Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 4, pp. 1141–1152
DOI : 10.47974/JIOS-1698

Developing sentiment lexicon for Marathi : A comprehensive survey and analysis

Pallavi V. Kulkarni *
Kalpana S. Thakre †
Department of Computer Engineering
Marathwada Mitra Mandal's College of Engineering
Pune
Maharashtra
India

and

Savitribai Phule Pune University
Karve Nagar
Pune
Maharashtra
India

Abstract

Sentiment Analysis plays an important role in developing AI applications involving human language and sense. Language-specific Sentiment Lexicon is important to accelerate Sentiment Analysis. Marathi is a morphologically rich but low-resource Indic Language. The language has a very strong grammatical base showing high resemblance to the human body. The paper elaborates on issues like Lexicon basics, word embedding, and Neural Network Models in the process of Lexicon Construction. Existing databases that are useful as seed databases are enlisted. A detailed study of various methods of Lexicon Construction is done. The methods highlight that Sentiment Neural Embedding is required and contextual information plays a vital role in detecting the polarity of a single word or document. The hybrid approach which uses lexicon-based features for deep learning provides better understanding of sentiment analysis. A method is proposed for MSL construction and sentiment analysis of Marathi text. Natural Language Processing Research for Indic

* *E-mail:* **pallavikulkarni.phdcomp@mmcoe.edu.in (Corresponding Author)**
† *E-mail:* **kalpanathakre@mmcoe.edu.in**

Language is its boom. For Marathi, the efforts are seen. Considering the depth and scope of this language, more resource creation is a must for its revolutionization.

Subject Classification: *68U15.*

Keywords: *Sentiment analysis, Natural language processing, Word embedding sentiment lexicon, Neural word embedding.*

1. Introduction

Affective Computing, Conversational Artificial Intelligence, Common Sense Reasoning, Computational Social Science, and Pervasive Computing are emerging streams of Computer Science that have originated to solve technical, social, and health-related problems in our life. Understanding emotion from text will improve communication technology in various fields such as telemedicine. [1]

Semantics is the Heart of the dialog System which requires multimodal analysis. Sentiment/Emotion/Opinion detection from text is one of the key techniques of Natural language understanding involved in semantics. Sentiment and emotions are important for building an empathetic social bot [2]. Google's DialogflowSystem, Microsoft's Bot, Facebook's Wit.ai, IBM's Watson Assistant, and Alexa assistant and smart speaker by Amazon are Frameworks that allow users to create their own chatbot. A conversational recommender System (CRS) needs the sentiment of the sentence to approximate the user's feelings about the item [3].

The sentiment lexicon is one of the crucial resources needed strongly for sentiment analysis. Such lexicons are available for English, Hindi, and some other Indian Languages. But to date, no open-source high coverage Sentiment lexicon is available which could feed advanced algorithms to have high accuracy in sentiment classification for Marathi. Manual construction of such a lexicon is a time and effort-consuming task. Construction by translation from other languages has been attempted but logically every language has its own characteristics. Language-specific Sentiment Lexicon should be the aim.

To Construct such a lexicon seed dataset is available. Using appropriate lexicon structure, a sentiment lexicon for Marathi can be constructed and incrementally upgraded by using corpus-based algorithms. For the revolution of this Indic Language, Language resource creation is the precondition [4]. Part of speech information of words is necessary to incorporate into the lexicon when the sentiment analysis algorithms work for the sentence and document [5]. POS tagging should be language-specific Inappropriate or incomplete POS tagging hurts or

decreases the performance of NLP algorithms [6]. After a rigorous survey on sentiment analysis of Indigenous languages, S. Shah concluded that quality sentiment wordnet specific to a particular language must be developed by using deep learning approaches [7].

2. Background

There is a scarcity of resources for applying Natural Language Processing to Indic Languages. Digitized text is also available in huge amounts in these languages. NLP processing on such domain-specific text has the potential to influence the lives and activities of a large amount of the population. For the revolution of this Indic Language, Language resource creation is the precondition [8]. Available sentiment databases are corpus-based where paragraphs or sentences are classified based on polarity. Some are derived from the English sentiment lexicon by translation. A language-specific attempt is necessary to enlighten its originality.

3. Literature Survey

The Marathi Language processing is initiated and the detail insight is given by P. Lahoti covering available tools, techniques used and resources created for MLP. The author concludes that there is a need of creating more resources including a Sentiment Lexicon [9]. The sentiment analysis for English is groomed by a large size dataset, advanced tools and GPU processing power. Useful seed datasets are elaborated in first section. Different word embedding techniques and modeling algorithms are elaborated in the second and third section. Finally, methodologies useful for low resource languages are summarized.

3.1 *Seed Dataset*

A large resource repository for Indic languages is available through joint efforts of Government, IIT's and Universities. Very few are available for Marathi sentiment analysis [10]. Following resources will serve the requirement of construction of Sentiment lexicon for Marathi.

3.1.1 Marathi-mr-NRC-Emotion-Intensity-Lexicon-v1

The file contains emotion scores for eight different emotions (anger, anticipation, disgust, joy, fear, sad-ness, surprise, trust) for 9922 Marathi words with its English synonym and emotion intensity scores (0 to 1) [11].

3.1.2 NRC-Emotion-Lexicon-v0.92-In105Languages-Nov2017Translations

Excel file contains positive, and negative sentiment scores and emotion scores (8 emotions) in the form of 0 or 1 for different languages including Marathi. Lexicons are constructed by translating English to Marathi from the original English lexicon. POS tags are not associated with words. There is no translation for some of the words [12].

3.1.3 SentiWordNet v3.0.0 (1 June 2010)

English SentiWordNet v3.0 is based on WordNet 3.0. The pair (POS, ID) uniquely identifies a WordNet synset. Labelling is fuzzy. The words positivity and negativity score are given by PosScore and NegScore from SentiWordNet. The objectivity score can be implicitly calculated. A word may have multiple sentiment corresponding to multiple meaning. In SynsetTerms, a column of SentiWordNet term is represented with sense number [13].

3.1.4 HSWN (Hindi SentiWordNet)

All fields are separated by Tab space. (# POS tag, #Field 2: Synset ID (HindiWN), #Positive score, #Negative score, #Related terms separated by comma). All words of sentence are not necessarily polar. Polar words like Adjective, Noun, Adverb, Verb are identified and feed to the algorithm. The words are grouped based on the lexical concept of Synonymy which is base for lexicon construction [14].

3.1.5 MarathiWN_1_3

Rich source of Marathi words with different POS tags (noun, adjective, verb, adverb, onto) and Updated Wordnet database with several new and updated synsets. Word morphology is also addressed. (IIT Bombay copyright) Marathi_Sense_Tagged_Data_Tourism Corpus containing files with positive and negative scores at document level.[15]

3.1.6 L3CubeMahaSent

A Marathi Tweet-based Sentiment Analysis Dataset: It is largest publicly available Marathi sentiment analysis dataset containing manually labeled tweets. Total 18378 tweets are classified into three classes: Positive (1), Negative (-1), and Neutral (0). fastText embedding model provided by IndicNLP and Facebook based on CNN and BiLSTM is used. Multilingual BERT and Indic BERT are also used for performing baseline experiments.

For two class classification, CNN with IndicFT trainable model gave the best performance 93.13% accuracy. IndicBERT performed well in 3 class classifications. Performance of BiLSTM with IndicFT trainable was average in both type of classification [16].

3.2 *Word Embedding Techniques*

Word embedding is a representation of word/s in real number or a vector form. Word embedding space captures some kind of relationship about the words like meaning, morphology, context etc. The survey is basically performed on the basis of techniques available in the literature for word embedding techniques. Based on word features word embedding techniques are classified into 1. Statistical-Based Word Embedding and 2. Neural-Network based word embedding.

3.2.1 Statistical-Based Word Embedding

Statistical methods capture the relevance of words with respect to the corpus of text. TF-IDF is one of the statistical techniques. It does not capture semantic word associations.

This method is used by Rini Wijayanti for developing Indonesian sentiment lexicon. Seed lexicon is constructed by extracting English SentiWordNet Bahasa. To ensure the quality of the seed lexicon good quality hand-checked transition and high-quality automatic translation is performed. Formal and informal words are added in lexicon by dictionary-based method and corpus-based methods respectively. The language-specific sentiment score of a word is calculated using term/synset relative frequency in the sentiment corpus. Context of the word is considered to calculate negation and booster score. Valence score of longest synset is used to calculate sentence score. The metrics used for evaluation are accuracy and F1 score. Constructed lexicon is tested for sentiment classification of Trusted Company corpus and Twitter corpus [17].

Lorenzo Gatti et al constructed a sentiment lexicon called Senti Word Net by using an ensemble learning framework in two steps. In the first step, prior polarity lexica are manually constructed to achieve high precision. In the second step, various formulae are ensembled to calculate prior polarity from the posterior of the polarity of unseen words resulting in high coverage. From the result it is concluded that selecting one sense is not a good choice, instead combining different posterior polarities in various ways using the ensemble method proved as the best performing formula. The idea is to view the same information from a different

perspective. SVMfs using SWNS outperforms among all other methods. Results are analyzed through three approaches 1. Random 2. Posterior to prior 3. Learning algorithm. The formula based on the "posterior polarity saliency "hypothesis performs best in all experiments [18].

3.2.2 Neural Network-based Word Embedding

Automatically features are extracted in this approach using neural network model. The reaschers are adapting it for improving performance.

By Siwei experimentation is done on seven different frequently used word embedding models along with neural word embedding, on corpora of different sizes (small 13M to large 2.8B) with varying size dimensions (10 to 200) as parameters. The observation was, that corpus size is important but the domain is more important for better performance. The average 50-dimensional feature vector gives better performance for analyzing the semantic properties. A pure in-domain corpus with a large corpus size achieves better performance compared with a mixed domain corpus [19].

Minglei Li obtained rich semantic meaning from word embedding. He assumed that different features in word embedding contribute differently to a particular affective dimension and a particular feature in word embedding contribute differently to different affective dimensions. Regression models featuring less computation cost were used to learn affective meaning in each dimension. It is based on Neural Network and skip gram with Negative Sampling [20].

3.3 *Modeling Algorithms*

A sentiment analysis can be done by main three approaches Lexicon Based, Deep Learning Based and Hybrid.

3.3.1 Lexicon-based Approach

They give lexical cues to handle negation and interrogate sentences. Can deal with a particular type of negation sentences. Dictionary of polar and negation words improves the accuracy. A set of rules for polar and negation words compiled to judge the sentiment orientation of a text document. Generalization capability is poor. This method cannot efficiently predict implicit orientation.

Emotion Lexicon introduced by Kanika Garg for Sentiment classification of Hindi Text is scalable. This Hindi Emotion Net was created in three steps. Using Parrot's emotion model and English Affect WordNet

seed word list for of Hindi Emotion bearing words is constructed. These words are arranged in hierarchy of emotions based on the WordNet affect. Manually emotion class of some words is assigned considering the language translation limitations. The root is labeled as "emotion" then nodes at level 2 are labeled "positive"," negative", "neutral" & "ambiguous" representing the respective classes. Affect WordNet hierarchy is used to construct the graph by translation. Finally Hindi Emotion Net is constructed.

In IndoWordNet the information is structurally organized. It is represented in graphical form called sense-based network. It is based on Psycholinguistic principles. It contains content words like nouns verbs, adjectives, and adverbs and relations defined over them. The Nouns, Adjectives, Lexical words, Adverbs carry affective meaning. Conjunction plays an important role. Degree centrality, closeness, betweenness, and page rank. The classification accuracy is 85.78% for dataset of screenplays, stories, and blogs in Hindi. Translated dataset SemEval gave 75.91% classification accuracy. Performance measures Precision, recall and F1 score are used. Highest accuracy and precision are given by page rank [21].

3.3.2 Deep Learning-based Approach

Deep learning-based models are designed from Artificial Neural Network based techniques. These methods are capable of automatic feature extraction and improving classification and prediction accuracy. When only deep learning-based methods are applied it is difficult to classify negation and interrogative sentences. A. Pingle Developed a Multidomain Marathi Sentiment annotated database using MahaBERT Model. But it could not handle semantic ambiguity due to homonymy, polysemy [4].

3.3.3 Hybrid Approach

Hybrid approach where involvement of both deep learning and lexicon-based methods, provides better understanding of sentiment analysis. Lexical Information for word embedding is considered such as key lexical words, sentimental words, POS, Conjunction.

Dong Deng found Self -Attention Mechanism useful for the sentiment classification task. Sentient Lexicon is primary and valuable in sentiment analysis tasks. There are two categories of Sentiment Lexicon 1. Domain independent 2. Domain specific lexicon. The architecture of Sparse Self-

Attention (SSA) LSTM processes the input in vector form through four different layers. Text data through word embedding is given to input layer as a real value vector. Contextual semantic relation between words is extracted in LSTM layer. Using attention mechanism weights are assigned to each word in documents in Sparse self-attention layer. Also, words which does not contribute are filtered. Sentiment-aware word embedding is done where weights are applied to each word. The role of different words in deciding documents semantic and sentiment is different. Weights are given to word according to their importance. L1 regularizer is introduced to ensure that most of the uninformative words are ignored. The Prediction (fully connected layer) layer is output layer. Weight summation of all words is input which represents the whole document to this layer. And it predicts sentiment polarity. First a seed list with positive (125) and negative (109) words is manually constructed. A Lexicon with 2003 positive and negative word derived from publicly available Liu's Lexicon [22].

Topic and sentiment these two dimensions are used by Dong Deng. In this TaSL topic-adaptive Sentiment Lexicon construction model Latent is introduced. Each word is generated over both topics and sentiments. Multinomial distribution is used to represent words. Generative probabilistic model is used to characterize the relationship between topic, words, and sentiment. This is based on Bernoulli distribution and Dirichlet

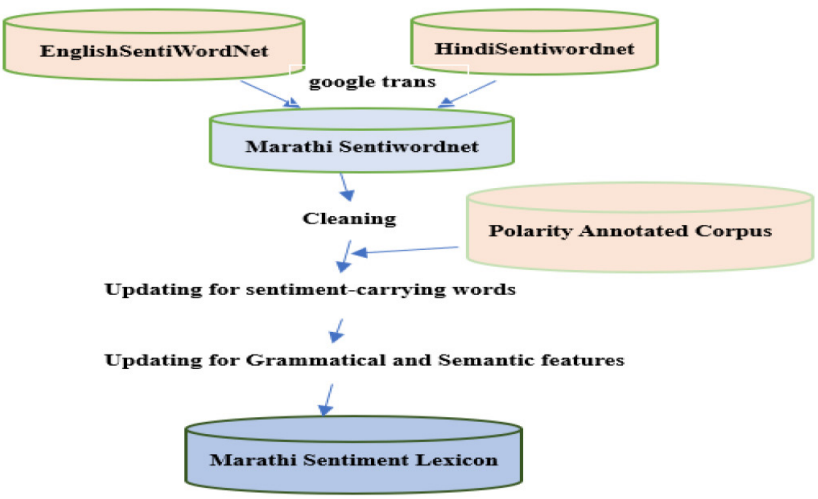

Figure 1

Construction of Marathi Sentiment Lexicon

distribution. Probability distribution of each word over sentiment polarities is used to calculate sentiment strength of each word. Absolute value of corresponding sentiment scores is used to determine sentiment polarity. The polarity of some of the words changes as the topic changes. The model converges faster and is more stable than ssLDA. These models are generated by an iterative algorithm with collapsed Gibbs sampling techniques. For constructing Domain specific sentiment lexicon this model is beneficial [23].

Context of some word will decide their probability. Wu developed a hybrid model and used Polar Flipping words to process such situations. Word level polar labels are considered noisy and the flipping model models the relationship between noisy and true labels. CNN–Lex is better than CNN under both character and word embedding which concludes that lexicon information found useful in Sentiment analysis. The performance of T1–LSTM (Two level LSTM) with both lexicon embedding and the flipping mode; are consistently better than those of T1 LTF than those of T1-LSTM without lexicon embedding. T1-LSTM with flipping also slightly outperforms [24].

The pragmatic study of text based different sentiment analysis models, is useful to choose an appropriate model for contextual sentiment analysis. Khan concluded the GRU (Gated recurrent Unit) outperforms all other Neural Network models. LDA and LISA reduces dimensions of feature vector. Genetic Algorithms are helpful in optimizing execution delay. [25]

4. The proposed System

Considering the low resource availability for Marathi Natural Language Processing a Marathi Sentiment Lexicon can be constructed from other language sentiment lexicon. It is divided in to two phases MSL Construction and Sentiment Recognition. MSL Construction will be carried out through three approaches: From English/Hindi by Translation, Corpus based, Crowdsourcing. The seed MSL construction from English and Hindi SentiWordNet is described in Figure 1. Word embedding techniques based on grammatical features like Part of Speech tagging and semantic features based on word similarity index can be used to construct feature vectors. These feature vectors will be fed to various Machine Learning, Neural Network Models for Marathi Sentiment Lexicon Construction.

5. Conclusion

Research in Sentiment Analysis for Marathi Text is in its naïve stage Researchers attempted to solve the task. Word embedding is important to describe features of words occurring in a corpus. Neural Network based word embedding allows the algorithm to automatically extract the features to construct feature vectors. Not only contextual features but sentiment features of words are important for developing a Sentiment Analysis System. Some translated dictionaries and annotated corpora are available in Marathi. Considering the scope and depth of this morphologically rich language more efforts are essential. The proposed system is a step towards creating an important resource required for Marathi Sentiment analysis.

References

[1] Busra Saylan , Songul Cinaroglu . Opinion mining and machine learning analysis : What emotions twitter data tell us about telemedicine?. *Journal of Statistics & Management Systems* ISSN 0972-0510 (Print), ISSN 2169-0014 (Online) Vol. 27, No. 3, pp. 605–633 (2024). DOI : 10.47974/JSMS-1028

[2] Huang, Minlie, et al. "Challenges in Building Intelligent Open-Domain Dialog Systems." *ACM Transactions on Information Systems*, vol. 38, no. 3 (2020), https://doi.org/10.1145/3383123.

[3] Zhao, Deji, and Bo Ning. A Survey on Conversational Question-Answering Systems. *Lecture Notes in Electrical Engineering*, vol. 654 LNEE, no. 5, pp. 1857–61 (2021), https://doi.org/10.1007/978-981-15-8411-4_244.

[4] Pingle Aabha, et al. L3Cube-MahaSent-MD: A Multi-Domain Marathi Sentiment Analysis Dataset and Transformer Models (2023). arXiv:2306.13888v1https://doi.org/10.48550/arXiv.2306.13888.

[5] Yin Fulian, et al. "The Construction of Sentiment Lexicon Based on Context-Dependent Part-of-Speech Chunks for Semantic." *IEEE Access*, vol. PP, p. 1 (2020), https://doi.org/10.1109/ACCESS.2020.2984284.

[6] Moeller Sarah, et al. "To POS Tag or Not to POS Tag: The Impact of POS Tags on Morphological Learning in Low-Resource Settings." ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on

Natural Language Processing, Proceedings of the Conference, pp. 966–78 (2021), https://doi.org/10.18653/v1/2021.acl-long.78.

[7] Shah Sonali Rajesh, and Abhishek Kaushik. Sentiment analysis on indian indigenous languages : a review on multilingual opinion mining. preprints 2019110338 https://doi.org/10.20944/preprints201911.0338.v1.

[8] Ranathunga, Surangika, and Ranjiva Munasinghe. *Indic Language Computing*. no. June 2020 (2019), https://doi.org/10.1145/3343456.

[9] Lahoti, Pawan, et al. A Survey on NLP Resources , Tools , and Techniques For Marathi Language Processing. ACM Transactions on Asian and Low-Resource Language Information Processing, Volume 22 Issue 2 Article No 4 7 pp 1–34. https://doi.org/10.1145/3548457.

[10] Kunchukuttan Anoop. GitHub - AI4Bharat/Indicnlp_catalog: A Collaborative Catalog of Resources for Indian Language NLP. https://github.com/AI4Bharat/indicnlp_catalog.30 March 2024.

[11] Mohammad Saif M., and Peter D. Turney. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, vol. 29, no. 3, pp. 436–65 (2013), https://doi.org/10.1111/j.1467-8640.2012.00460.x.

[12] Mohammad Saif M., and Peter D. Turney. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. CAAGET '10 Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, no.June, pp. 26–34 (2010).

[13] Baccianella, Stefano et al. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. International Conference on Language Resources and Evaluation (2010).

[14] Das Amitava, and Sivaji Bandyopadhyay. "SentiWordNet for Indian Languages." The 8th Workshop on Asian Language Resources (ALR), August, no. August, pp. 56–63 (2010), http://www.aclweb.org/anthology/W/W10/W10-3208.pdf.

[15] Aditya Joshi, Sagar Ahire, Pushpak Bhattacharyya, 'Sentiment Resources: Lexicons and Datasets', Book Chapter, 'A Practical Guide to Sentiment Analysis', Editors: Dr. Dipankar Das, Dr. Erik Cambria, Springer Publication

[16] Kulkarni, Atharva, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. L3CubeMahaSent: A Marathi Tweet-

Based Sentiment Analysis Dataset. April (2021), http://arxiv.org/abs/2103.11408.

[17] Wijayanti Rini, and Andria Arisal. Automatic Indonesian Sentiment Lexicon Curation with Sentiment Valence Tuning for Social Media Sentiment Analysis. ACM Transactions on Asian and Low-Resource Language Information Processing. Volume 20 Issue 1Article No.:15 pp 1–16 https://doi.org/10.1145/3425632

[18] Gatti Lorenzo, et al. SentiWords : Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. arXiv:1510.09079v1 no. 4, pp. 409–21 (2016). https://doi.org/10.1109/TAFFC.2015.2476456.

[19] Lai Siwei, et al. How to Generate a Good Word Embedding. arXiv:1507.05523v1 https://doi.org/10.48550/arXiv.1507.05523

[20] Li Minglei, et al. "Inferring Affective Meanings of Words from Word Embedding." *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 443–56 (2017), https://doi.org/10.1109/TAFFC.2017.2723012.

[21] Altameem, Ayman, et al. "P-ROCK: a sustainable clustering algorithm for large categorical datasets." Intell. Autom. Soft Comput 35.1: 553-566 ((2023)).

[22] D. Deng, L. Jing, J. Yu and S. Sun, "Sparse Self-Attention LSTM for Sentiment Lexicon Construction," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1777-1790, Nov. (2019), doi: 10.1109/TASLP.2019.2933326.

[23] D. Deng, L. Jing, J. Yu, S. Sun and M. K. Ng, "Sentiment Lexicon Construction With Hierarchical Supervision Topic Model," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 704-718, April (2019), doi: 10.1109/TASLP.2019.2892232.

[24] O. Wu, T. Yang, M. Li and M. Li, "Two-Level LSTM for Sentiment Analysis With Lexicon Embedding and Polar Flipping," in *IEEE Transactions on Cybernetics*, vol. 52, no. 5, pp. 3867-3879, May (2022), doi: 10.1109/TCYB.2020.3017378.

[25] Khan H. T, Ridhorkar Sonali. A pragmatic analysis of text-based sentiment analysis models from an analytical perspective. *Journal of Statistics & Management Systems* ISSN 0972-0510 (Print), ISSN 2169-0014 (Online) Vol. 27, No. 2, pp. 285–294 (2024). DOI: 10.47974/JSMS-1254 https://doi.org/10.47974/JSMS-1254.