexities inherent in the task.

5. Results and Discussion

The evaluation of the systems submitted by Team BITS and Team TextTitans offers insight into their performance in terms of various metrics and approaches.

Team BITS tested several pre-processing and stemming techniques with their results. They also tried re-ranking the base model results with SBERT and independently applied an SBERT-based information retrieval model. With significant effort, the integration of a GNN-based model for re-ranking SBERT results was disappointing. The performance of GNN model was very unsatisfactory and not good as initially expected. This should mean there might be something amiss with the relation of the task to architecture or requires more tuning towards optimization. The team holds that further investigation is also necessary in order to highlight what exactly contributes to underperforming GNN based approach. Alternative strategies and further fine-tuning the GNN parameters would be explored in future work to make its ranking effectiveness potentially better.

Team TextTitans evaluated their system's performance using a set of standard information retrieval metrics: Mean Average Precision (MAP), normalized Discounted Cumulative Gain (NDCG), Precision at

Team Name	Submission File	MAP Score	nDCG Score	P@5 Score	P@10 Score
TextTitans	submit_cmir	0.701773	0.797937	0.793333	0.766667
TextTitans	submit_cmir_1	0.701773	0.797937	0.793333	0.766667
TextTitans	submit_cmir_2	0.701773	0.797937	0.793333	0.766667
TextTitans	submit_cmir_3	0.701773	0.797937	0.793333	0.766667
TextTitans	submit_cmir_4	0.703734	0.799196	0.793333	0.766667
Team BITS	submission_1	0.184110	0.429291	0.340000	0.250000
Team BITS	submission_2	0.233033	0.508931	0.426667	0.350000
Team BITS	submission_3	0.108895	0.374311	0.226667	0.173333
Team BITS	submission_4	0.012254	0.212133	0.013333	0.010000

Table 2
Performance Scores of Team Submissions

5 (P@5), and Precision at 10 (P@10). The results across all their submissions were very consistent, with very minor differences. For MAP, the first four submissions all returned the same score of 0.701, while the fifth submission scored slightly higher at 0.703. The NDCG scores for the first four submissions were identical at 0.797 and had a slight increase to 0.799 in the fifth submission. P@5 scores for all submissions were 0.793, which meant that all runs produced equal accuracy for the top five ranked documents. P@10 scores were identical across all submissions at 0.766. Although the fifth submission showed only a slight gain in terms of MAP and NDCG, precision metrics (P@5 and P@10) remained unchanged, which implies stability in performance for relevant documents retrieval in top-ranked results.

Analyzing both teams, the system of Team TextTitans had better performance consistency as observed with minute rank quality improvements by their fifth submission (See Table 2). Their usage of MAP, NDCG, and precision-based metrics implies that the retrieval system of Team TextTitans was stable, ranking most of the relevant documents atop all queries used. Meanwhile, the GNN-based re-ranking approach of Team BITS faced a problem:. This may have had further scope for improvement. Experiments performed with SBERT re-ranking for Team BITS indicated some possible improvement, but the addition of the GNN model did not improve performance and needed further investigation.

6. Conclusion

In conclusion, The Code-Mixed Information Retrieval Shared Task at FIRE-2024 showcased core challenges and opportunities arising during the retrieval of relevant documents in a code-mixed scenario, especially with regards to Bengali-English text. The task did well to present complexities regarding informal language usage and management through multiple scripts in the given code-mixed data. Only two teams provided system predictions, and the results give useful insight into how different models might work on this task. MAP score evaluation indicates that though there is some progress in this area, there is still much to be researched and modeled in order to catch the semantic subtleties of code-mixed languages. This shared task forms the foundation for further work in the area of code-mixed information retrieval and encourages more advanced techniques and broader participation in future editions.

Acknowledgments

We would like to express our sincere gratitude to Prof. Kripabandhu Ghosh (IISER Kolkata, India) and Prof. Thomas Mandl (Universitat Hildesheim, Germany) for providing us with the opportunity to organize this task as part of FIRE 2024. We deeply appreciate their trust and collaboration, which has significantly contributed to the growth and recognition of our work.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] S. Chanda, A. Misha, S. Pal, Advancing language identification in code-mixed tulu texts: Harnessing deep learning techniques., in: FIRE (Working Notes), 2023, pp. 223–230.

[2] S. Chanda, S. Pal, Irlab@ iitbhu@ dravidian-codemix-fire2020: Sentiment analysis for dravidian languages in code-mixed text., in: FIRE (Working Notes), 2020, pp. 535–540.

[3] S. Chanda, A. Mishra, S. Pal, Sentiment analysis and homophobia detection of code-mixed dravidian languages leveraging pre-trained model and word-level language tag, in: Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR, 2022.

[4] S. Chanda, A. Mishra, S. Pal, Sentiment analysis of code-mixed dravidian languages leveraging pretrained model and word-level language tag, Natural Language Processing (2024) 1–23. doi:10.1017/nlp.2024.30.

[5] S. Chanda, S. Sheth, S. Pal, Coarse and fine-grained conversational hate speech and offensive content identification in code-mixed languages using fine-tuned multilingual embedding, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE). CEUR-WS. org, 2022, pp. 502–512.

[6] S. Chanda, A. Mishra, S. Pal, Sarcasm detection in tamil and malayalam dravidian code-mixed text., in: FIRE (Working Notes), 2023.

[7] S. Chanda, S. Pal, The effect of stopword removal on information retrieval for code-mixed data obtained via social media, SN Comput. Sci. 4 (2023) 494. URL: https://doi.org/10.1007/s42979-023-01942-7. doi:10.1007/S42979-023-01942-7.