L3CubeMahaSent: A Marathi Tweet-based Sentiment Analysis Dataset

Atharva Kulkarni[1,3], Meet Mandhane[1,3], Manali Likhitkar[1,3], Gayatri Kshirsagar[1,3], and Raviraj Joshi[2,3]
[1]Pune Institute of Computer Technology, Pune
[2]Indian Institute of Technology Madras, Chennai
[3] L3Cube, Pune
{k.atharva4899,meetmandhanemnm,manalil1806,gayatrimohan7}@gmail.com
ravirajoshi@gmail.com

Abstract

Sentiment analysis is one of the most fundamental tasks in Natural Language Processing. Popular languages like English, Arabic, Russian, Mandarin, and also Indian languages such as Hindi, Bengali, Tamil have seen a significant amount of work in this area. However, the Marathi language which is the third most popular language in India still lags behind due to the absence of proper datasets. In this paper, we present the first major publicly available Marathi Sentiment Analysis Dataset - L3CubeMahaSent. It is curated using tweets extracted from various Maharashtrian personalities' Twitter accounts. Our dataset consists of ~16,000 distinct tweets classified in three broad classes viz. positive, negative, and neutral. We also present the guidelines using which we annotated the tweets. Finally, we present the statistics of our dataset and baseline classification results using CNN, LSTM, ULMFiT, and BERT-based deep learning models.

1 Introduction

The use of social media has seen a sharp upward trend in recent years. It plays a big role in forming and shaping the views of people on various issues. From sharing facts and opinions to voicing dissent and grievances, the platform has gained popularity amongst many users (Nielsen and Schrøder, 2014). Twitter is a significant social media platform. It has been quite popular in India for the past few years. It has been used by many politicians, journalists, and activists to connect with people directly. These kinds of interactions are generally strong on emotions, and can be used for developing sentiment analysis systems (Pak and Paroubek, 2010; Mathew et al., 2019). Such systems have proven to be important for political analysis as well as identifying and curbing more complex issues such as fake news, harassment, hate speech, and bullying (Schmidt and Wiegand, 2017; Joshi et al., 2021; Wani et al., 2021). In this work, we consider basic sentiment analysis or polarity identification tasks.

Popular languages such as English, Arabic, Russian, Mandarin (Rogers et al., 2018; Nabil et al., 2015; Yu et al., 2020) as well as Indian languages such as Hindi, Bengali and Tamil have been explored on the sentiment task for a long time (Arora, 2013; Patra et al., 2015; Akhtar et al., 2016; Mukku and Mamidi, 2017; Ravishankar and Raghunathan, 2017). Many resources such as properly annotated datasets, SentiWordNets, annotation guidelines have been developed for these languages (Socher et al., 2013; Saif et al., 2013; Hu and Liu, 2004). Alternatively due to the low resource nature of many languages, translated versions of the English datasets were used for analysis (Joshi et al., 2019; Refaee and Rieser, 2015; Mohammad et al., 2016). However, such translated datasets are often noisy due to the limitation of current translation systems for low resource languages.

Marathi is an Indian language spoken by around 83 million people and ranks as the third most spoken language in India. But surprisingly, there is no significant work or resource for the task of sentiment analysis in Marathi (Kulkarni et al., 2021). A sentiment analysis dataset curated by IIT-Bombay is available, but it has a very small size consisting of only 150 samples (Balamurali et al., 2012). In this paper, we present L3CubeMahaSent[1] - the largest publicly available Marathi Sentiment Analysis dataset to date. This dataset is gathered using Twitter. Our work is summarized as follows:

1. We present a ~16,000 tweets strong Marathi Sentiment Analysis Dataset, manually tagged into three classes viz. positive, negative, and

[1]https://github.com/l3cube-pune/MarathiNLP

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 213–220
April 19, 2021.

neutral.

2. We provide a comprehensive annotation policy useful for tagging sentences by their sentiment. We also provide statistics for our dataset and a balanced split for experimentation.

3. We present the result of our experiments on this dataset on recent deep learning approaches to create a benchmark for future comparisons.

2 Related Work

Sentiment analysis is a fundamental task of Natural Language Processing (Medhat et al., 2014). The absence of a proper sentiment analysis dataset for the Marathi language has led to limited research in this area. In this section, we will review some of the works introducing data resources in Indian and other languages. Balamurali et al. (2012) presented an approach for cross-lingual sentiment analysis using linked wordnets for Marathi and Hindi languages. For this purpose, they used various blogs and travel editorials as a dataset which consisted of about 75 positive and 75 negative reviews. The WordNet approach showed an improvement of 14-15 percent over the approach using a bilingual dictionary. The Marathi dataset created in this work is very small and cannot be used to train existing deep learning algorithms.

Bhardwaj et al. (2020) presented a hostility detection dataset in Hindi. Data was collected from various online platforms like Twitter, Facebook, Whatsapp, etc., and was benchmarked using machine learning algorithms namely, support vector machine (SVM), decision tree, random forest, and logistic regression. They also labeled each hostile post as either fake, hateful, offensive, or defamation.

Patra et al. (2018) presented details of a shared task in a competition on sentiment analysis of code-mixed data pairs of Hindi-English and Bengali-English. The best performing team used SVM for sentence classification. The sentiment analysis of code-mixed English-Hindi and English-Marathi text is also studied in (Ansari and Govilkar, 2018).

Nabil et al. (2015) introduced Arabic sentiments tweets dataset consisting of 10,000 tweets classified as objective, subjective positive, subjective negative, and subjective mixed. They tried 4-class sentiment analysis as well as 3-class sentiment analysis on the dataset and found that the former was more challenging. They also concluded that SVM performed well on the dataset for the task.

Rogers et al. (2018) presented RuSentiment, a dataset for sentiment analysis in the Russian language. They performed experiments on their dataset using algorithms like logistic regression, linear SVM, and neural networks. The best performance was observed in the case of neural networks. They also released the fastText embeddings that they have used for experimentation.

Ikoro et al. (2018) presented results of analyzing sentiments of UK energy consumers on Twitter. They proposed a method in which they combined functions from two sentiment lexica. The first lexicon was used to extract the sentiment-bearing terms and the negative sentiments. The second lexicon was used to classify the rest of the data. This method improved the accuracy compared to the general method of using one lexicon.

3 Curation of Dataset

3.1 Dataset Collection

For creating our dataset, we first manually created a list of various famous personalities who actively tweet about current affairs. Twitter profiles were shortlisted based on their frequency, relevance of activity, and degree of the sentiment of the tweets. Hence a majority of the tweets are from political personalities' profiles and activists as they express a wide range of emotions and sentiments. We attempted to improve the diversity of the points of view contained in the dataset.

All tweets in this dataset are specifically in the Marathi language. All hashtags, mentions, special symbols, and the occasional English words are kept in the tweets in the publicly available version of the dataset. We think it is best to keep the original dataset unhampered for anyone to experiment on it. However, while performing experiments, we have removed the aforementioned tokens from the tweets during data pre-processing. Also, the dataset does not retain any context of the tweets such as the tweeting profile, time of posting, region, etc.

As far as scraping the tweets is concerned, there are multiple python libraries available. Some of them are Tweepy (the official open-source library provided by Twitter)[2], GetOldTweets3[3], Twint[4], and Snscrape[5]. We have used the Twint library to scrape tweets.

3.2 Dataset Annotation

We have manually labeled the entire dataset into three classes: positive, negative, and neutral. These three classes have been denoted by '1','-1', and '0' respectively. The dataset was split among the entire team to tag in parallel. In order to maintain consistency while tagging tweets, we developed an annotation policy.

To begin with, we ensured not to take into account the author of the tweet, thereby eliminating any bias towards any author. Tweets are tagged by a general assumption that they are posted by any random person. Positive emotions such as happiness, gratitude, respect, inspiration, support are tagged as positive. Negative emotions such as hate, disrespect, grief, insult, disagreement, the opposition are tagged as negative. Tweets that do not convey a strong feeling, such as simple facts, statistics, or statements are tagged as neutral.

Tweets containing sarcasm, irony which clearly depict a negative sentiment are tagged as negative. Congratulatory and thank-you tweets are tagged as positive. A tweet that criticizes something or someone, or which states a fact stating an adverse event or reaction is termed negative. However, if the criticism comes as constructive and healthy, mentioning possible solutions, then it is tagged as positive. Finally, tweets containing mixed sentiments are labeled by the more dominant emotion expressed.

Even though these rules were laid down, there were some tweets that simply were difficult to tag by a single individual and needed to be reviewed. In such cases, we took a vote amongst the team and tagged the tweet according to the majority votes. Tweets for which no consensus could be formed were removed.

Some examples of tagged tweets are mentioned for more clarity in Table 3 given in the Appendix.

3.3 Dataset Statistics

Initially, we annotated a total of 18,378 tweets. But, in order to ensure that the classes are balanced, we randomly selected an equal number of tweets for each class. Hence, the final version of L3CubeMahaSent consists of 15,864 tweets. Table 1 shows class-wise distribution and the train-test-validation split. The remaining 2,514 annotated tweets will also be published along with the dataset. It consists of 2,355 positive and 159 negative tweets. These extra tweets have not been used for model evaluation. Commonly occurring words in each class can be visualized in the form of word-clouds as shown in Figure 2.

Split	Total Tweets	1	-1	0
Train	12114	4038	4038	4038
Test	2250	750	750	750
Validation	1500	500	500	500

Table 1: Dataset Statistics.

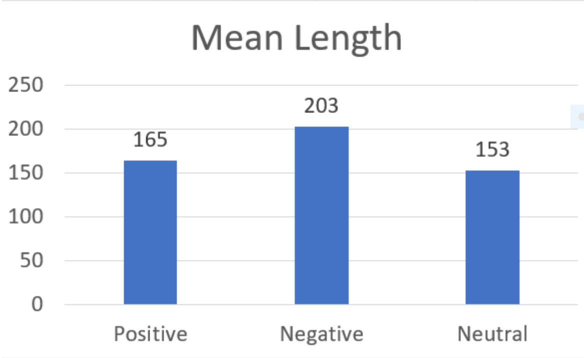

Figure 1: Mean length of records per class (in words)

4 Evaluation

4.1 Experimentations

We performed 2-class and 3-class sentiment analysis on our dataset. For conducting baseline experiments on our dataset, hashtags, mentions, and special symbols were removed during preprocessing. We used some of the widely used text classification architectures for sentiment classification (Kulkarni et al., 2021; Kowsari et al., 2019; Kim, 2014; Sun et al., 2019). The text is tokenized as words or sub-words and passed to the algorithms mentioned below:

- **CNN:** The initial embedding layer outputs word embeddings of size 300. These embeddings are passed to a Conv1D layer having 300 filters and kernel size 3. A global max-pooling is applied to the output sequences to get a sentence representation. This is then

[2]https://www.tweepy.org/
[3]https://pypi.org/project/GetOldTweets3/
[4]https://pypi.org/project/twint/
[5]https://github.com/JustAnotherArchivist/snscrape

Figure 2: Positive, Neutral and Negative wordclouds

passed on to a dense layer having size 100. A final dense layer having size equal to the number of classes is added to give classification results. We have experimented with various types of embedding layers having random initialization (word and subword), original Facebook fastText embeddings (trainable and non-trainable) (Mikolov et al., 2018), and Indic fastText embeddings (trainable and non-trainable) by IndicNLP (Kakwani et al., 2020).

- **BiLSTM+GlobalMaxPool:** This is similar to the CNN network with Conv1D layer replaced by a Bi-LSTM layer. Inputs are fed to an embedding layer which outputs word embeddings of size 300. These embeddings are given to a bi-directional LSTM layer with cell size 300 and then output is max pooled over time. A dense layer of size 100 and a subsequent dense layer of size equal to the number of classes complete the architecture. Embeddings same as those mentioned in the CNN section are also experimented with.

- **ULMFiT:** ULMFiT is also a LSTM based model (Howard and Ruder, 2018). It uses transfer learning which allows the model to be finetuned quickly on the target dataset using even a small sample set. We use a publicly available ULMFiT model for the Marathi language released by iNLTK and finetune it on our dataset (Arora, 2020).

- **BERT:** The BERT is a transformer based model pretrained on a huge text corpora, which can be finetuned for any target dataset (Devlin et al., 2019). Many publicly available flavours of BERT are available, and we use two specific multilingual models:
  - Multilingual-BERT (mBERT)
  - Indic-BERT by IndicNLP (Kakwani et al., 2020)

  For both of these models, we used the CLS token for sequence classification.

4.2 Results

We experimented with a variety of architectures such as CNN and BiLSTM for text classification on our dataset along with different embeddings. We have used random word and subword initializations, and also used pre-trained word embeddings made public by Facebook and IndicNLP. Both of these pre-trained embeddings were used in trainable and static modes. Along with these architectures, pre-trained models such as ULMFiT, mBERT, and IndicBERT were also used.

The results from our experiments were in line with previous works done in Marathi text classification. Pretrained word embeddings give a definite edge over random initializations. The CNN-based models have a slight advantage over BiLSTM based models. It was observed that though models using random initializations generally give good results, they tend to quickly overfit while training. The use of pre-trained embeddings significantly reduces overfitting. The results for 3-way and 2-way classification are shown in Table 2. The neutral class is dropped for 2-way classification.

The Marathi word embeddings provided by IndicNLP perform better than the original versions released by Facebook. Keeping word embeddings trainable further increases the accuracy. The ULMFiT gives results that are comparable with simple CNN and BiLSTM models. The CNN model combined with trainable Indic fastText word embeddings gives the best results in the 2-class classifi-

Model	Variant	3-class Accuracy	2-class Accuracy
CNN	random-word	79.47	90.00
	random-subword	81.56	91.73
	FB fastText-Trainable	81.02	92.67
	FB fastText-Static	80.18	90.93
	Indic fastText-Trainable	**83.24**	**93.13**
	Indic fastText-Static	83.00	92.47
BiLSTM	random-word	80.93	90.87
	random-subword	79.42	89.80
	FB fastText-Trainable	81.78	92.33
	FB fastText-Static	79.87	89.67
	Indic fastText-Trainable	**82.89**	91.8
	Indic fastText-Static	82.41	**92.67**
ULMFiT	(iNLTK)	80.80	91.40
BERT	mBERT	80.66	91.40
	IndicBERT (INLP)	**84.13**	**92.93**

Table 2: Classification accuracies over different architectures. The 2-class accuracy corresponds to positive and negative class only.

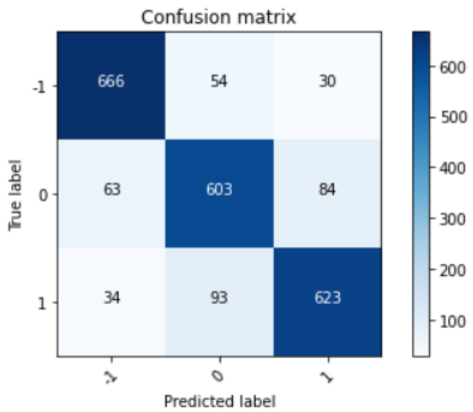

Figure 3: Confusion Matrix for IndicBERT - 3 class classification

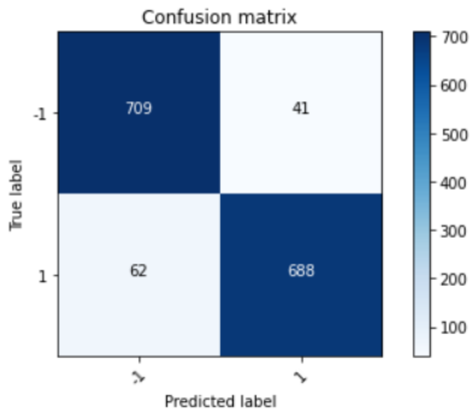

Figure 4: Confusion Matrix for CNN with Indic fastText (trainable) - 2 class classification

cation experiments, slightly outperforming the IndicBERT. The IndicBERT was the best performing model for the more difficult 3-class classification experiments. The respective confusion matrices are shown in Figure 4 and Figure 3.

5 Conclusion

In this paper, we have presented L3CubeMahaSent - the first major publicly available dataset for Marathi Sentiment Analysis which consists of ~16000 distinct tweets. We also describe the annotation policy which we used for manually labeling the entire dataset. We performed 2-class and 3-class sentiment classification to provide a benchmark for future studies. The deep learning models used for sentiment prediction were CNN, BiLSTM, ULMFiT, mBERT, and IndicBERT. The publicly available Marathi fastText embeddings were used with word-based models. We report the best accuracy using IndicBERT and CNN with Indic fastText word embeddings. We hope that our dataset will play a crucial role in advancing NLP research for the Marathi language.

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.

References

Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 482–493.

Mohammed Arshad Ansari and Sharvari Govilkar. 2018. Sentiment analysis of mixed code for the transliterated hindi and marathi texts. *International Journal on Natural Language Computing (IJNLC) Vol*, 7.

Gaurav Arora. 2020. inltk: Natural language toolkit for indic languages. *arXiv preprint arXiv:2009.12534*.

Piyush Arora. 2013. Sentiment analysis for hindi language. *MS by Research in Computer Science*.

AR Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. 2012. Cross-lingual sentiment analysis for indian languages using linked wordnets. In *Proceedings of COLING 2012: Posters*, pages 73–82.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi. *arXiv preprint arXiv:2011.03588*.

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.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Victoria Ikoro, Maria Sharmina, Khaleel Malik, and Riza Batista-Navarro. 2018. Analyzing sentiments expressed on twitter by uk energy company consumers. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 95–98. IEEE.

Ramchandra Joshi, Purvi Goel, and Raviraj Joshi. 2019. Deep learning for hindi text classification: A comparison. In *International Conference on Intelligent Human Computer Interaction*, pages 94–101. Springer.

Ramchandra Joshi, Rushabh Karnavat, Kaustubh Jirapure, and Raviraj Joshi. 2021. Evaluation of deep learning models for hostility detection in hindi text. *arXiv preprint arXiv:2101.04144*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4948–4961.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, Jayashree Jagdale, and Raviraj Joshi. 2021. Experimental evaluation of deep learning models for marathi text classification. *arXiv preprint arXiv:2101.04899*.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Tomáš Mikolov, Édouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Sandeep Sricharan Mukku and Radhika Mamidi. 2017. Actsa: Annotated corpus for telugu sentiment analysis. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 54–58.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2515–2519.

Rasmus Kleis Nielsen and Kim Christian Schrøder. 2014. The relative importance of social media for accessing, finding, and engaging with news: An eight-country cross-media comparison. *Digital journalism*, 2(4):472–489.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: an overview of sail_code-mixed shared task@ icon-2017. *arXiv preprint arXiv:1803.06745*.

Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. 2015. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 650–655. Springer.

Nadana Ravishankar and Shriram Raghunathan. 2017. Corpus based sentiment classification of tamil movie tweets using syntactic patterns. *IIOAB Journal: A Journal of Multidisciplinary Science and Technology*, 8(2):172–178.

Eshrag Refaee and Verena Rieser. 2015. Benchmarking machine translated sentiment analysis for arabic tweets. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 71–78.

Anna Rogers, Alexey Romanov, Anna Rumshisky, Svitlana Volkova, Mikhail Gronas, and Alex Gribov. 2018. Rusentiment: An enriched sentiment analysis dataset for social media in russian. In *Proceedings of the 27th international conference on computational linguistics*, pages 755–763.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2013. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. 2021. Evaluating deep learning approaches for covid19 fake news detection. *arXiv preprint arXiv:2101.04012*.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727.

6 Appendix

This section lists some sample annotated tweets as shown in Table 3.

S.No.	Tweet	English Translation	Tag
1	विधान परिषदेचे उपसभापती आणि प्रदेश काँग्रेस कमिटीचे माजी अध्यक्ष आमचे लाडके नेते आ.माणिकराव ठाकरे यांना वाढदिवसानिमित्त हार्दिक शुभेच्छा	A very happy birthday to the deputy chairman of the Assembly, former president of Regional Congress Committee and our beloved leader Mr. Manikrao Thackrey	1
2	९ August ला मराठा काय असतो ते सरकारला कळेल. मराठा झुकता नहीं झुकाता है! खोट्या सरकारला झुकावेच लागेल.लढाई न्यायाची,हक्काची	Government will come to know the power of Maratha community. Marathas do not bend but make others bend. This fake Government will have to accept defeat. This is a battle of justice and our rights!	1
3	खा.अशोक चव्हाण यांच्या नेतृत्वाखाली @INCIndia चा जनसागर आज जनआक्रोश हल्ला बोल मोर्चात सहभागी झाला. बहुसंख्य कॉंग्रेस कार्यकर्त्यांनी मोर्चा यशस्वी करण्यासाठी सिंहांचा वाटा उचलल्याबाबत हार्दिक अभिनंदन!फडणवीस सरकारच्या पायाखालची जमीन हादरली. काउंट डाऊन सुरू.	Under the guidance of MP Ashok Chavan, the sea of @INCIndia supporters participated in a ‘Halla Bol’ demonstration. A very hearty congratulations to majority of the congress workers who gave their lion’s share to make this demonstration successful. The ground under Fadanvis Government is shaken. Count down begins!	1
4	सलग तिस-या वर्षी महाराष्ट्र भ्रष्टाचारात पहिला का आहे ? याचे उत्तर चिक्की घोटाळ्याच्या चौकशीवरून लक्षात येत आहे. काँग्रेस पक्ष या भयंकर घोटाळ्याचा सतत पाठपुरावा करत राहून दोषींना शासन होत नाही तोपर्यंत स्वस्थ बसणार नाही. राज्यातले सरकार ‘क्लीन-चीटर’ सरकार आहे.	Why does Maharashtra rank first in corruption consecutively for the third year? The answer to this question unfolds from the chikki scam investigation. Congress will not remain calm until these people get punsihed. State government is a clean cheater government.	-1
5	महाराष्ट्र विधानपरिषदेचे माजी उपसभापती वसंत डावखरे यांच्या निधनाने एक अतिशय हसतमुख , अजातशत्रू आणि कर्तबगार नेता आपल्यातून निघून गेला. त्यांच्या परीवारास हे दुःख सहन करण्याची शक्ती मिळो ही प्रार्थना! भावपूर्ण श्रद्धांजली!	With the demise of former leader of Maharashtra’s Assembly Mr. Vasant Davkhare, an ever-smiling, friendly and capable leader has left us. I pray his family gets the strength to cope with this grief. My deepest condolences.	-1
6	मातीयुक्त चिक्की खाण्यायोग्य असल्याचे लॅबचे प्रमाणपत्र दाखवले आता पोषण आहारातील शियात सापडलेला मृत बेडूक पौष्टीक म्हणून सरकारने सांगावे	Chikki containing soil has been deemed fit to eat by a lab certification. Now, the government should say that a dead frog found in food served at school lunch is nutritious.	-1
7	मित्रा, यूपी मध्ये भाजपाचे 71 खासदार आहेत आणि त्यातले केंद्रात मंत्रीही आहेत	Dear friend, there are 71 MPs of BJP in UP and some of them are central ministers as well.	0
8	हा हा, स्टुडियोत वेगळे भाषण होते का?	Haha, was there a different speech at the studio?	0
9	सिंघम प्रवीण पोटे यांची शासनाने तात्काळ पोलिस दलात नियुक्ति करावी व पोलिसांना माशा मारण्याचे महत्वपूर्ण काम द्यावे	Singham Pravin Pote should be appointed in the police immediately and the police should be given the work of swatting flies.	-1

Table 3: Sample annotated tweets.