GSAC: A Gujarati Sentiment Analysis Corpus from Twitter

Monil Gokani and **Radhika Mamidi**
Language Technologies Research Center (LTRC)
Kohli Center on Intelligent Systems
International Institute of Information Technology, Hyderabad
monil.gokani@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract

Sentiment Analysis is an important task for analysing online content across languages for tasks such as content moderation and opinion mining. Though a significant amount of resources are available for Sentiment Analysis in several Indian languages, there do not exist any large-scale, open-access corpora for Gujarati. Our paper presents and describes the Gujarati Sentiment Analysis Corpus (GSAC), which has been sourced from Twitter and manually annotated by native speakers of the language. We describe in detail our collection and annotation processes and conduct extensive experiments on our corpus to provide reliable baselines for future work using our dataset.

1 Introduction

Sentiment Analysis is an NLP task that involves identifying the sentiment or the polarity (such as positive vs negative) of a piece of text (Pang et al., 2008). It has received significant attention in recent years, with ever-increasing internet access and social media usage, even in Indian languages such as Hindi (Akhtar et al., 2016; Patra et al., 2018) and Marathi (Kulkarni et al., 2021a) which are typologically and geographically close to Gujarati. However, there is hardly any work done in Gujarati itself.

Gujarati is a very prominent language in Western India, with over 55 million first-language speakers and a significant presence in the states of Gujarat, Maharashtra, and Rajasthan (Eberhard et al., 2023). It is also the official language of the state of Gujarat. Despite a large online community active on social media and a significant mainstream media presence, there is a lack of large-scale, publicly available resources for sentiment classification (see Section 2).

Hence, we describe a new, gold-standard, manually annotated Gujarati Sentiment Analysis Corpus (GSAC) for monolingual sentiment classification. The dataset is sourced from Twitter and labelled by native speakers. We describe our annotation process and also run extensive experiments on the dataset using feature-based and deep-learning architectures to establish a reliable baseline for GSAC and compare the performances of various model architectures. The dataset is available on GitHub.[1]

2 Related Work

Significant work has been done on coarse-grained and aspect-based sentiment analysis (SA) in various Indian languages. Datasets have been created for SA in Hindi (Akhtar et al., 2016; Patra et al., 2018), Telugu (Mukku and Mamidi, 2017), Marathi (Kulkarni et al., 2021b), Bengali (Islam et al., 2021; Patra et al., 2018) and Tamil (Jenarthanan et al., 2019), and Tamil and Malyalam (Chakravarthi et al., 2021). However, SA in Gujarati has been scarcely explored, and no standard, publicly available dataset exists.

One of the earliest works in SA in Gujarati was by Joshi and Vekariya (2017), who used a POS tag-based feature set for an SVM classifier on a small sample of 40 tweets. Since then, Gohil and Patel (2019) developed and experimented with a Gujarati SentiWordNet to classify tweets, creating a Twitter dataset with 1120 samples. Other approaches included scraping movie-review websites to create a dataset (Shah and Swaminarayan, 2021; Shah et al., 2022a), even translating reviews from English to Gujarati to expand the dataset (Shah and Swaminarayan, 2022; Shah et al., 2022b). Mehta and Rajyagor (2021) attempted classifying a set of 300 poems into nine different emotional categories using machine learning-based approaches. However, none of the datasets used in these experiments have been released to open access, which makes it difficult to reproduce any of these results or compare the performance of new models with

[1] https://github.com/MG1800/gsac

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 129–137
July 14, 2023 ©2023 Association for Computational Linguistics

Work(s)	Source	Size	Annotation	Open Access
(Joshi and Vekariya, 2017)	Twitter	40	Manual	No
(Mehta and Rajyagor, 2021)	Poems	300	Manual	No
(Gohil and Patel, 2019)	Twitter	1120	Manual	No
(Shah et al., 2022a), (Shah and Swaminarayan, 2021)	Movie Reviews	500	Manual	No
(Shah and Swaminarayan, 2022), (Shah et al., 2022b)	Movie Reviews (Gujarati + translated from English)	2085	Automated, based on website rating	No
GSAC	**Twitter**	**6575**	**Manual**	**Yes**

Table 1: Comparison of previous datasets on Gujarati Sentiment Analysis with our dataset - GSAC

them.

Gujarati was a part of the set of languages included in the training data for XLM-T (Barbieri et al., 2022), a highly multilingual effort for creating a unified Twitter-based language model for sentiment classification. However, Gujarati was not a part of the monolingual evaluation reported by the authors. Additionally, Gujarati has been included in some research on multilingual lexical level sentiment classification (Zhao and Schütze, 2019; Buechel et al., 2020).

Efforts in dataset creation for Sentiment Analysis have been varied. We mainly focused on Twitter datasets or datasets in Indian languages for reference when deciding our annotation process. Jenarthanan et al. (2019) created a Twitter-based emotion classification dataset in Tamil and English and used a set of emotion words as queries for collecting tweets, an approach that we also use for collecting our data. Mukku and Mamidi (2017) classify sentences from a news corpus into three sentiment categories - positive, negative, and neutral, similar to what we aim for, and hence are a good source of reference for annotation guidelines. We also refer to Muhammad et al. (2022), which is a more recent effort at creating a sentiment classification dataset for resource-poor languages, collecting and annotating a dataset for 4 African languages with multiple human annotators.

Table 1 compares our dataset to the existing SA datasets in Gujarati.

3 Dataset Creation

The dataset was created in two main steps - collecting and sampling the dataset from Twitter to create a subset for annotation and getting the data annotated by native speakers, which included creating the annotation guidelines and training them for the task.

3.1 Collection

We source our data from Twitter, which has a large active user base of Gujarati speakers. We scraped the initial dataset using Twitter API [2], which supports filtering the results for Gujarati using the language tag. We also used the API parameters to exclude retweets and quotes, to reduce the number of duplicates in our dataset. To ensure we had a desirable mix of sentiments in the dataset, the search queries were based on a hand-picked subset of sentiment words [3] based on a machine-translated English sentiment lexicon (Chen and Skiena, 2014). We chose a subset so as to remove words that were either not translated or translated incorrectly in the list, selecting ~250 words. The start times are varied to ensure the tweets are spread out over time, with the final set having tweets ranging from August 2010 to February 2022. We then preprocessed, filtered, and sampled from this large dataset to generate subsets for each of our annotators to label. The complete process we followed is described below:

1. Create a list of prompts by hand-picking samples from machine-translated sentiment vocabulary.

2. Scrape tweets using these prompts using Twitter API, using the API parameters to ensure collected tweets are in Gujarati script, spread out over several years, and do not include any retweets or quotes.

3. Preprocess these tweets, normalising whitespaces and newlines, lower-casing, and replacing all user mentions and URLs with the tokens `@user` and `<url>` respectively.

[2] https://developer.twitter.com/en/docs/twitter-api

[3] https://www.kaggle.com/datasets/rtatman/sentiment-lexicons-for-81-languages

4. Drop any tweets with identical text or fewer than 10 tokens after preprocessing. This step eliminated a significant number of gibberish tweets that were not useful for the task, such as the one shown in row 4 of Figure 1.

5. Randomly sampled 10% of the tweets for each prompt to create a subset of approximately 22,000 samples from the larger set that retained the same distribution as the original set.

6. From this smaller representative subset, we randomly sampled 7,000 tweets for annotation based on the annotation resources available to us.

The statistics for this process are provided in Table 2. We labelled approximately 7000 tweets from the representative set, with the final dataset containing 6,575 tweets after dropping undesirable samples as described in Section 3.2.

Stage	Count
Initial set from scraping	320,978
Filtering out duplicates	247,226
Dropping tweets with <10 tokens	226,482
Representative Set after Sampling	22,630
Annotated	6,575

Table 2: Data collection statistics

3.2 Annotation

We first developed the annotation schema and tested it by annotating a small sample of the dataset ourselves. Once the dataset was finalised, we recruited four annotators and trained them over several rounds of labelling and discussion before providing them with independent subsets to annotate.

3.2.1 Annotation Schema

We classified each tweet in our dataset as `positive`, `negative`, or `neutral`. We also gave our annotators an `unfit` tag for tweets that they think cannot be used for the task. We define each of the labels as follows:

- `positive` - Tweets were classified as positive if they expressed a positive sentiment about some subject (a product or a movie, for example) or if they showed support for a subject, such as a person or a policy. Tweets about events inherently associated with positive sentiments (such as reporting a sports team's victory) are also labelled positive.

- `negative` - Tweets that expressed a negative opinion about a subject (such as criticising a policy or an official) were labelled negative. Tweets talking about events with an inherently negative connotation - such as reporting the death of a celebrity or the loss of a sports team, and tweets containing any kind of derogatory remarks or threats towards a subject were also labelled negative.

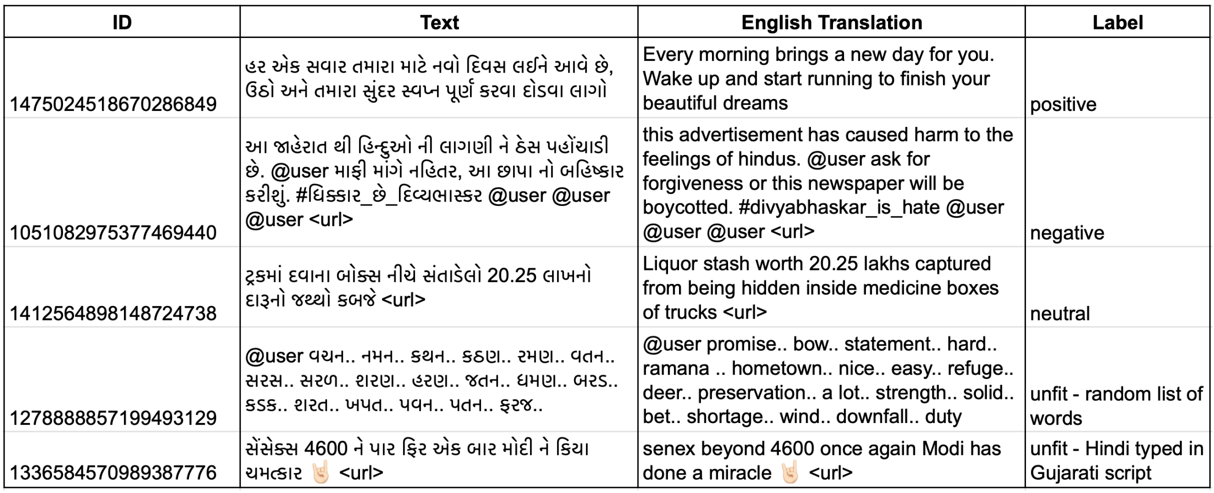
- `neutral` - Tweets were labelled as neutral in two cases - if they contained no sentiment about the subject or if they contained a mix of both positive and negative sentiment about a subject (such as praising one aspect but criticising another of a product).

- `unfit` - Tweets were marked unfit if the annotator could not assign one of the three labels to it. This happened in several cases, such as cases where it was a different language tweet that was typed in Gujarati script, or there was not enough context in the tweet to label it (if it required a media attachment to understand, for example). Any tweets marked unfit by any of the annotators were dropped from the dataset.

Figure 1 illustrates some of the tweets and their labels, along with an approximate English translation of the tweet.

3.2.2 Annotation Process

We manually annotated 7000 samples across four annotators. The annotators were linguistics students who were native speakers of Gujarati, aged between 19 and 23. The annotators were trained for the task over three rounds of annotation on small subsets of 50 tweets each, followed by a session of doubt clarification and discussion after every round. To measure the annotation quality, we calculate inter-annotator agreement using Fleiss' Kappa coefficient (Fleiss, 1971). Over the three rounds of training, it improved from 0.48 to 0.52 and finally to 0.58, which suggests moderately strong agreement. The tweets used for these training rounds were discarded and not included in the final dataset. Each annotator then labelled data in subsets of 500 samples.



ID	Text	English Translation	Label
1475024518670286849	હર એક સવાર તમારા માટે નવો દિવસ લઈને આવે છે, ઉઠો અને તમારા સુંદર સ્વપ્ન પૂર્ણ કરવા દોડવા લાગો	Every morning brings a new day for you. Wake up and start running to finish your beautiful dreams	positive
1051082975377469440	આ જાહેરાત થી હિન્દુઓ ની લાગણી ને ઠેસ પહોંચાડી છે. @user માફી માંગે નહિતર, આ છાપા નો બહિષ્કાર કરીશું. #ધિક્કાર_છે_દિવ્યભાસ્કર @user @user @user <url>	this advertisement has caused harm to the feelings of hindus. @user ask for forgiveness or this newspaper will be boycotted. #divyabhaskar_is_hate @user @user @user <url>	negative
1412564898148724738	ટ્રકમાં દવાના બોક્સ નીચે સંતાડેલો 20.25 લાખનો દારૂનો જથ્થો કબજે <url>	Liquor stash worth 20.25 lakhs captured from being hidden inside medicine boxes of trucks <url>	neutral
1278888857199493129	@user વચન.. નમન.. કથન.. કઠણ.. રમણ.. વતન.. સરસ.. સરળ.. શરણ.. હરણ.. જતન.. ધમણ.. બરડ.. કડક.. શરત.. ખપત.. પવન.. પતન.. ફરજ..	@user promise.. bow.. statement.. hard.. ramana .. hometown.. nice.. easy.. refuge.. deer.. preservation.. a lot.. strength.. solid.. bet.. shortage.. wind.. downfall.. duty	unfit - random list of words
1336584570989387776	સેંસેક્સ 4600 ને પાર ફિર એક બાર મોદી ને કિયા ચમત્કાર 🤘🏻 <url>	senex beyond 4600 once again Modi has done a miracle 🤘🏻 <url>	unfit - Hindi typed in Gujarati script

Figure 1: Some samples from the GSAC dataset

3.3 Statistics

Our final dataset contains a total of 6575 tweets after dropping the tweets labelled unfit. We divide the dataset into training, development, and test sets in a 70:10:20 ratio, respectively. Within the complete dataset, the `neutral` class has the highest representation, comprising about 45.12% of the total dataset, followed by `positive` at 30.05% and finally `negative` at 24.83%. Additional details about the class distribution are reported in Table 3.

The average word count for the combined dataset is 27.77, with a standard deviation of 13.86. The average word count (excluding whitespaces) is 136.07, with a standard deviation of 67.55, as shown in Table 4, which also reports the same values for each class. Figures 2 and 3 illustrate the class-wise and split-wise distribution of word counts in the dataset, respectively.

Split	Positive	Neutral	Negative	Total Count
Train	1374	2100	1128	4602
Dev	201	287	163	651
Test	401	580	341	1322
Total	1976	2967	1632	6575

Table 3: Split-wise Class Distribution of Dataset

Split	Tokens	Characters
Positive	27.86 (11.65)	141.79 (58.92)
Neutral	27.25 (15.52)	132.39 (77.02)
Negative	28.60 (13.07)	135.62 (57.82)
Overall	**27.77 (13.86)**	**136.01 (67.55)**

Table 4: Mean Token and Character Counts for each label (brackets contain standard deviation)

4 Experiments

We train two sets of models to test how different models perform on our dataset and to set baselines for it. The first set of models consists of feature vector-based models, which we train on two different variants based on different sets of features - Bag-of-Words and TF-IDF. The second set is a set of deep contextualised models, where we fine-tune various transformer-based pre-trained language models for classification on this dataset.

4.1 Feature Vector Models

We train five classifiers - Naive Bayes, Logistic Regression, Support Vector Machines, Random Forests, and a Multi-Layer Perceptron - each on two different feature vectors - Bag-of-Words and TF-IDF for a total of 10 models.

Bag-of-Words (BoW) or Count Vectorizer represents a document (in this case, a tweet) as a vector of the counts of each word present in the document. Even though it ignores word order, bag-of-words features can still be useful as feature vectors for tasks such as text classification (McCallum and Nigam, 2001).

TF-IDF (Term Frequency - Inverse Document Frequency) (Spärck Jones, 1972) is a method to represent documents that factors in the relative frequency of a word across documents by calculating a score based on two parameters - term frequency, which is the frequency of a term in the current document, and inverse document frequency - which is based on the frequency of the term across all documents.

The models we train for each of these are:

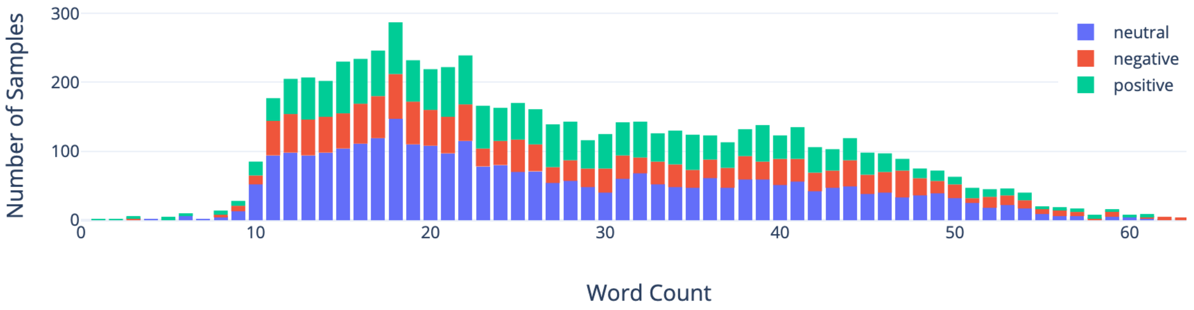


Figure 2: Class-wise distribution of word counts

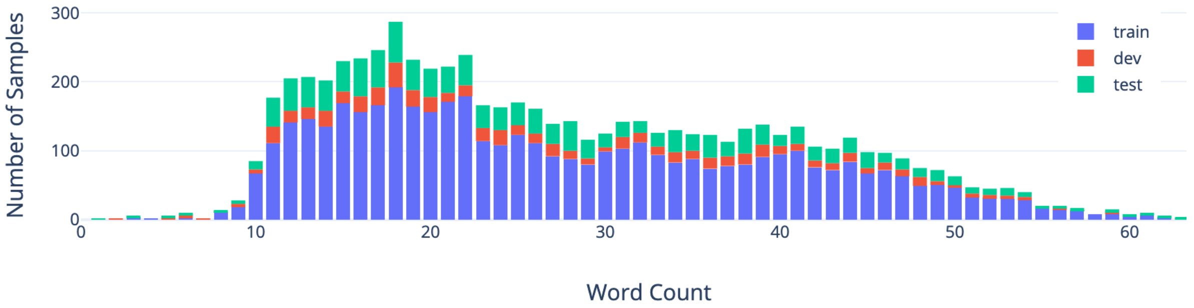


Figure 3: Split-wise distribution of word counts

- **Naive Bayes Classifier** - A Naive Bayes classifier is a simple classifier that estimates the probability of each label under the assumption of input features being conditionally independent, which has been shown to perform well on text classification (McCallum and Nigam, 2001). We train the classifier for 200 epochs or until convergence.

- **Logistic Regression** - Logistic regression (Cox, 1958) is a classification algorithm that estimates a logistic function to calculate the probability of an input feature belonging to a certain class. We train an LR classifier over 100 epochs or convergence using a one-vs-all approach.

- **Support Vector Machine** - A support vector machine (Cortes and Vapnik, 1995) is a classifier that tries to find the hyper-plane that most optimally divides the training data according to the labels. This is also trained using a one-vs-all approach, over 200 maximum epochs.

- **Random Forests** - Random Forests (Breiman, 2001) are a type of ensemble classifier that use a large number of decision trees (set to 100 for our model), each using a subset of the input features and training data, to estimate the most likely label for the given input.

- **Multi-Layer Perceptron** is a simple feed forward neural network (Rosenblatt, 1958; Rumelhart et al., 1986). Our model uses a single 100-dimension hidden layer, with a ReLU activation, for 300 maximum epochs.

We use the scikit-learn python library (Pedregosa et al., 2011) to create feature vectors from the text and train and test this set of models.

4.2 Deep Contextualised Models

Multilingual transformer-based language models trained on multiple languages such as BERT (Devlin et al., 2018) and RoBERTa (Conneau et al., 2019) have been shown to perform well on downstream tasks (Pires et al., 2019). We fine-tune the following language models on our dataset:

- **Multilingual BERT** - mBERT is a multilingual version of BERT (Devlin et al., 2018), and is a language model trained on the top 100 languages with the largest Wikipedia corpora, which includes Gujarati. We use the `bert_base_multilingual_uncased` version of BERT.

- **XLM-RoBERTa** - XLM-RoBERTa is a multilingual version of RoBERTa (Conneau et al., 2019), which is itself a more optimised version of BERT, trained on a larger dataset, and

Model	Precision	Recall	Accuracy	Weighted F1	Macro F1
Bag of Words					
Naive Bayes	0.59	**0.58**	**0.58**	**0.57**	**0.56**
Logistic Regression	0.55	0.55	0.55	0.55	0.54
SVM	0.55	0.52	0.52	0.49	0.46
Random Forests	**0.61**	**0.58**	**0.58**	0.55	0.53
MLP	0.52	0.52	0.52	0.52	0.51
TF-IDF					
Naive Bayes	**0.66**	0.52	0.52	0.43	0.38
Logistic Regression	0.58	**0.57**	**0.57**	**0.56**	**0.55**
SVM	0.57	0.56	0.56	0.54	0.53
Random Forests	0.59	0.55	0.52	0.50	0.50
MLP	0.50	0.50	0.50	0.50	0.49
Pretrained LMs					
mBERT	0.38	0.51	0.51	0.43	0.38
XLM-RoBERTa	0.41	0.52	0.52	0.43	0.39
XLM-T	0.64	0.62	0.63	0.64	0.63
GujaratiBERT	0.64	0.64	0.64	0.64	0.64
IndicBERT	**0.65**	**0.67**	**0.66**	**0.66**	**0.66**

Table 5: Results of the various models on the test set. **Bold** indicates best score for each set of models. **Underline** indicates best score across all models.

a modified training task. XLM-RoBERTa also includes Gujarati as a part of its training set. We fine-tune the `xlm-roberta-base` variant of the model.

- **XLM-T** - XLM-T (Buechel et al., 2020)is a variant of XLM-RoBERTa that was trained on a Twitter dataset consisting of 198M tweets in a large set of languages, including over 10,000 samples in Gujarati. It was further finetuned for Sentiment Classification on a set of 8 languages, which included Hindi, which is closely related typologically to Gujarati. We further finetune the `twitter-xlm-roberta-base-sentiment` variant of the model on HuggingFace.

- **GujaratiBERT** - GujaratiBERT (Joshi, 2022) is an mBERT (base variant) model that has been fine-tuned for Gujarati using publicly available monolingual Gujarati corpora. Since it is specifically fine-tuned for Gujarati, we expected it to perform better than mBERT and XLM-RoBERTa.

- **IndicBERT** - IndicBERT (Kakwani et al., 2020) is an ALBERT (Lan et al., 2019) model pre-trained on a combined corpus of 12 different Indian languages (including Gujarati), which has been shown to achieve state-of-the-art performance on multiple downstream tasks in several Indian languages on the IndicGLUE benchmark (Kakwani et al., 2020), including sentiment analysis in Hindi (Akhtar et al., 2016) and Telugu (Mukku and Mamidi, 2017). We fine-tune this model for classification on our dataset.

All of our transformer models are trained for 5 epochs, with a learning rate of 4e-5 and batch size of 8. We set up our training and testing scripts using the simpletransformers (Rajapakse, 2019) library, which is based on the transformers library from HugggingFace (Wolf et al., 2020).

5 Results

We report the detailed results for each model in Table 5. We make a few observations from observing the weighted and macro F1 scores for each model:

- We observe that GujaratiBERT and IndicBERT achieve the best performance compared to all other models. This could be because compared to the rest of the pretrained language models, these two models have been trained on a significantly higher amount of Gujarati data (during pretraining for IndicBERT, and during fine-tuning for GujaratiBERT).

- mBERT and XLM-RoBERTa perform very poorly compared to other pretrained language models. This could be because they are trained on a very large set of languages, due to which Gujarati might not have sufficient representation in the corpus and the model vocabulary causing it to underperform.

- XLM-T contained only ∼10,000 samples in Gujarati out of a total ∼198M samples in its training data. However, it still achieves comparable performance to GujaratiBERT and IndicBERT. This may be because the training data for XLM-T comes exclusively from the same domain as our dataset (Twitter), which suggests pretraining or fine-tuning models on similar domain data in multiple languages can help improve model performance in low-resource languages.

- Despite not achieving the same performance as XLM-T, GujaratiBERT, or IndicBERT, the Naive Bayes model using TF-IDF features achieves the highest precision out of all the models trained. Other statistical models (such as Random Forests and Naive Bayes on both feature sets) also achieve reasonably high average precision (>= 0.59) while taking significantly less computational resources and time.

6 Conclusion and Future Work

In this paper, we present the Gujarati Sentiment Analysis Corpus (GSAC), which contains over 6500 manually annotated tweets. To the best of our knowledge, it is the first significant publicly available corpus for this task in Gujarati. We also present our annotation schema and conduct extensive experimentation to establish baselines for this new dataset. We find that pre-trained language models that included Gujarati as a part of pre-training or fine-tuning achieve better performance on this dataset compared to other models, with IndicBERT achieving the best weighted and macro F1 scores. As a part of future work, we plan to explore methods to extend this dataset automatically by using this dataset as a seed dataset to label additional data (such as by bootstrapping) or by exploring other avenues of acquiring data, such as via machine translation of existing datasets in other languages such as English or Hindi.

7 Ethical Consideration

Sentiments in a dataset sourced from social media platforms can be susceptible to inherent bias due to public opinion being biased in favour of or against certain subjects, depending on external factors like demographics. During the collection and annotation process for our dataset, we switched our collection strategy from querying tweets for particular topics (events) during the initial stages to querying them using a sentiment lexicon because we observed that the topics we queried were frequently heavily biased towards either positive or negative sentiments. The privacy of platform users is another concern that is raised when collecting data from social media. To ensure that no identifying details about any Twitter user were presented to our annotators, we removed any identifying characteristics such as user mentions and URLs from the tweets, as well as the original Tweet IDs and used internally generated IDs for the annotation process. We also only release the Tweet IDs and corresponding labels in our dataset in compliance with Twitter's data-sharing policy.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Leo Breiman. 2001. *Machine Learning*, 45(1):5–32.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. *CoRR*, abs/2005.05672.

Bharathi Raja Chakravarthi, KP Soman, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kingston Pal Thamburaj, John P McCrae, et al. 2021. Dravidianmultimodality: A dataset for multi-modal sentiment analysis in tamil and malayalam. *arXiv preprint arXiv:2106.04853*.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. *Ethnologue: Languages of the World*, 26 edition. SIL International, Dallas, Texas.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5):378–382.

Lata Gohil and Dharmendra Patel. 2019. A sentiment analysis of gujarati text using gujarati senti word net. *International Journal of Innovative Technology and Exploring Engineering*.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rajenthiran Jenarthanan, Yasas Senarath, and Uthayasanker Thayasivam. 2019. Actsea: Annotated corpus for tamil sinhala emotion analysis. In *2019 Moratuwa Engineering Research Conference (MERCon)*, pages 49–53.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Vrunda C Joshi and Vipul M Vekariya. 2017. An approach to sentiment analysis on gujarati tweets. *Advances in Computational Sciences and Technology*, 10(5):1487–1493.

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 EMNLP*.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021a. 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, Online. Association for Computational Linguistics.

Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021b. L3cubemahasent: A marathi tweet-based sentiment analysis dataset. *CoRR*, abs/2103.11408.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Andrew McCallum and Kamal Nigam. 2001. A comparison of event models for naive bayes text classification. *Work Learn Text Categ*, 752.

Bhavin Mehta and Bhargav Rajyagor. 2021. Gujarati poetry classification based on emotions using deep learning.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Abdullahi Salahudeen, Aremu Anuoluwapo, Alípio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *CoRR*, abs/2201.08277.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Braja Gopal Patra, Dipankar Das, and Amitava Das. 2018. Sentiment analysis of code-mixed indian languages: An overview of sail$_{c}ode-mixedsharedtask@icon-2017$.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

T. C. Rajapakse. 2019. Simple transformers. `https://github.com/ThilinaRajapakse/simpletransformers`.

F Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65(6):386–408.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning internal representations by error propagation.

P. Shah, P. Swaminarayan, and Maitri Patel. 2022a. Sentiment analysis on film review in gujarati language using machine learning. *International Journal of Electrical and Computer Engineering (IJECE)*.

Parita Shah and Priya Swaminarayan. 2022. Machine learning-based sentiment analysis of gujarati reviews. *International Journal of Data Analysis Techniques and Strategies*.

Parita Shah, Priya Swaminarayan, and Maitri Patel. 2022b. Sentiment analysis on film review in gujarati language using machine learning. *International Journal of Electrical and Computer Engineering*, 12(1):1030.

Parita Vishal Shah and Priya Swaminarayan. 2021. Lexicon-based sentiment analysis on movie review in the gujarati language. *Int. J. Inf. Technol. Commun. Convergence*.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mengjie Zhao and Hinrich Schütze. 2019. A multilingual BPE embedding space for universal sentiment lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3506–3517, Florence, Italy. Association for Computational Linguistics.