MMT: A Multilingual and Multi-Topic Indian Social Media Dataset

Dwip Dalal[1], **Vivek Srivastava**[2], **Mayank Singh**[1]
[1]IIT Gandhinagar, Gandhinagar, India
[2]TCS Research, Pune, India
{dwip.dalal, singh.mayank}@iitgn.ac.in
srivastava.vivek2@tcs.com

Abstract

Social media plays a significant role in cross-cultural communication. A vast amount of this occurs in code-mixed and multilingual form, posing a significant challenge to Natural Language Processing (NLP) tools for processing such information, like language identification, topic modeling, and named-entity recognition. To address this, we introduce a large-scale multilingual, and multi-topic dataset (*MMT*) collected from Twitter (≈1.7 million Tweets), encompassing 13 coarse-grained and 63 fine-grained topics in the Indian context. We further annotate a subset of 5,346 tweets from the *MMT* dataset with various Indian languages and their code-mixed counterparts. Also, we demonstrate that the currently existing tools fail to capture the linguistic diversity in *MMT* on two downstream tasks, i.e., *topic modeling* and *language identification*. To facilitate future research, we will make the anonymized and annotated dataset available in the public domain.

1 Introduction

In the last decade, we have observed high growth in the number of available social media platforms, as well as the user engagement on these platforms (Liu et al., 2014). Such widespread usage of these platforms makes them the primary means of information spread within as well as across cultures in any socially engaging event such as elections (Jungherr, 2016), entertainment (Antelmi et al., 2018), sports (Wang, 2020), science (López-Goñi and Sánchez-Angulo, 2018), and technology (Kreiss and McGregor, 2018).

India, with a population of over 1.3 billion, attracts the attention of all major social media firms(Aneez et al., 2019); various studies (Bharucha, 2018; Singh et al., 2019) reaffirm the active participation of Indians on these platforms. With diversity and multilingualism deeply ingrained in the culture of India (Ishwaran, 1969), it is no wonder that we find huge volumes of code-mixed data (Thara and Poornachandran, 2018) in the Indian social media space – which consequently makes it a goldmine for the NLP research community(Conway et al., 2019).

The NLP community has always been interested in solving problems in multilinguality (Xue et al., 2021) and multi-topicality (Yuan et al., 2018). In most of the research, the two problems are addressed separately. However, several interesting questions emerge in multilingual-multitopical datasets. Here, we explore three research questions:

- ***RQ1***: how traditional topic modeling tools perform in multilingual settings?
- ***RQ2***: can we achieve better topic modeling with the multilingual data using the contextual topic models?
- ***RQ3***: how do multilingual language identification tools perform in multi-topical text?

To the best of our knowledge, we have not found extensive investigation into the answers to the above questions. This paper explores these pertinent questions supported by robust evaluations and presents interesting anecdotal examples.

2 Constructing The Multilingual and Multi-topic Dataset

2.1 MMT

The large-scale multilingual and multi-topic dataset is constructed in four phases as listed below:

1. **Annotator selection and grouping**: We selected a diverse group of 49 students who were either undergraduates, masters, or postgraduates from different regions and cultural backgrounds in India. These students hailed from various states across India, representing different parts of the country from north to south, east to west. The 49 students were self-organized into 13 teams, with 10 teams consisting of 4

Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 47–52
May 5, 2023 ©2023 Association for Computational Linguistics

members each and 3 teams consisting of 3 members each. All the students were native Indians and active Twitter users with high proficiency in English and knowledge of at least one Indian language. This selection criterion ensured a diverse and representative sample.

2. **Topic identification**: As an initial step, we identify 13 topics relevant to the Indian context to capture and cater to various dimensions of discussions on social media, specifically Twitter. We enlist all 13 topics in Table 1. The choice of seed topics is also motivated by the most frequently discussed and relevant areas to the Indian community, as it helps get quality large-scale data easily from Twitter.
3. **Subtopic selection**: Next, we collect the fine-grained categorization for each of the 13 seed topics. We assign one seed topic to each team and ask them to develop a set of subtopics within each seed topic. The teams have the flexibility to do their own study (within and outside the Twitter community) to come up with a set of subtopics. We provided teams with constructive feedback and suggestions for improvement to ensure the accuracy and relevance of the selected subtopics. We fostered a collaborative process to arrive at a consensus on 63 subtopics that encompass diversity and exhaustiveness. The selected subtopics for all 13 seed topics are presented in Table 1.
4. **Data collection**: We curate data from Twitter based on the assigned subtopics for each seed topic. For this task, we employ the same set of 13 teams with a task of scraping at least 100K tweets (and the associated data and metadata) using the TWINT tool[1] . The teams are encouraged and rewarded to curate more than 100K tweets. We further preprocess and remove the tweets with missing values.

   In total, MMT comprises 1,755,145 tweets, with 135K tweets on average for each topic (Table 1). We observe a high degree of multilingualism, with tweets coming from 47 languages (as identified by Twitter). Based on manual inspection, we observe that the Twitter language identification system (hereafter *"TLID"*) assigns incorrect language tags to a large number of non-English tweets.

[1] https://github.com/twintproject/twint

Example 1

TWEET: In Taj Mahotsav Mukatakashiy Manch our Guru ji Dr. Sadanand Brahmbhatt hamein unke sath stage share karne ko mila ...
TWITTER ASSIGNED LANGUAGE: Hindi(Hi)
ANNOTATOR ASSIGNED LANGUAGE: Hindi-English (Hi-En)

Example 2

TWEET: @HarrietTurle @HarrietTurle Kaziranga Assam in East Ranthambore in North, Kahana national park in Madhya pradesh Bandhavgarh in Rajeshthan
TWITTER ASSIGNED LANGUAGE: Hindi(Hi)
ANNOTATOR ASSIGNED LANGUAGE: English (En)

Figure 1: Tweets from the *MMT-LID* dataset with language tags from Twitter and the human annotator.

2.2 MMT-LID

We construct this dataset using a language annotation task on the MMT dataset. We assign each team member (of the 13 teams) a randomly selected set of 500 tweets (with no duplicates) from the same seed topic as assigned in the *MMT's data collection* step. We provide the following guidelines for the annotation task:

- For each selected tweet, mark if the Twitter-assigned language tag is correct. In case the tag is incorrect, identify the correct language tag. In case the text mixes multiple languages, assign a combined tag by separating them using a hyphen. For example, if the tweet text mixes Hindi (either in Devanagari or Roman) and English tokens, the first answer will be '*No*', and the second answer will be '*Hi-En*'.
- In case the tweets are code-mixed, identify and annotate the main language (whose grammar is followed) and the embedded language (whose few tokens are embedded in the main language). For example, in the tweet "*items ko cart me daal ke app band kar dena is not funny*", the main language is '*Hi*' and embedded language is '*En*'.

As a result of the annotation, we obtain 5,346 tweets with human-annotated language tags. To evaluate the annotator's performance on this task, we evaluate the inter-annotator agreement (IAA) for each of the 13 topics using Cohen's Kappa (CK) score. We re-annotate 325 tweets (25 randomly selected tweets from each topic of the MMT-LID dataset) with the language tags and then calculate CK for IAA. Overall, we achieve an IAA score of 0.94. In Table 2, we report IAA scores per topic.

Topic	Subtopics	# Tweets	Avg len
Environment	Pollution, Climate Change, Eco Friendly, Floods	142208	216.58
Food	Online food delivery, Food security, Indian desserts	195086	140.75
Economics and Retail	Initial Public Offering (IPO), SEBI and New margin rules, Unicorns, Unemployment in India	158016	179.16
Natural Disaster	Cyclone:, Earthquake, Pandemic, Flood	75591	161.18
Art and Literature	Forms of Indian Art, Art festivals, Literature festivals, Book Fairs	111909	150.43
Sports	Olympics, Indian Premier League (IPL), Indian Super League (ISL), Pro Kabaddi League (PKL)	122740	117.06
Politics	Pegasus Snooping, Farmer Agitation, West Bengal Elections, 2021	119963	155.0
R&D and Technology	Mobile Technology, Health-Tech and Medical Innovations, ISRO	111615	166.43
Wildlife and Vegetation	Kaziranga National Park, Bandhavgarh National Park, Nilgiri National Park, Corbett National Park, Ranthambore National Park, Gir National Park, Nanda Devi National Park, Save Tiger Project, Save Elephants, Save the Great Indian Bustard, Wildlife Tourism and Heritage, Forest Cover, River Rejuvenation, Restoration, Wildlife Crime, Climate Change	280091	155.03
Manufacturing	Make in India, Steel Manufacturing, Automobile Manufacturing, Electronics and electrical manufacturing	100969	125.14
Films and OTT	OTT platforms such as Netflix, Amazon Prime Video, OTT Censorship, OTT Voicecalling, Nepotism in Film Industry, NationalFilmAwards	89760	142.12
Journalism & Media	Policy and Trends, Print Media & TV, Criminal Journalism, Social Movements and News	139563	170.88
Education	Exams, IIT, Online Education, Education System	107634	186.39

Table 1: Distribution of topics, subtopics, the number of tweets, and the average length of tweets in the *MMT* dataset. By incentivizing teams to collect over 0.1 million, we obtained more than 0.1 million tweets for 11 seed topics.

Topic	English	Hindi	Bengali	Marathi	Telugu	Unidentified	#L	Avg len	IAA
Environment	136929	575	5	22	8	2667	45	216.58	0.96
Food	135094	17141	125	366	96	10731	45	140.75	0.91
Economics & Retail	141766	4295	19	71	15	4801	45	179.16	0.94
Natural Disaster	37081	16670	740	547	1257	2915	43	161.18	0.91
Art and Literature	85389	2955	139	63	79	6977	44	150.43	0.93
Sports	56952	5493	519	129	83	9996	45	117.06	0.94
Politics	56469	25112	666	537	99	20532	43	155.0	0.89
R&D and Technology	75176	5428	99	195	129	3377	45	166.43	0.93
Wildlife & Vegetation	203024	18831	42	591	24	7063	45	155.03	0.90
Manufacturing	48421	1805	19	50	221	3220	47	125.14	0.94
Films & OTT	70314	2808	7	30	49	6274	44	142.12	0.94
Journalism & Media	80762	29663	838	1725	481	12703	46	170.88	0.90
Education	80848	7937	50	216	147	3435	45	186.39	0.92

Table 2: Topic-wise distribution of top-5 most spoken Indian languages (according to 2011 Census of India). #L: number of unique languages, and Avg len: average length of tweets.

2.3 Dataset Analysis

We make several interesting observations from the ***MMT*** and ***MMT-LID*** datasets. We list these observations below:

- Table 2 showcases that tweets for topics such as *'Environment'*, *'Education'*, and *'Economics & Retail'* are significantly longer than topics such as *'Sports'*, *'Manufacturing'*, and *'Food'*. The significant difference in the average lengths illustrates the diversity in the discussions; for example, agendas, news, and political topics represent lengthier conversations than match updates, movie reviews, and product launches.
- Figure 2 shows the distribution of top-5 languages (as identified by the human annotators) in the ***MMT-LID*** dataset. We observe that the majority (≈95%) of the English language tweets are correctly identified by Twitter. We identify that code-mixed language Hinglish is the second most frequent language in the dataset. TLID identifies the majority of the Hinglish tweets as either English or Hindi. We observe that 11.45% of tweets in ***MMT-LID*** dataset are code-mixed. This also includes tweets that mix English with other (non-Hindi) languages. Interestingly, we found 175 annotated tweets where none of the languages in the code-mixed pair were identified by TLID.

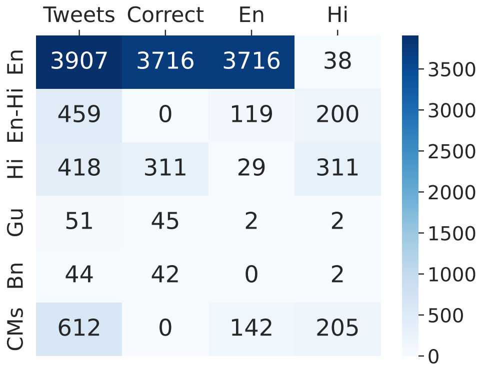

Figure 2: Distribution of language annotation by human annotators in the *MMT-LID* dataset. Here, we report the top-5 identified languages by the human annotators in the *MMT-LID* dataset. Here, `Correct` shows the number of tweets with correct language identification by Twitter. The column name `En` and `Hi` show the language identified by Twitter. `CMs` show all tweets in code-mixed languages.

Language	Metric	13 topics		63 subtopics	
		LDA	CTM	LDA	CTM
All	Accuracy	0.424	0.492	0.095	0.130
	W-F1	0.408	0.469	0.091	0.124
	Coherence	0.534	0.629	0.542	0.636
En	Accuracy	0.443	0.521	0.102	0.144
	W-F1	0.399	0.478	0.089	0.128
	Coherence	0.573	0.654	0.590	0.659
Non-En	Accuracy	0.398	0.461	0.084	0.119
	W-F1	0.379	0.437	0.086	0.113
	Coherence	0.384	0.512	0.407	0.563

Table 3: Perfomance evaluation of the topic modeling systems on the *MMT* dataset.

3 Answering the Pertinent Questions

In this section, we explore the three research questions posed in Section 1.

3.1 *RQ1*: how do traditional topic modeling tools perform in multilingual settings?

We answer this question by exploring the traditional topic modeling algorithm LDA (Blei et al., 2003). We conduct experiments on MMT and MMT-LID datasets based on the coarse and fine-grained topic categorization. For each experiment, we randomly partition the dataset into a 95:5 ratio, wherein a 95% split is used for training the LDA model and 5% for inference. We report the model's accuracy, weighted F1-score (W-F1), and coherence score (Röder et al., 2015) for each experiment.

3.1.1 Inferring topics in MMT dataset

In the first experiment, we separately train the LDA model on the MMT dataset's train split with 13 topics and 63 subtopics. Each of the trained topics (and subtopics) is manually assigned to one of the 13 original topics (and 63 subtopics). In Table 3, we report the result of our experiment with the LDA topic model on the inference split of the MMT dataset.

In the second experiment, we partition the MMT dataset into two partitions based on language tags assigned by Twitter's language identification tool. The first partition comprises English tweets (1,208,225 tweets), and another partition comprises non-English tweets (546,920 tweets). For each partition, we follow the same steps as the first experiment (described above). The scores (see Table 3) for the English partition are better than the non-English partition. We witness a significant drop in the accuracy and coherence scores in the non-English partition. This showcases the inefficacy of LDA in handling multilingual datasets. As English tweets are present in majority in the MMT dataset, we attribute this imbalance for higher scores of English against the full MMT dataset.

3.1.2 Inferring topics in MMT-LID dataset

Next, we conducted two similar experiments (described in the previous section) on the MMT-LID dataset. The main motivation for conducting these experiments is to bypass the errors introduced by Twitter's language identification tool. The results (see Table 4) follow the experimental observations conducted in the previous section. In comparison to non-English multilingual datasets, LDA performs better on monolingual English datasets. We believe that the small size of the dataset led to the discrepancy in the coherence score. The small size dataset limits the number of words for the model to learn. Thereby limiting the number of coherent words in a topic cluster, making the coherence score very volatile and dataset dependent (Syed and Spruit, 2017).

3.2 *RQ2*: can we achieve better topic modeling with the cross-lingual contextual topic model (CTM)?

The pertinent problem in the traditional LDA model lies with the bag-of-word (BoW) assumption, which disregards grammar and word order and only considers the frequency of words. As a result, such topic models cannot effectively deal with unseen words in the document. Additionally, such topic models do not perform well on multilin-

Language	Metric	13 topics		63 subtopics	
		LDA	CTM	LDA	CTM
All	Accuracy	0.395	0.488	0.090	0.141
	W-F1	0.363	0.434	0.082	0.129
	Coherence	0.447	0.602	0.442	0.619
En	Accuracy	0.472	0.637	0.139	0.193
	W-F1	0.448	0.591	0.126	0.179
	Coherence	0.386	0.589	0.418	0.624
Non-En	Accuracy	0.297	0.442	0.061	0.110
	W-F1	0.301	0.381	0.064	0.102
	Coherence	0.546	0.667	0.553	0.676

Table 4: Perfomance evaluation of the topic modeling systems on the *MMT-LID* dataset.

gual corpora without combining the vocabulary of multiple languages. To overcome these challenges, we experiment with **ZeroShotTM** (Bianchi et al., 2021), which is a cross-lingual contextual topic model supporting multilingual embeddings.

We conduct similar experiments described in Section 3.1 by replacing traditional LDA with ZeroShotTM. Tables 3 and 4 showcase the higher of ZeroShotTM (labeled as CTM) against LDA. However, the performance under the multilingual non-English partition is still significantly lower than the monolingual English partition.

3.3 *RQ3*: how do multilingual language identification tools perform in the multi-topical text?

Here, we explore the performance of the multilingual language identification systems on *MMT-LID* dataset. We experiment with four language identification systems as given in (Srivastava and Singh, 2021), i.e., Polyglot, FastText, Langdetect, and CLD3.

In addition, we report the performance of the TLID. We use the language tags assigned by the human annotators as a reference for evaluation. To report the system performance, we use two evaluation metrics, i.e., accuracy and weighted F1 score. Table 5 shows the results of multilingual language identification systems on the *MMT-LID* dataset. We observe that all the systems perform extremely well on the English dataset. We observe a drop in system performance with the entire *MMT-LID* dataset. Also, with only non-English data, all the systems show extremely poor results. These results indicate that multilingual language identification tools perform poorly in real-world settings where data from multiple languages and topics co-exist.

Language	Metric	TW	PG	FT	LD	CLD3
All	Accuracy	0.816	0.812	0.820	0.797	0.721
	W-F1	0.795	0.777	0.780	0.781	0.755
En	Accuracy	0.945	0.973	0.983	0.957	0.856
	W-F1	0.972	0.986	0.991	0.978	0.922
Non-En	Accuracy	0.462	0.372	0.379	0.360	0.352
	W-F1	0.392	0.362	0.349	0.352	0.348

Table 5: TW: Twitter, PG: Polyglot, FT: FastText, LD: Langdetect and CLD3: Compact Language Detector v3.

4 Limitations and Future Works

We collected the dataset from Twitter without language-specific constraints to reflect the real-world distribution of languages. This means that English, as a primary language, is over-represented in the dataset, while under-spoken languages such as Assamese are under-represented due to their limited use on the platform. This difference in distribution presents a challenge for building a robust multilingual system that performs well for such underrepresented languages. To overcome this, data augmentation techniques such as paraphrasing and oversampling, as well as transfer learning methods, can be utilized. These techniques can help balance the representation of languages in the dataset and further improve the performance of the multilingual system.

5 Concluding Remarks

In this paper, we present a multilingual and multi-topic dataset collected from Twitter for the Indian community spanning various Indian languages, including but not limited to the popular code-mixed languages. This could prove useful for further understanding and exploring the natural phenomenon of the co-existence of multilingual and multi-topical data. We also showcased several issues in topic modeling the multilingual dataset using traditional algorithms like LDA. We believe that the availability of such a large-scale and quality dataset will be useful in building systems for numerous downstream tasks such as multilingual topic modeling, language identification, machine translation, etc.

References

Zeenab Aneez, Taberez Ahmed Neyazi, Antonis Kalogeropoulos, and Rasmus Kleis Nielsen. 2019. India digital news report. (2019).

Alessia Antelmi, John Breslin, and Karen Young. 2018. Understanding user engagement with entertainment media: a case study of the twitter behaviour of Game of Thrones (GoT) fans. In *2018 IEEE Games, Entertainment, Media Conference (GEM)*. IEEE, 1–9.

Jehangir Bharucha. 2018. Social media and young consumers behavior. *International Journal of Supply Chain Management* 7, 6 (2018), 72–81.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1676–1683.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.

Mike Conway, Mengke Hu, and Wendy W Chapman. 2019. Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. *Yearbook of medical informatics* 28, 01 (2019), 208–217.

K Ishwaran. 1969. Multilingualism in India. In *Studies in Multilingualism*. Brill, 122–150.

Andreas Jungherr. 2016. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics* 13, 1 (2016), 72–91.

Daniel Kreiss and Shannon C McGregor. 2018. Technology firms shape political communication: The work of Microsoft, Facebook, Twitter, and Google with campaigns during the 2016 US presidential cycle. *Political Communication* 35, 2 (2018), 155–177.

Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. 2014. The tweets they are a-changin': Evolution of twitter users and behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Ignacio López-Goñi and Manuel Sánchez-Angulo. 2018. Social networks as a tool for science communication and public engagement: focus on Twitter. *FEMS Microbiology letters* 365, 2 (2018), fnx246.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*. 399–408.

Shiwangi Singh, Akshay Chauhan, and Sanjay Dhir. 2019. Analyzing the startup ecosystem of India: a Twitter analytics perspective. *Journal of Advances in Management Research* (2019).

Vivek Srivastava and Mayank Singh. 2021. Challenges and Limitations with the Metrics Measuring the Complexity of Code-Mixed Text. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. 6–14.

Shaheen Syed and Marco Spruit. 2017. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE, 165–174.

S Thara and Prabaharan Poornachandran. 2018. Code-mixing: A brief survey. In *2018 International conference on advances in computing, communications and informatics (ICACCI)*. IEEE, 2382–2388.

Yuan Wang. 2020. Building relationships with fans: how sports organizations used twitter as a communication tool. *Sport in Society* (2020), 1–15.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 483–498.

Michelle Yuan, Benjamin Van Durme, and Jordan L Ying. 2018. Multilingual Anchoring: Interactive Topic Modeling and Alignment Across Languages.. In *NeurIPS*. 8667–8677.