L3Cube-MahaSent-MD: A Multi-domain Marathi Sentiment Analysis Dataset and Transformer Models

Aabha Pingle[1,3], Aditya Vyawahare[1,3], Isha Joshi[1,3], Rahul Tangsali[1,3]
Raviraj Joshi[2,3]
[1] Pune Institute of Computer Technology, Pune, Maharashtra India
[2] Indian Institute of Technology Madras, Chennai, Tamil Nadu India
[3] L3Cube Pune
{aabhapingle, aditya.vyawahare07, joshiishaa, rahuul2001}@gmail.com
ravirajoshi@gmail.com

Abstract

The exploration of sentiment analysis in low-resource languages, such as Marathi, has been limited due to the availability of suitable datasets. In this work, we present L3Cube-MahaSent-MD, a multi-domain Marathi sentiment analysis dataset, with four different domains - movie reviews, general tweets, TV show subtitles, and political tweets. The dataset consists of around 60,000 manually tagged samples covering 3 distinct sentiments - positive, negative, and neutral. We create a sub-dataset for each domain comprising 15k samples. The MahaSent-MD is the first comprehensive multi-domain sentiment analysis dataset within the Indic sentiment landscape. We fine-tune different monolingual and multilingual BERT models on these datasets and report the best accuracy with the MahaBERT model. We also present an extensive in-domain and cross-domain analysis thus highlighting the need for low-resource multi-domain datasets. The data and models will be shared publicly.

Keywords: Marathi Sentiment Analysis, Multi-Domain Dataset, Low Resource Language, Movie Reviews, Subtitles, Tweets, BERT, Transformers

1 Introduction

Sentiment analysis is the process of determining the sentiment or emotional tone expressed in a piece of text, such as a sentence or a group of sentences. It involves analyzing the subjective content to identify whether it conveys positive, negative, or neutral sentiments (Pang et al., 2002). Sentiment analysis is a valuable technique in natural language processing (NLP) and finds applications across different domains (Blitzer et al., 2007). It can be applied to various forms of text data, including tweets, social media posts, customer reviews, surveys, news articles, and more (Roccabruna et al., 2022). Common uses of sentiment analysis include customer feedback analysis, social media monitoring, market research, financial analysis, and more.

Sentiment analysis can be utilized to aggregate sentiments expressed in movie reviews and calculate an overall sentiment score or rating for a movie (Maas et al., 2011). This enables users to quickly grasp the general sentiment towards a movie prior to watching it. Real-time monitoring of tweets (Ekbal et al., 2020) and detection of sentiment trends related to specific topics, events, or hashtags can be achieved through sentiment analysis. This aids businesses and marketers in comprehending public opinion, identifying emerging trends, and adjusting their strategies accordingly. Categorizing the emotional tone of different scenes or segments through sentiment analysis of subtitles is valuable. This can enhance content searchability and retrieval, making it easier for viewers to find specific types of content based on their moods. While these applications have been extensively researched in high-resource languages like English (Zhang et al., 2018), they are lacking in low-resource languages like Marathi (Joshi, 2022b; Kulkarni et al., 2022; Joshi, 2022a; Lahoti et al., 2022).

In this work, we present the L3Cube-MahaSent-MD dataset, an expanded version of the original L3CubeMahaSent (Kulkarni et al., 2021) dataset that specifically focuses on sentiment analysis in the Marathi language. This dataset includes sub-datasets in four domains, with three new domains introduced in this work: MahaSent-MR (movie reviews), MahaSent-GT (generic tweets), MahaSent-ST (TV show subtitles), and MahaSent-PT (political tweets). MahaSent-MR consists of Marathi movie reviews obtained by scraping various websites. MahaSent-GT includes a collection of Marathi tweets covering diverse topics, collected from Twitter using advanced techniques. MahaSent-ST consists of translated Marathi subtitles from the popular English TV show "Friends". Lastly, MahaSent-PT represents the original L3Cube-MahaSent dataset, specifically containing political tweets. Each dataset comprises approximately 15,000 examples in native Devanagari script, divided into training, validation, and test sets. The annotation policies employed for all these datasets are thoroughly described. Further, we also present the results of standard transformer-based BERT models on these datasets. The models used to evaluate the accuracies are MuRIL [1] (Khanuja et al., 2021), mBERT [2] (Devlin et al., 2019), MahaBERT [3] (Joshi, 2022a), and IndicBERT [4] (Kakwani et al., 2020). We also perform cross-domain analysis on these datasets using MahaBERT, the best-performing model. The main contributions of this work are as follows:

- We present **L3Cube-MahaSent-MD**, the first comprehensive multi-domain sentiment analysis dataset in Marathi, an Indian language. It comprises approximately 60,000 manually tagged sentences across four different domains, with positive, negative, and neutral labels. Specifically, our work contributes three new domains: movie reviews, generic tweets, and TV show subtitles, encompassing around 45,000 manually tagged sentences.

- This study introduces sub-datasets for sentiment analysis, namely MahaSent-MR, a dataset for Marathi Movie Reviews, MahaSent-GT, a dataset of Marathi general tweets, and MahaSent-ST, a dataset of Marathi subtitles. These datasets are the first of their kind and address a gap in the existing literature.

- We evaluate various monolingual and multilingual BERT models like MahaBERT, MuRIL, mBERT, and IndicBERT and demonstrate the superior performance of the monolingual MahaBERT model. Additionally, we release MahaBERT-based Marathi sentiment models for the domains considered in this study.

- We also conduct cross-domain analysis to demonstrate that domain-specific models do not generalize well across different domains. Thus, there is a significant need for multi-domain datasets in low-resource languages. We show that a multi-domain MahaBERT model trained on all the domains works competitively with the domain-specific models.

The datasets and models are shared publicly[5]. The individual models for MahaSent-MR[6], MahaSent-GT[7], MahaSent-ST[8], MahaSent-PT[9], and MahaSent-All[10] are available on Huggingface.

2 Related Work

In this section, we will discuss multi-domain datasets and low-resource datasets in the field of sentiment analysis. A vast amount of content is published on the internet every day, which has led to platforms like Twitter gaining significant attention for sentiment analysis tasks (Ekbal et al., 2020). Additionally, domains such as movie reviews have remained highly popular for sentiment analysis (Maas et al., 2011).

(Roccabruna et al., 2022) emphasizes the importance of enhancing the performance of BERT models across multiple sources and domains. The authors conduct an extensive evaluation of BERT-based models using sentiment analysis corpora from various domains and sources. Their research indicates that jointly fine-tuning the model on multi-source and multi-domain corpora yields superior performance compared to fine-tuning it solely on single-source and single-domain settings.

We conducted a thorough review of numerous sentiment analysis datasets available for low-resource languages. One of these datasets is HindiMD, a multi-domain corpus specifically developed for Hindi sentiment analysis (Ekbal et al., 2022). This dataset consists of a total of 9,090 tweets written in Hindi, which is considered a low-resource Indic language. The authors obtained these texts from Twitter by utilizing various keywords for data crawling. Through the creation of multiple baselines using this dataset, they effectively showcased its usefulness. However, it should be noted that this work does not differentiate between different domains and provides a single dataset encompassing all domains. As a result, the usability of the datasets for cross-domain analysis is limited.

(R et al., 2012) describes their work on cross-lingual sentiment analysis for Indian languages using linked wordnets. The dataset utilized in their study comprised user-written travel destination re-

[1] muril-base-cased
[2] bert-base-multilingual-cased
[3] marathi-bert-v2
[4] ai4bharat/indic-bert
[5] marathi-nlp
[6] marathi-sentiment-movie-reviews
[7] marathi-sentiment-tweets
[8] marathi-sentiment-subtitles
[9] marathi-sentiment-political-tweets
[10] marathi-sentiment-md

(a) MahaSent-MR dataset (b) MahaSent-ST dataset (c) MahaSent-GT dataset

Figure 1: Word clouds for the movie reviews, subtitles, and generic tweets domain

views. The authors conducted several experiments on this dataset, specifically focusing on the Hindi and Marathi languages.

(Ansari and Govilkar, 2018) utilized a dataset for their own experimentation on sentiment analysis in Marathi. This dataset included Marathi and Hindi documents sourced from social media platforms, encompassing data from chats, tweets, and YouTube comments. Notably, the dataset consisted of text in the transliterated (Romanized) format.

Released in 2021, the L3CubeMahaSent dataset (Kulkarni et al., 2021) is a manually labeled dataset comprising Marathi tweets. It consists of approximately 18,378 tweets that were classified into three sentiment classes. This dataset stands as the largest publicly available resource for Marathi sentiment analysis to date. However, it should be noted that the dataset exclusively contains tweets from the political domain. A similar tweet-based Marathi hate speech detection corpus, MahaHate was released in (Patil et al., 2022). Both MahaSent and MahaHate datasets have also been evaluated in a separate study (Velankar et al., 2022). A dataset for named entity recognition in Marathi, MahaNER was released in (Litake et al., 2022).

3 Dataset Collection

3.1 MahaSent-MR (Marathi Movie Reviews)

We identified a number of Marathi movie review websites in native Devanangari script for the MahaSent-MR dataset. The Marathi language movie reviews were scraped from TV9 [11] , eSakal [12] , Loksatta [13] , ABP Majha [14] , Lokmat [15], Saamana [16] and Maharashtra Times [17]. The movie reviews were scraped using the Requests module [18] in Python. The content of the articles was extracted from the webpages using the re (regular expressions) [19] Python module. The movie review articles obtained were then split into individual sentences using the period symbol (".") as the delimiter. All special characters were removed from the text.

3.2 MahaSent-ST (Marathi Subtitles)

The MahaSent-ST dataset includes English subtitles that have been translated into Marathi using the deep-translator tool [20]. All translations we manually verified and problematic generations were discarded. Subtitles from the situational comedy TV show F.R.I.E.N.D.S. have been used to create the dataset. The dataset was created using subtitles from seasons 1 through 4. The subtitle files had the .sub and .srt file extensions and were converted to text (.txt) files. The text file subtitles were then divided into individual sentences in the dataset. Special characters have not been removed from the dataset.

3.3 MahaSent-GT (Marathi Generic Tweets)

The MahaSent-GT dataset contains tweets in the Marathi language (native script). To scrape tweets,

[11] https://www.tv9marathi.com/live-tv

[12] https://www.esakal.com/

[13] https://www.loksatta.com/

[14] https://marathi.abplive.com/

[15] https://www.lokmat.com/

[16] https://www.saamana.com/

[17] https://maharashtratimes.com/

[18] https://pypi.org/project/requests/

[19] https://docs.python.org/3/library/re.html

[20] https://pypi.org/project/deep-translator/

we utilized the Python tools twint [21] and snscrape [22]. In order to ensure that the dataset contained generic tweets from wide range of topics, we scraped tweets using non-specific (stopwords in Marathi) keywords. We restricted the length of tweets to 15 words or less.

3.4 MahaSent-PT (Marathi Political Tweets)

The original MahaSent dataset (Kulkarni et al., 2021), referred to as MahaSent-PT in this work, contains tweets regarding current affairs. It features tweets in Marathi from political figures and activists presenting a variety of thoughts and perspectives. The dataset was manually categorized into three categories: negative, positive, and neutral. In the publicly available version of the dataset, hashtags, mentions, symbols and occasional English words have been retained. The tweets were scraped using the twint library.

4 Dataset Annotation

The sentences in the datasets were classified into three categories: positive (1), negative (-1) and neutral (0). All the datasets were manually annotated by four native Marathi speakers who are proficient in reading and writing Marathi. The Cohen's Kappa (Cohen, 1960) for the annotators is 0.86. The scraped setences were initially annotated using existing out-of-domain sentiment models in-order to aid the manual process. The Marathi MahaSent-MR and MahaSent-GT sentences were tagged (as positive, negative or neutral) using the MarathiSentiment model [23]. Similarly, the non-traslated English sentences from the MahaSent-ST dataset were tagged using VADER (Valence Aware Dictionary and sEntiment Reasoner)(Hutto and Gilbert, 2014) from the NLTK [24] library. The sentences were then sorted according to the tags, and the negative and positive sentences were annotated first. This was done to expedite the annotation process, establish a balanced dataset, and prevent the annotation of excessive neutral sentences. The initial out-domain models utilized for pseudo-labeling achieved an accuracy of 60-70% on the final datasets, highlighting the importance of manual annotation for new domains.

A set of guidelines were employed for the annotation of the datasets to ensure consistency in labeling. For the MahaSent-MR dataset, a sentence was labeled negative if a reviewer appeared to complain about an aspect of the film or reported a negative incident in the film. Similarly, each sentence that featured praise for the film or a description of some favorable developments in the story was labeled as a positive sentence. Sentences that simply provided facts about the movie or its events were considered neutral sentences.

One-word or small subtitles in the MahaSent-ST dataset were categorized based on the sentiment of the words present, regardless of context. "Yes", for example, was labeled as a positive subtitle. Grammatical errors in the subtitle translations have been rectified by the annotators. Incorrect but grammatically sound translations have not been altered.

Sarcastic tweets annotated for the MahaSent-GT dataset were marked negative. If the context of a tweet could not be ascertained, it was labeled as neutral. While tagging, English hashtags and emoticons were ignored, but Marathi hashtags were taken into account to determine the sentiment. Tweets offering condolences or expressing regret were categorized as negative.

Figures 2(a), 2(b) and 2(c) illustrate some sample sentences from the MahaSent-MR, MahaSent-GT and MahaSent-ST datasets, respectively.

5 Dataset Statistics

The updated version of L3CubeMahaSent consists of a total of 15000 records for MahaSent-MR, MahaSent-GT, and MahaSent-ST each. The MahaSent-MR domain consists of movie reviews, MahaSent-GT consists of general tweets, and MahaSent-ST comprises of TV show subtitles. Each of these domains has been annotated for sentiment analysis in the Marathi language. The dataset has been curated to ensure that the classes are balanced by randomly selecting an equal number of tweets for each class.

The MahaSent-MR, MahaSent-GT, and MahaSent-ST datasets have been split into train, test, and validation sets, where 12000 records of the data were used for training and 1500 records each were used for validation and testing.

We also create a single corpus by merging all four datasets (MahaSent-PT, MahaSent-MR, MahaSent-GT, and MahaSent-ST) together and maintaining an equal number of labels of each sentiment in the training, validation, and test sets. There are 48114 total examples in the training set con-

[21] https://pypi.org/project/twint/

[22] https://github.com/JustAnotherArchivist/snscrape

[23] https://huggingface.co/l3cube-pune/MarathiSentiment

[24] https://www.nltk.org/

Movie Review Sentence	English Translation	Label
विद्युतचे ॲक्शन सीन्स डोळे दिपवणारे आहेत	Electric action scenes are eye -catching	1
म्हणून तर ब्रम्हास्त्रचा पहिला भाग हा कमालीचा गोंधळून टाकणारा आहे	So the first part of the Brahmastra is the most confusing	-1
तो तिच्या घराचा पत्ता विसरला	He forgot the address of her house	0

(a) MahaSent-MR dataset

General Tweet	English Translation	Label
तुम्ही पण गुरुकुलमध्ये शिकला होतात का लहानपणापासून?	Have you also studied in Gurukul since childhood?	0
हे चुकीचं आहे अध्यक्ष महोदय	This is wrong sir	-1
अप्रतिम चित्रपट आहे, दोन वेळा बघितला	Amazing movie, watched it twice	1

(b) MahaSent-GT dataset

Subtitle	Marathi Translation	Label
Well, she has issues.	बरं, तिला समस्या आहेत.	-1
I cannot sleep in a public place.	मी सार्वजनिक ठिकाणी झोपू शकत नाही.	0
You look great.	तू छान दिसतोस	1

(c) MahaSent-ST dataset

Figure 2: Sample sentences from the movie reviews, subtitles, and generic tweets domain

Table 1: Dataset Statistics

Dataset	Total Samples	Train	Valid	Test	Average Sentence Length
MahaSent-PT (political tweets)	15864	12114	1500	2250	26.8658
MahaSent-GT (generic tweets)	15000	12000	1500	1500	10.4528
MahaSent-ST (subtitles)	15000	12000	1500	1500	6.3753
MahaSent-MR (movie review)	15000	12000	1500	1500	12.6323
MahaSent-All (all)	60864	48114	6000	6750	14.2630

taining 16038 examples of each sentiment, 6000 total examples in the validation set containing 2000 examples of each sentiment, and 6750 total examples in the test set containing 2250 examples of each sentiment. To further ensure the quality of the dataset, the commonly occurring words in each class have been visualized in the form of word clouds, which can be seen in Figures 1(a), 1(b), and 1(c). Table 1 gives a detailed view of the statistics of each of the datasets that have been used. The L3Cube-MahaSent-MD dataset is expected to facilitate sentiment analysis research in the Marathi language across multiple domains.

6 Baseline Models

6.1 mBERT

mBERT, also known as multilingual BERT, is a BERT-based model trained on and applicable to 104 languages (Devlin et al., 2019). It can be effectively utilized for downstream sentiment analysis tasks and was trained using the masked language modeling (MLM) and next sentence prediction (NSP) objectives.

To conduct sentiment analysis on Marathi text using mBERT, one can fine-tune the model on a Marathi sentiment analysis dataset. In this process, a new classification layer can be trained on top of mBERT. This new layer maps the representations learned by mBERT to the appropriate sentiment labels.

6.2 IndicBERT

IndicBERT is a language model based on the ALBERT architecture (Lan et al., 2020) and has been trained on a substantial corpus covering 12 major Indian languages: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu. The training data for IndicBERT was sourced from the IndicCorp dataset (Kakwani et al., 2020) and utilized joint training, which enables effective usage for underrepresented languages. In comparison to the XLM-R (Conneau et al., 2020) and mBERT models, IndicBERT generally demonstrates superior performance. Two variations of the model are available: IndicBERT base, trained on 12 million tokens, and IndicBERT large, trained on 18 million tokens.

6.3 MuRIL

MuRIL is a language model specifically developed for Indian languages and is trained exclusively on a substantial amount of Indian text data (Khanuja et al., 2021). To provide supervised cross-lingual signals during training, translated and transliterated document pairings are incorporated into the training set. This approach enhances MuRIL's ability to capture the nuances of Indian languages and effectively handle transliterated input. The performance of MuRIL was evaluated using the XTREME benchmark (Hu et al., 2020), which involves challenging cross-lingual evaluation tasks. Across the board, MuRIL outperformed multilingual BERT (mBERT) based on the evaluation results. Furthermore, transliterated test sets were utilized to assess the model's proficiency in handling such data.

6.4 MahaBERT

MahaBERT is a multilingual BERT model that has been fine-tuned on L3Cube-MahaCorpus, as well as other publicly available Marathi monolingual datasets (Joshi, 2022a). It was trained using the masked language modeling objective and trained on a corpus comprising 752M tokens.

7 Results

We conducted experiments using BERT and MuRIL-based models on four diverse datasets, and the corresponding results are presented in Table 2. The four models trained were IndicBERT, mBERT, MuRIL, and MahaBERT. The three-class classification was performed on four domain-specific datasets: MahaSent-PT (political tweets), MahaSent-MR (movie reviews), MahaSent-GT (general tweets), and MahaSent-ST (TV show subtitles). Additionally, we also evaluated the full mixed-domain corpus MahaSent-All. For the datasets MahaSent-PT, MahaSent-GT, and MahaSent-All, preprocessing involved the removal of hashtags, mentions, special symbols, and emoticons. However, MahaSent-ST and MahaSent-MR were clean datasets and did not require any specific preprocessing.

7.1 MahaBERT, a monolingual Marathi BERT model works the best across datasets

Upon analyzing the results, it was observed that the MahaBERT model achieved the highest accuracies for the MahaSent-PT, MahaSent-GT, and MahaSent-MR datasets compared to the MuRIL and other BERT-based models. Additionally, the Mahabert All model demonstrated an interesting ability to effectively learn from the shorter sentence structures present in TV show subtitles, resulting in the highest accuracy for the MahaSent-ST dataset.

Subsequently, we conducted further analysis on the performance of the best-performing model, MahaBERT, by comparing it across different domains. Initially, the model was trained on individual datasets, resulting in models named MahaBERT-PT, MahaBERT-GT, MahaBERT-ST, and MahaBERT-MR. Additionally, a model trained on the combined dataset was named MahaBERT-All. We then evaluated the accuracy scores of these models on the test sets of other domain datasets. The accuracy scores for all the models, with a maximum possible score of 100, are provided in Table 3.

7.2 Domain-specific models exhibit poor generalization, a mixed-domain model is more desirable

Upon comparing the performance of different models, we observed that the MahaBERT-PT, MahaBERT-GT, and MahaBERT-MR models displayed impressive accuracy on their respective datasets. Interestingly, the MahaBERT-All model, trained on the combined dataset, exhibited the highest accuracy scores for both the MahaSent-ST and MahaSent-All datasets. This indicates the MahaBERT-All model's remarkable ability to generalize well across various datasets, highlighting its versatility and consistently achieving high or near the highest accuracy levels.

Additionally, we noticed that the domain-specific models did not perform well on out-domain test sets. This observation emphasizes the unique intricacies associated with each domain and underscores the importance of having a multi-domain dataset for comprehensive sentiment analysis. Out of the four domain-specific models MahaSent-ST trained on subtitles corpus gave the best cross-domain numbers. However, the numbers were significantly lower than the MahaBERT-All model.

8 Conclusion

In this paper, we present an enhanced version of the L3CubeMahaSent dataset, which expands beyond political tweets (MahaSent-PT) to include three additional domains: movie reviews (MahaSent-MR), general tweets (MahaSent-GT), and TV show subtitles (MahaSent-ST), each containing 15,000 examples. We fine-tuned four distinct models, namely MuRIL, mBERT, MahaBERT, and IndicBERT, on

Table 2: Classification accuracies for the datasets using different BERT models

Model	MahaSent-PT	MahaSent-GT	MahaSent-ST	MahaSent-MR
mBERT	80.66	70.07	75.26	70.53
IndicBERT	84.13	76.06	77.00	74.13
MuRIL	84.30	77.13	78.73	78.00
MahaBERT	**84.90**	**78.80**	79.07	**78.53**
MahaBERT-All	83.95	77.86	**79.20**	77.73

Table 3: Cross-domain analysis of models trained on different datasets

Model	MahaSent-PT	MahaSent-GT	MahaSent-ST	MahaSent-MR	MahaSent-All
MahaBERT-PT	**84.90**	69.87	60.60	62.93	70.80
MahaBERT-GT	70.40	**78.80**	67.73	67.13	70.91
MahaBERT-ST	75.24	72.80	79.07	70.33	74.46
MahaBERT-MR	73.33	66.46	61.73	**78.53**	70.42
MahaBERT-All	83.95	77.86	**79.20**	77.73	**80.13**

these datasets for evaluation purposes.

Our analysis revealed that the MahaBERT model consistently achieved the highest accuracies on the MahaSent-PT, MahaSent-GT, and MahaSent-MR datasets. To further investigate its performance across different domains, we conducted a cross-domain analysis using MahaBERT as the base model. Impressively, the MahaBERT-All model, trained on the combined dataset, demonstrated excellent generalization abilities, consistently achieving the highest or near-highest accuracies across diverse domains.

By providing a low-resource multi-domain dataset and models trained on specific domains, we aim to equip researchers and practitioners with valuable resources to analyze sentiment across diverse domains.

Acknowledgments

This work was done under the L3Cube Pune mentorship program. We would like to express our gratitude towards our mentors at L3Cube for their continuous support and encouragement. This work is a part of the L3Cube-MahaNLP project (Joshi, 2022b).

References

Mohammed Arshad Ansari and Sharvari Govilkar. 2018. Sentiment analysis of mixed code for the transliterated hindi and marathi texts. volume 7.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar, Shikha Srivastava, et al. 2022. Hindimd: A multi-domain corpora for low-resource sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7061–7070.

Asif Ekbal, Pushpak Bhattacharyya, Shikha Srivastava, Alka Kumar, Tista Saha, et al. 2020. Multi-domain tweet corpora for sentiment analysis: resource creation and evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5046–5054.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Raviraj Joshi. 2022a. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101.

Raviraj Joshi. 2022b. L3cube-mahanlp: Marathi natural language processing datasets, models, and library. *arXiv preprint arXiv:2205.14728*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vishnu Subramanian, and Partha Pratim Talukdar. 2021. Muril: Multilingual representations for indian languages. *ArXiv*, abs/2103.10730.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, Jayashree Jagdale, and Raviraj Joshi. 2022. Experimental evaluation of deep learning models for marathi text classification. In *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021*, pages 605–613. Springer.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220.

Pawan Lahoti, Namita Mittal, and Girdhari Singh. 2022. A survey on nlp resources, tools, and techniques for marathi language processing. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–34.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. L3cube-mahaner: A marathi named entity recognition dataset and bert models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 29–34.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Hrushikesh Patil, Abhishek Velankar, and Raviraj Joshi. 2022. L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 1–9.

Balamurali R, Aditya Joshi, and Pushpak Bhattacharyya. 2012. Cross-lingual sentiment analysis for indian languages using linked wordnets. pages 73–82.

Gabriel Roccabruna, Steve Azzolin, and Giuseppe Riccardi. 2022. Multi-source multi-domain sentiment analysis with BERT-based models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 581–589, Marseille, France. European Language Resources Association.

Abhishek Velankar, Hrushikesh Patil, and Raviraj Joshi. 2022. Mono vs multilingual bert for hate speech detection and text classification: A case study in marathi. In *Artificial Neural Networks in Pattern Recognition: 10th IAPR TC3 Workshop, ANNPR 2022, Dubai, United Arab Emirates, November 24–26, 2022, Proceedings*, pages 121–128. Springer.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.