Grenze International Journal of Engineering and Technology, June Issue

MarathiSarc: A Marathi Tweets Dataset for Automatic Sarcasm Detection of Marathi Tweets

Pravin K. Patil[1]and Prof. S. R. Kolhe[2]
[1-2]School of Computer Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University Jalgaon, India
Email: patilprvin555@gmail.com, srkolhe2000@gmail.com

***Abstract*— Sarcasm Detection is a task of predicting whether the given text is sarcastic or not. Considering the challenges in detecting sarcasm in a sentiment bearing text, sarcasm detection has become one of the hot research areas in Natural Language Processing. Considerable amount of research has been done in this area for foreign languages such as English, Czech, Italian, Dutch, and Indonesian etc. Small amount of work in this area is also available for Indian languages such as Hindi, Tamil, and Bengali etc. However, Marathi being the third most popular language in India lags far behind in this area. One of the most crucial reasons for this is the absence of proper dataset. In this paper, we present MarathiSarc - a dataset of labeled Marathi tweets for sarcasm detection. This dataset contains 2400 tweets in Marathi language. Further we also discuss our dataset collection and the annotation policy. Finally we present some of the baseline sarcasm detection experiments performed on this dataset. The dataset is made publicly available at https://github.com/patilpravin91/MarathiSarc.**

***Index Terms*— sarcasm, marathi, irony, tweets, sarcasm detection, text classification, sarcasm dataset.**

I. INTRODUCTION

Today social media has become one of the most widely used platforms for the people to share their views on various issues. From sharing facts to voicing strong opinion about a particular issue, people have started expressing themselves very freely on social media platform. Over the time, Twitter has become is one such kind of social media platforms that has gained a lot popularity across the globe. Many social activists, politicians, celebrities connect directly with the people through twitter. These kinds of interactions are full of strong emotions on various issues and this can be used as a base for sarcasm detection.

English is one of the highly explored languages for sarcasm detection. This has been mainly possible because of the availability of large annotated and curated datasets. Recently, this problem has also been approached for some of the popular foreign languages like Chinese [1], Czech [3], Dutch [4], Italian [2], Indonesian [5], French, etc. [1]. Some amount of research work can also been seen for Indian languages like Hindi, Bengali Tamil etc. [6] [7] [8][10]. Though these languages did not have benchmark datasets, synthetic datasets were generated with either English translated versions or a mixed version of Indian and English language for performing the sarcasm detection experiments [11].

With fourth largest number of native speakers in India, and being morphologically so rich, Marathi is oneof the most popular regional spoken languages in India. People have now started expressing themselves purely ina particular regional language also. Still, to the best of our knowledge, there is no strong

Grenze ID: 01.GIJET.8.2.16

benchmarking resource or a well-curated Marathi dataset readily available (containing tweets in Marathi language) that can be used to analyze sarcasm detection for Marathi tweets. On the contrary, the research community has shown significant interest and efforts in preparing such datasets for other language processing tasks e.g. Sentiment Analysis [12]. In this paper, we propose a Marathi tweets dataset forthat can be used by the community for exploring sarcasm detection for Marathi language. This dataset, to the best of our knowledge, is the very first and the largest publicly available Marathi sarcasm detection dataset. We provide corresponding labels for each of the tweets in the dataset, which have been curated based on a well-defined annotation policy. We also conduct some benchmarking experiments and evaluate the performance of machine learning models on the proposed dataset for the task of automatic sarcasm detection.
Our major contributions include:

- We present a strong corpus of Marathi Tweets whichis publicly available and can be readily used for automatic sarcasm detection of Marathi tweets.
- We annotate the cleaned tweets by human annotatorsbased on a well-defined annotation policy and label them as either sarcastic or non-sarcastic.
- We conduct baseline experiments using state-of-the-art language models and machine learning classifiersto perform a benchmarking study on the proposed dataset.

II. Related Work

Adequate availability of English datasets has made it possible for the researchers to explore sarcasm detection for English language. But there are some datasets available in few non-English languages which has helped there searchers to carry out some work for these language. Lieu et.al [1] proposed a Chinese dataset and carried out sarcasm detection in social media based on imbalance classification. Ptacek et al. 2014 [3], created a large Czech Twitter corpus consisting of 7,000 manually-labeled tweets and had a first attempt at sarcasm detection in the Czech language. Barbieri et al. 2014 [2], proposed a corpus of 25,450 Italian tweets consisting of sarcastic and non-sarcastic tweets. The sarcastic tweets were collected from the posts of Twitter accounts of Spinozait and LiveSpinoza which are popular Italian blogs for sarcastic posts on politics and the non-sarcastic tweets were taken from twitter account of some famous Italian newspapers. Charalampakis et al. 2016 [13], proposed a dataset containing 61.427Greek tweets related to Parliamentary elections of Greece. Lunando and Purwarianti 2013 [14], proposed a dataset of 980 Indonesian tweets gathered manually from the twitter. Liebrecht et al. 2013 [15], proposed a technique where they collected 78000 dutch tweets containing *#sarcasme* which is a dutch word for sarcasm. They classified tweets as a sarcastic or non-sarcastic using the Winnov classification technique.
Coming to the sarcasm detection for English Language, Mikhail Khodak et.al,(2018) [16] proposed a large corpus of 1.3 million tweets that were annotated by the author including the topic, user and context of the conversation for each tweet. This dataset proved useful for both balanced and unbalanced classification. Filatova (2012) [17] came up with an experiment of creating a corpus consisting of sarcastic and non-sarcastic reviews on Amazon products and services. Abercrombie and Hovy (2016) [18] created an unbalanced dataset of 2,240 Twitter conversations annotated manually. Their experiments indicated that imbalance of class and manual labeling both should be taken into consideration while sarcasm detection. Riloff et al. (2013) [19] proposed a technique where the first half of the dataset is labeled by using hashtags and second half of the dataset is annotated manually by humans. Using the bootstrapping algorithm, sarcasm was identified automatically. A similar approach is employed by Van Hee et al. (2018) [20].It is a balanced dataset of 4,792 tweets.
As far as Indian regional languages are concerned, SK Bharati et.al,(2017) [21], proposed a dataset of five thousand single line Hindi news on recent topics and developed a automatic sarcasm detection system for Hindi news. Sahil et.al [11], proposed a novel dataset consisting of mixture of English and Hindi tweets, scripted in English language.

III. DATASET Collection

So far, there have been two prominent approaches used for labelling the tweets for sarcasm detection task:

- **Manual Annotation**: In this method, the tweets are collected and presented to the human annotators for labelling. Filatova et.al [17], used a group of human annotators to label the tweets related to Amazon review as sarcastic or non-sarcastic. Out of all the tweets, 486 were labelled as sarcastic and 844 were labelled as non-sarcastic. Abercrombie and Hovy [18] labelled 2,240 tweets, out of which 448 were

positive and 1732 were negative labels. The drawback of this technique is that with the increase in the size of the dataset, it becomes very time consuming and tedious to use this technique for labelling.

- **Hashtag based Technique**: The second and by far, the most popular technique is the use of hashtags to label the tweets as sarcastic or non-sarcastic. Tweets containing specific hashtags such as #sarcasm, #irony etc is considered as sarcastic while those not containing such hashtags are considered to be non-sarcastic. The advantage of this technique is that, it helps us to create a large dataset very quickly. Also only the author of the tweet can give assure whether the tweet being really sarcastic or not.

But using hashtags based technique also has a disadvantage. The quality of data used for training here may get altered due to inappropriate used of hashtags.

Non sarcastic	Sarcastic
साहेब, एका मराठी माणसाचा जीव गेलाय,एक मराठी महिला विधवा झालीय, एका मराठी मुलीचं छत्र हरपले आहे.आणि तुम्ही #sarcasm च्या गोष्टी करता . असा होईल का महिला सबलीकरण..	आत्मनिर्भर च्या गोष्टी करणारे भक्ताड #SputnikV च्या स्वागतासाठी सतरंज्या टाकून रेडी हाय #sarcastic

Figure 1. Example of Sarcastic and Non Sarcastic Tweet

In Fig 1 both the tweets contain the hashtags *#sarcasm*, but only one of them is found to be actually sarcastic. So to overcome this, the right strategy would be to use the combination of both hashtag-based supervision along with manual annotation.

Considering the limitation of Twitter API, we preferred to use the Twint library of twitter for collecting the tweets. Using this, we were able to collect 2361 tweets in Marathi language. In the first stage, using the hashtag based supervision technique we collected Marathi tweets containing hashtags such as *#sarcasm, #sarcastic, #sarcasmic #irony, #ironic* etc. The time period of the corpus is from December 2011 to July 2021. As a sanity check, we remove the duplicate tweets from the collection.

IV. Annotation Policy

We have manually labelled the entire dataset into three classes as follows:

- Tweets that contained the hashtags such as *#sarcasm, #sarcastic, #sarcasmic, #irony, #ironic, #**व्यंग**,* and found to be actually sarcastic are labelled as **sarcastic**.
- Tweets that contained the hashtags such as *#sarcasm, #sarcastic, #sarcasmic, ,#irony, #ironic, #**व्यंग**,* but are found to be actually non sarcastic are labelled as **non-sarcastic**.
- Tweets which can be possibly sarcastic depending on the conversational history and the context are marked as **possibly sarcastic**

TABLE I. NUMBER OF SAMPLES IN EACH CATEGORY

Sarcastic	Non Sarcastic	Unsure	Total Tweets
647	1477 Table	276	2400

As part of the annotation policy we assured the following things:

- We ensured that we do not consider the author name while labelling the tweet to avoid a human bias.
- Applying hashtag based supervision; all the collected tweets were containing either of the hashtags such as as *#sarcasm, #sarcastic, #sarcasmic, #irony, #ironic, #**व्यंग**.*
- Out of these collected tweets, as part of manual annotation, only the tweets that were found to be actually sarcastic are labelled as sarcastic.
- Though, all the collected tweets were containing the hashtags such as *#sarcasm, #sarcastic, #sarcasmic, #irony, #ironic, #**व्यंग**,* but the tweets which didn't demonstrate sarcastic emotions were labelled as non-sarcastic.
- There were few tweets about which we were unsure about while classifying them as sarcastic or non-sarcastic. However, after referring the history of the conservation, we may classify such tweets. Such tweets are currently labelled as possibly sarcastic or unsure.

V. Dataset Analysis

A. Class Distribution

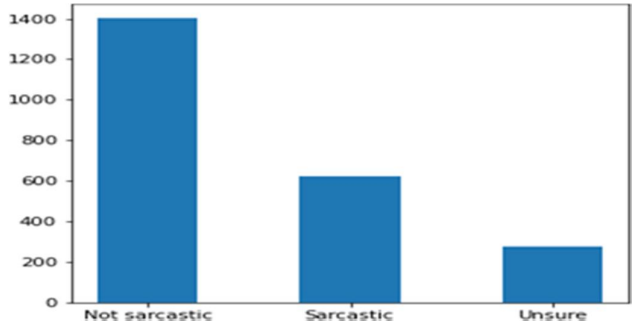

Figure 2. Class Distribution

In Fig 2, the class distribution analysis after manual annotation is shown. From Table 1, we can say that, The skewness of the dataset shows that out of all the 2400 tweets containing hashtags such as *#sarcasm, #sarcastic, #sarcasmic, ,#irony, #ironic, #**व्यंग***, only 647 were found to be actually sarcastic. This proves hashtag based supervision technique alone is not enough to collect a good quality dataset. In combination with hashtag, manual annotation is necessary for building quality models.

B. Hashtag Analysis

Fig 3 shows the top 10 frequently occurring hashtags from the identified sarcastic tweets. From the statistics,it is evident that no common topic is found in the tweets. This means that the dataset is not biased for any topic, which will help in learning general Marathi sarcastic semantics.

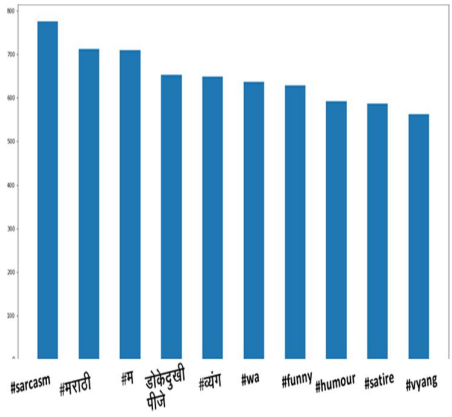

Figure 3. Top 10 frequently occurring hashtags

Hashtag	Number of Occurrences
#sarcasm	776
#मराठी	712
#म	710
#डोकेदुखीपीजे	653
#व्यंग	649
#wa	636
#funny	629
#humor	592
#satire	587
#vyang	562

Figure 4 Occurrences of top 10 hashtags

C. Emoji Analysis

Fig 5 shows top ten frequently used emojis in the identified sarcastic tweets. From the dataset, it is observed that most of the tweets contain emojis. This can be also be potentially used as one of the discriminative features for sarcasm detection task. This can be explored as a future work using this dataset.

VI. Experimental Results

In this section, we look at the baseline experiments performed on the collected dataset and the performance achieved using different machine learning classification algorithms. After removing duplicates, empty-valued and invalid tweets, we have a total of 2361 tweets out of which 647 are labeled as sarcastic whereas 1714 are labeled as non-sarcastic. To conduct baseline classification analysis, we performed two class

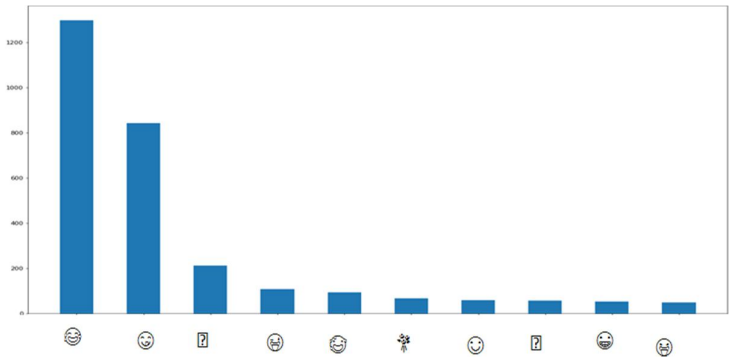

Figure 5. Top 10 frequently occurring Emojis

Emojis	Number of Occurrences
😂	1299
😊	843
😭	212
😍	109
😩	95
😒	67
🙂	60
😉	58
😁	53
😔	50

Figure 6 Occurrences of top 10 Emojis

classification experiments on our dataset. As part of preprocessinsteps, the following operations were performed –

- Removal of stopwords: Stop words occur in abundance, hence providing little to no unique information that can be used for classification. So from the collected tweets we removed all such stop words. To do so we first prepared a list of stop words. Then all the words of a particular tweet were comparedwith the words in the list of stopwords. If the tweet contains the words that are present in the stopwords list, such words are removed.
- Removal of Twitter Handles: The name of the twitter handle does not play any role for detecting the sarcasm of the tweet. So we removed the names of the twitter handle present as part of the tweets.
- Removal of HTML tags and hyperlinks: Most of the tweets contain some HTML tags and hyperlinks to some other tweets or related information. These HTML tags and hyperlinks do not add anything to the semantics of the tweets. So we removed such hyperlinks and HTML tags.

- Removal of special characters: Special characters do not add anything to the semantics of tweet. So these special characters do not have any significance in detecting the sarcasm. So we also removed such special characters present in the tweet.
- Removal of duplicate tweets: We also removed the redundant tweets that were scraped redundantly.

These preprocessed cleaned tweets were then subsequently used for further analysis. To obtain word embeddings, we used ULMFit [22], an LSTM-based model, as the language encoding model for our experiments. A ULMFit model pretrained for Marathi language on Marathi Wikipedia articles has been made publicly available byiNLTK [22]. For each tweet, we get a vector embedding which can be then used for further analysis.

To perform baseline classification experiments, we use three classical machine learning models - Gradient-boosted decision tree, Random Forest, and SVM with radial basisfunction (RBF) kernel, and compare their performance based on AUROC curve metric. Since we have a very limited dataset, we use k-fold cross validation to report the performance numbers. To perform these experiments, we use scikit-learn[24] and XGBoost[23] libraries for their respective algorithm implementations. The performance for each of these algorithms is shown below. We observed that XGBoost algorithm outperformed the other two classifiers and achieved an AUROC of 0.87 on our dataset.

TABLE II. RESULTS FROM BASELINE EXPERIMENTS

Sr. No	Classifier	Average AUROC	Average F1 Score
1	XGBoost	0.87	0.65
2	SVM	0.85	0.63
3	Random Forest	0.86	0.60

VII. CONCLUSION

In this paper, we have presented MarathiSarc, which to the best of our knowledge, is the first major Marathi tweets-based dataset for sarcasm detection of Marathi tweets. Our dataset consists of approximately 2400 Marathi tweets. We also describe the annotation policy which we used for manual labelling of the entire dataset. We performed two class classification experiments on our dataset using classical machine learning models - Gradient-boosted decision tree, Random Forest, and SVM with radial basis function (RBF) kernel, and compare their performance based on AUROC curve metric. We observed that XGBoost algorithm outperformed the other two classifiers and achieved an AUROC of 0.87 on our dataset. Our dataset will really have a vital role as far as research in natural language processing tasks for Indian regional languages is concerned. In future, we intend to perform well-informed feature engineering on the existing dataset to study the impact of various important features that can be extracted from the existing data and possibly improve the sarcasm detection performance.

REFERENCES

[1] Peng Liu, Wei Chen, Gaoyan Ou, Tengjiao Wang, Dongqing Yang, and Kai Lei. 2014. Sarcasm Detection in Social Media Based on Imbalanced Classification. In Web-Age Information Management. Springer,459–471.

[2] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2014a. Italian irony detection in twitter: a first approach. In The First Italian Conference on Computational Linguistics CLiC-it 2014 the Fourth International Workshop EVALITA. 28–32

[3] Tomas Ptacek, Ivan Habernal, and Jun Hong. 2014. Sarcasm Detection on Czech and English Twitter. In Proceedings COL- ING 2014. COLING.

[4] CC Liebrecht, FA Kunneman, and APJ van den Bosch. 2013. The perfect solution for detecting sarcasm intweets not. (2013).

[5] Edwin Lunando and Ayu Purwarianti. 2013. Indonesian social media sentiment analysis with sarcasm detection. In AdvancedComputer Science and Information Systems (ICACSIS), 2013 International Conferenceon. IEEE, 195–198.

[6] Piyush Arora. 2013. Sentiment analysis for hindi language. MSby Research in Computer Science.

[7] 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 preprintarXiv:1803.06745.

[8] 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.

[9] Sandeep Sricharan Mukku and Radhika Mamidi. 2017. Actsa: Annotated corpus for telugu sentiment analysis. In

Proceedingsof the First Workshop on Building Linguistically GeneralizableNLP Systems, pages 54–58.
[10] Nadana Ravishankar and Shriram Raghunathan. 2017. Corpusbased sentiment classification of tamil movie tweets using syn- tactic patterns. IIOAB Journal: A Journal of Multidisciplinary Science and Technology,8(2):172–178.
[11] Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed S. Akhtar, Manish Shrivastava A Corpus of English-Hindi Code-Mixed Tweets for Sarcasm Detection 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2018)
[12] Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi, L3CubeMahaSent: A Marathi Tweet-based Sentiment Analysis Dataset, Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 213–220 April 19, 2021. ©2021 Association for Computational Linguistics
[13] Basilis Charalampakis, Dimitris Spathis, Elias Kouslis, and Katia Kermanidis. 2016. A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. Engi- neering Applications of Artificial Intelligence (2016), DOI:http://dx.doi.org/10.1016/j.engappai.2016.01.007
[14] Edwin Lunando and Ayu Purwarianti. 2013. Indonesian social media sentiment analysis with sarcasm detection. In AdvancedComputer Science and Information Systems (ICACSIS), 2013 International Conference on. IEEE, 195–198.
[15] CC Liebrecht, FA Kunneman, and APJ van den Bosch. 2013. The perfect solution for detecting sarcasm intweets# not. (2013).
[16] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In LREC. ELRA.
[17] Elena Filatova. 2012. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing.. In LREC. 392–398.
[18] Gavin Abercrombie and Dirk Hovy. 2016. Putting Sarcasm Detection into Context: The Effects of Class Imbalance and Manual Labelling on Supervised Machine Classification of Twit- ter Conversations. ACL 2016 (2016), 107.
[19] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as Con- trast between a Positive Sentiment and Negative Situation.. In EMNLP. 704–714.
[20] Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In SemEval, pages 39–50. ACL
[21] S. K. Bharti, K. S. Babu, and S. K. Jena. 2015. Parsing based sarcasm sentiment recognition in twitter data. In ASONAM, pages 1373–1380. ACM.
[22] Gaurav Arora. 2020. inltk: Natural language toolkit for indic languages. arXiv preprint arXiv:2009.12534.
[23] Chen, Tianqi and Guestrin, Carlos, “XGBoost: A Scalable Tree Boosting System” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[24] Pedregosa, Fabian and Varoquaux, and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent, “Scikit-learn: Machine learning in Python”, Journal of Machine Learning Research,2011.