Lexical Resource Creation and Evaluation: Sentiment Analysis in Marathi

Mahesh B. Shelke, Saleh Nagi Alsubari, D. S. Panchal, and Sachin N. Deshmukh

Abstract In India, the raise of regional language contents over social media, websites, blogs and news article are exponentially increasing because of ease of use of technology, and people are expressing their thoughts, opinion more conveniently and powerfully over Internet. In this paper, we evaluate the challenges of sentiment analysis in Marathi by setting up a baseline, where we produced an annotated dataset, however, initially, we created an annotated dataset consisting of Marathi news scraped from various newspaper/channel websites. Furthermore, domain experts annotated Marathi news with positive, negative and neutral polarity. And we used machine learning models such as logistic regression, Stochastic Gradient Decent (SGD), support vector machine (SVM), nearest neighbour, neural network, decision tree (DT), Naïve Bayes (NB) and proposed ensemble-based model for sentiment analysis to demonstrate effectiveness. In experimentation, the proposed ensemble classifier outperforms other classifiers with an accuracy of 94.16% and an F-score of 97.02% for fivefold validation. Also, for tenfold validation, the accuracy is 95.07%, and the F-score is 96.93%.

Keywords Marathi · Sentiment dataset · Machine learning · Indian language

1 Introduction

It is very easy at present to express thoughts on the web or on social media. After viewing films, using a product or visiting a locality, we can write movie reviews, product reviews or tourist reviews. This opinion-rich data will be interested both in the decision makers for the entities concerned and in enterprises seeking to improve their products or services. It provides people the opportunity to express themselves rather than media personalities speaking on behalf of the general majority of the

M. B. Shelke (✉) · S. N. Alsubari · D. S. Panchal · S. N. Deshmukh
Department of Computer Science & IT, Dr. Babasaheb Ambedkar, Marathwada University, Aurangabad, M.S., India
e-mail: mahesh_shelke21@hotmail.com

Y.-D. Zhang et al. (eds.), *Smart Trends in Computing and Communications*, Lecture Notes in Networks and Systems 396, https://doi.org/10.1007/978-981-16-9967-2_19

population. You can make your feelings heard by posting your thoughts on the Internet. As a result, vast amounts of information, including the author's point of view, are available on the Internet. It is now a challenge that these documents contain useful information. This enhances the application of sentiment analysis/ opinion mining.

Emotions play an important role in the communication and decision-making processes in our intellectual activities. Emotions are a succession of events made up of feedback loops. Feelings and behaviours, like cognition, may aspect cognition. In addition, the detection and interpretation of emotional data are vital in a wide range of fields, including human–computer interactions, e-learning, e-Health, automotive, security, user profiling and customisation. The analysis of textual sentiment is one of the popular computer linguistic methodologies.

Sentiment analysis is a process of recognising and categorising views expressed in a piece of text in a computational way, particularly to assess whether the person has a positive, negative and neutral opinion towards any certain topic, product, etc. Product reviews, Film reviews, blog and articles are the popular and available opinion-rich contents. There are three levels of sentiment analysis can be performed: Document level, Phrase/Sentence level and Entity/Aspect level. In document level sentimental analysis, the polarity is determined for the complete document. In sentence level sentiment analysis, polarity is determined for the individual sentences of the document. In the aspect level sentiment analysis, the polarity is determined for the entity/aspect of the document. Following methods can be used for sentiment analysis:

- **Using Lexicon-based Sentiment Analysis**: It is a polarity-based information dataset of phrases or word, in which scores are assigned to each word. This score describes the characteristics related to the term as positive, negative or neutral sentiment expressed in text.
- **Using N-Gram Model-based Sentiment Analysis**: It forms and uses the N-Gram model (unigram, bigram, trigram or mixed) using the categorization trainings data.
- **Using Machine Learning-based Sentiment Analysis:** Supervised and Unsupervised learning model can be used to perform prediction on data by extracting features from text.

The main objective of this paper is to develop and evaluate lexical resources for sentiment analysis in Marathi, as there are few lexical resources, libraries, Corpus and tools available in Marathi, which signifies that Marathi has not been explored in the field of sentiment analysis. The proposed approach is designed in combination with machine learning-based algorithm. The lack of Marathi preprocessing tools and annotated dataset for training the model was the main challenge during model development. As a result, we created a preprocessing tool using the Indic NLP Library and inltk tools [1, 2], as well as collecting Marathi news from various online newspapers/channels and manually annotating and validating it with domain

experts. A comparison of all classification algorithms was also evaluated and discussed. In addition, an annotated dataset of Marathi news with sentiment orientations of positive, negative and neutral was developed [3, 4].

2 Related Work

Sentiment analysis (SA) and Opinion Mining are become popular in the field of natural language processing for Indian languages such as Hindi, Bengali, Tamil, Telugu and so on, but there are still some languages that remain unexplored due to a lack of lexical resources, such as Marathi, Gujarati, and Punjabi [5]. Three types of sentiment classification techniques are machine learning approaches, lexicon-based approaches and hybrid approaches. Machine learning approaches are classified into three types: supervised learning, semi-supervised and unsupervised learning. When there is annotated dataset unavailable to train the model then unsupervised learning is used, whereas supervised learning is used when there is a substantial annotated dataset available. The authors proposed a sentiment analysis (SA) method for Gujarati tweets that used POS tagging to extract features and SVM to classify the tweets. They have collected 40 tweets as dataset [6].

For performing sentiment analysis of Malayalam movie reviews, the authors devised a hybrid approach based on fuzzy logic. This consists of a tagging machine learning system and a fuzzy logic approach to determining review membership. TnT Tagger was also used to train datasets that had been manually tagged [7]. For Tamil and Bengali tweets, the authors developed machine learning-based sentiment analysis model based on probabilistic and decision tree classification. The amount of the dataset used, variations in writing style, and the incorrect use of punctuation marks all have an impact on the model's performance [8]. Using machine learning approaches such as support vector machine, maximum entropy, decision tree and Naïve Bayes, the author developed a system for predicting emotion in Tamil films. When it comes to accuracy, SVM outperforms other algorithms [9].

3 Proposed Methodology

This section describes the corpus creation process in detail, beginning with data collection, preprocessing, corpus annotation and inter-annotator agreement for measurement of agreement between annotators. And also, this section covers sentiment classification approach.

3.1 Lexical Corpus Creation

Web Scraping
We created a python-based web scrapper to collect Marathi news from various online newspapers/channels and collected 1649 news headlines from categories such as general news, current affairs, sports, entertainment, art, culture, science and technology, health and medicine and so on.

Data Preprocessing
We pre-processed the crawled data into the desired forms, following steps are carried out in preprocessing:

- Manually identify and remove duplicate and irrelevant news.
- Identify news items that contains English words and transliterated them.
- Remove improper punctuation marks, smileys, hashtags and photo tags.
- Remove complex sentences because they are unsuitable for sentiment analysis.

Data Annotation and Inter-annotator agreement
For manual data annotation, we chose three Marathi language domain experts who are academicians and researchers to annotate a Marathi news dataset with polarity scores −1, 0 and 1 (Negative, Neutral and Positive, respectively). For evaluation of manual data annotation, we have used the Fleiss' Kappa inter-annotator agreement score. Following formula is used to calculate Fleiss' kappa score [10].

$$k = \frac{\overline{P}_x - \overline{P}_x}{1 - \overline{P}_x}. \tag{1}$$

where the factor $1 - \overline{P}_x$. represents the degree of agreement that can be obtained other than by chance, The degree of agreement that was achieved above chance is given by $\overline{P}_x - \overline{P}_x$. and if the evaluators are totally in agreement, Kappa $k = 1$ and $k = 0$ if there is no agreement amongst the evaluators (other than what would expected by chance).

And the inter-annotator agreement score for Marathi news dataset is $k = 0.957$, which is almost perfect agreement. Above Table 1. Fleiss's Kappa Inter-annotator agreement score. And following Table 2 shows examples of data annotation. Table 3 shows the statistics for Marathi news dataset after preprocessing and data annotation.

Table 1 Fleiss's Kappa inter-annotator agreement score

Notator (i with j)	K-score
R_{12}	0.953
R_{23}	0.965
R_{13}	0.954
Fleiss's Kappa score (R_{123})	**0.957**

Table 2 Shows examples of data annotation

News	Polarity
पकिअप पलटी होऊन भीषण अपघात, ११ जणांचा मृत्यू ७ जखमी	−1
ऐतहिासकि अ्रथसंकल्प, स्क्रॅप पॉलसिीमुळे देशात ५० हजार नवे जॉब	1
वॉर्ड आरक्षणासंद्रभात औरंगाबाद महापालकिने शपथपत्र दल्लिीला पाठवलि	0

Table 3 Shows the statistics for Marathi news dataset after preprocessing and data annotation

S. No.	Statistics	No. of news
1	Initial	1649
2	After preprocessing	1321
3	Positive	538
4	Negative	536
5	Neutral	237

3.2 Feature Extraction

To train the model, supervised machine learning algorithms require a text document in the form of a feature vector. Feature extraction techniques reduce the length of the feature vector by transforming all of the features in a lower-dimensional feature vector. Unigram features are bag-of-words (BoW) features obtained by removing unnecessary spaces and noisy characters between two words.

3.3 Sentiment Classification

Experiment work is based on supervised machine learning algorithms such as logistic regression, Stochastic Gradient Decent (SGD), support vector machine (SVM), nearest neighbour, neural network, decision tree (DT), Naive Bayes (NB), and the proposed ensemble-based model.

Individual algorithms compose an ensemble-based sentiment analysis in order to develop a highly accurate predictive model that classifies Marathi news in terms of sentiment orientation. The ensemble classifier system uses the average predicted probability, which is a soft voting approach, to determine the sentiment orientation. Figure 1 shows sentiment classification approach for Marathi news.

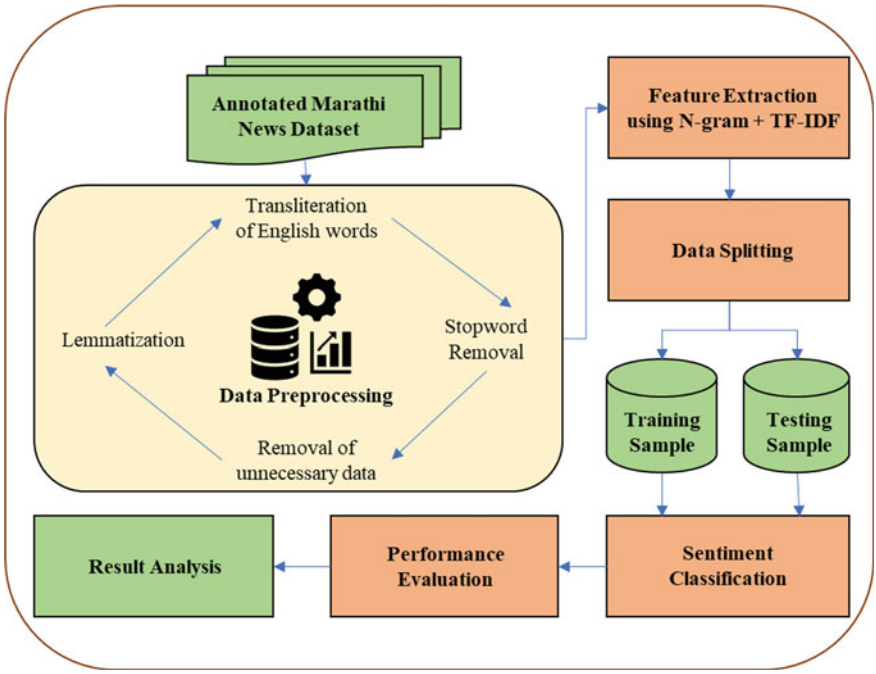

Fig. 1 Shows sentiment classification approach

4 Experimental Evaluation

In the experiment, we focused on the three types of class problems: positive, neutral and negative. We collected Marathi news headlines from a variety of newspaper and channel websites. In addition, the Marathi news dataset is divided into three categories based on the sentiment expressed in the sentences. If the expressed opinion is positive, it is labelled as 1, neutral, 0, and negative, it is labelled as −1. For training and testing samples, the dataset is divided into 80:20 ratios and various preprocessing techniques, such as data cleaning, URL and Hashtag removal, extra blank spaces, punctuation mark removal, emoticons, transliteration of English words into Marathi, Stopword removal and lemmatization, are used on the dataset. In experimentation validation, we used k-fold cross validation with $k = 5$ and $k = 10$. A classification model can be evaluated using a variety of measures, with accuracy being one of the most straightforward. The number of correctly classified examples divided by the total number of examples is the definition of accuracy. Accuracy is useful, but it ignores the complexity of class imbalances and the different costs of false negatives and false positives. Following Table 4. Shows performance evaluation of individual classifier with k-fold validation.

We performed fivefold cross validation on training dataset, and obtained highest accuracy for ensemble classifier as 95.19%, and also performed tenfold cross

Table 4 Shows performance evaluation of individual classifier with k-fold validation

S. No.	Classifier	$k = 5$		$k = 10$	
		Accuracy	F-score	Accuracy	F-score
1	Logistic regression	86.15	92.53	86.72	92.84
2	Stochastic Gradient Decent	94.12	96.84	94.44	95.32
3	SVM	88.78	93.87	90.95	95.01
4	Nearest neighbour	92.56	95.81	93.01	96.02
5	Neural network	94.16	97.02	95.07	96.93
6	Decision tree	93.12	94.16	94.21	95.16
7	Naïve Byes	87.30	93.11	88.44	93.71
8	Ensemble classifier	95.19	97.15	96.33	97.92

validation on training dataset, and we obtained highest accuracy for ensemble classifier as 96.33% for all Marathi news dataset. Figures 2 and 3 shows performance evaluation of individual classifier with fivefold and tenfold validation.

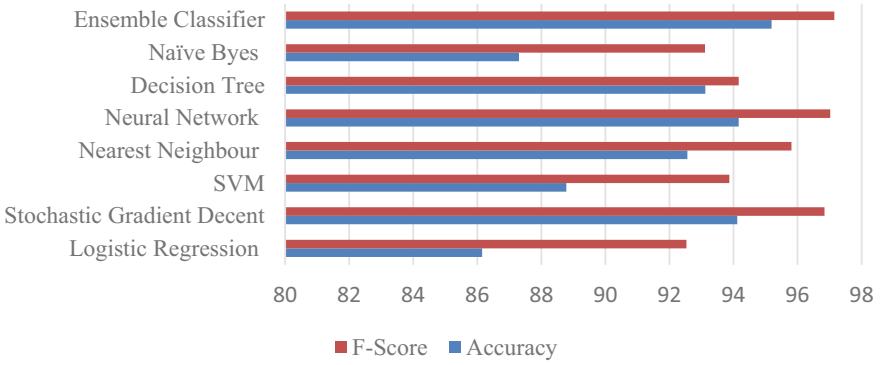

Fig. 2 Shows performance evaluation of individual classifier with fivefold cross validation

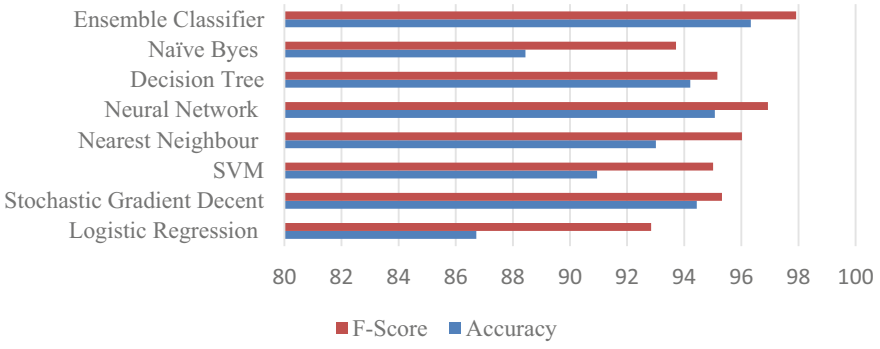

Fig. 3 Shows performance evaluation of individual classifier with tenfold cross validation

5 Conclusions and Future Work

This paper presents a baseline model for sentiment analysis in Marathi. We created an annotated dataset of Marathi news scraped from various newspaper/channel websites, and domain experts annotated Marathi news with positive, negative and neutral polarity. For sentiment analysis, we used machine learning models such as logistic regression, Stochastic Gradient Decent (SGD), support vector machine (SVM), nearest neighbour, neural network, decision tree (DT), Naive Bayes (NB) and the proposed ensemble-based model. The proposed ensemble classifier outperforms other classifiers in experiments, with an accuracy of 94.16% and an F-score of 97.02% for fivefold validation. In addition, the accuracy for tenfold validation is 95.07%, and the F-score is 96.93% and this dataset will be made available publicly for advancement in research.

In the future, we can create more domain-specific annotated lexical resources to expand resources for Indian languages, increase dataset size and implement using a deep learning-based model.

Acknowledgements Authors acknowledges the Chh. Shahu Maharaj National Research Fellowship (CSMNRF-2019), Pune, Maharashtra.

References

1. G. Arora, iNLTK: natural language toolkit for Indic languages, in *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, 2020
2. A. Kunchukuttan, *The IndicNLP Library* (2020)
3. S. Badugu, Telugu movie review sentiment analysis using natural language processing approach, in *Data Engineering and Communication Technology. Advances in Intelligent Systems and Computing*, 2020
4. A. Rajan, A. Salgaonkar, Sentiment analysis for Konkani Language: Konkani Poetry, a case study, in *ICT Systems and Sustainability. Advances in Intelligent Systems and Computing*, 2020
5. S. Rani, P. Kumar, A journey of Indian languages over sentiment analysis: a systematic review, pp. 1415–1462 (2018)
6. V.C. Joshi, V.M. Vekariya, An approach to sentiment analysis on Gujarati Tweets. Adv. Comput. Sci. Technol. **10**(5),1487–1493 (2017)
7. M. Anagha, R.R. Kumar, K. Sreetha, P.C. Reghu Raj, Fuzzy logic based hybrid approach for sentiment analysis of Malayalam movie reviews, in *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, Kozhikode, 2015
8. S. Se, R. Vinayakumar, M. Anand Kumar, K.P. Soman, Predicting the sentimental reviews in tamil movie using machine learning algorithms. Indian J. Sci. Technol. **9**(45), 1–5 (2016)
9. S.S. Prasad, J. Kumar, D.K. Prabhakar, S. Tripathi, Sentiment mining: an approach for Bengali and Tamil tweets, in *2016 Ninth International Conference on Contemporary Computing (IC3*, Noida, 2016
10. "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Fleiss%27_kappa. [Accessed May 2021]

11. M.G. Jhanwar, A. Das, An ensemble model for sentiment analysis of Hindi-English code-mixed data, in *Workshop on Humanizing AI (HAI)*, Stockholm, Sweden, 2018
12. M.A. Ansari, S. Govilkar, Sentiment analysis of transliterated Hindi and Marathi Script, in *Sixth International Conference on Computational Intelligence and Information*, Cochin, India