Int. j. inf. tecnol. (October 2021) 13(5):2069–2074
https://doi.org/10.1007/s41870-021-00702-1

ORIGINAL RESEARCH

Sentiment analysis of Marathi news using LSTM

Manisha Satish Divate[1]

Received: 31 January 2021 / Accepted: 29 April 2021 / Published online: 31 August 2021

Abstract Sentiment analysis of online contents related to e-news, product, services etc., become very important in this digital era in order to improve the quality of the service provided. The proposed sentiment analysis of Marathi e-news will help the online readers to read the positive news to avoid the depression which may be caused by reading the negative news. The system will be also used to filter out the news before uploading it online. Machine learning based, knowledge based and hybrid are the three approaches use to perform the sentiment analysis of a text, audio, emotions. The proposed system is polarity-based sentiment analysis of the e-news in Marathi. Marathi ranks third in most spoken languages used in India. Computationally it is low resource language. To compute the polarity of the Marathi e-news text, LSTM, deep learning algorithm is used. The model identifies the polarity with accuracy of 72%.

Keywords Natural Language Processing · Long short-term memory · Recurrent Neural Network · SentiWordNet

1 Introduction

Sentiment analysis uses the Natural Language Processing (NLP) techniques to extract the opinion, sentiments of the user about the product, service they are using. This helps to understand the popularity, utilization of a product or a service and necessary changes needed to improve its quality [1, 2]. Table 1 shows the types of sentiment analysis used.

Based on the methodology used to perform the sentiment analysis, they are classified as:

1. Machine learning based sentiment analysis
2. Knowledge based sentiment analysis
3. Hybrid sentiment analysis

Machine learning approach, is a statistical approach, where surface features such as word frequency, bag of words, are used for sentiment prediction. Algorithm such as Support vector machine (SVM), Naïve Bayes, LSTM, CNN, Combine CNN-LSTM are used to extract and classify the text into polarity classes defined [12–14].

Knowledge/Dictionary based approach where the lexicon-based semantic information (ex. SentiWordNet, WordNet), grammatical dependency of words is used for the prediction a sentiment [4].

Hybrid technique is the combination of Machine Learning algorithms and Knowledge based tools used for the polarity identification.

LSTM method of deep learning, remembers the long-distance relationship between the words which is helpful in deciding the sentiment of a sentence. The paper proposes a study of sentiment analysis of Marathi e-news using LSTM method. The rest of the paper is organized as follow: Sect. 2 provides the current state of art in sentiment analysis. Section 3 explains the LSTM techniques for sentiment analysis. Results and conclusion of the research is in Sect. 4 and Sect. 5 respectively.

✉ Manisha Satish Divate
manisha.divate@upgcm.ac.in

[1] Usha Pravin Gandhi College of Arts Science and Commerce, Vile Parle, Mumbai 400056, India

Table 1 Types of sentiment analysis

Type	Purpose
Subjectivity/objectivity [3]	To classify the text into objective sentence and the subjective sentence
Aspect based [4–7]	Identifying the sentiments for the given aspect. Here aspect is the category or the feature of the product or service
Grading/polarity based sentiment analysis [8]	To identify the positive, negative polarity of the sentence or document
Multilingual sentiment analysis [9–11]	Analyzing the sentiments of a low resource language using a resourceful language

2 Literature survey

Classification of the review, based on the sentiment orientation of the text was the first work found in the literature [15]. Similarity between the pair of the words is measure using Point-wise mutual information (PMI)-Information Retrieval (IR) algorithm. The algorithm first extracts the adjective and adverb phrases present in the sentence. Semantic orientation for each phase is computed and further based on the average semantic value of the phrase, review is classified.

Next notable work in 2004 was feature based opinion summarization by [16]. It identifies the features of the product, count the number of positive and negative opinion for each feature. Using Apriori algorithm, frequently occurring phrases are identified.

[8] performs the sentiment classification of the Bengali and the Hindi tweets using Support Vector Machine (SVM). The research considers the unigram, bigram to compute the word overlapping. Use of distributional thesaurus which compare the words automatically for the similarity and group them. SentiWordNet is used to count the positive sentiments (mark with numeric value 1) for the words which found in SentiWordNet. If the word is negative in SentiWordNet then count it as negative sentiment.

[17] had used Arabic health care data for the sentiment analysis. Arabic corpora is rich in morphology but has limited tools available for the preprocessing. A combined CNN and LSTM model is used in the research. The Convolution layer uses the filters to detect the multiple features and represent the sequence of vectors in feature maps. Extracted featured vector is given to LSTM model for sentiment analysis. By considering the previous data, LSTM layers captures the new sequence data. The recorded accuracy of the model is 94.24%.

Another approach of sentiment classification as mentioned in Table 1 is Aspect based sentiment analysis (ABSA). In ABSA, Aspect Terms are the features, attributes or categories such as food, electronics, marketing, travel, movie etc. Here text is annotated as per the AT. There are two approaches used in ABSA; 1. Assign polarity to AT and 2. Polarity to the sentence is the count of polarities of ATs present in the sentence. But sentence having ATs with opposite polarity is a challenge [4, 7, 18].

Use of bilingual dictionary was the another approach found in literature for the low resource languages like Telugu, Hindi, Bengali (English–Indian Languages) [3]. Translation software is used to translate the low resource language text (Ex. Hindi) in to high resource language text (Ex. English) [5].

[19] had considered three languages Hindi, Bengali, Tamil for the opinion mining. Each tweet is human annotated. 2-class classification (positive, negative) and 3 class classification (positive, negative and neutral) had performed on the tweet data. For classification, features such as word n-gram, SentiWordNet features, surface features (such as *number of hashtags, number of @ symbols, number of characters, number of words in tweet* etc.,) are derived manually. Various different classifiers such as naïve bayes, logistic regression, decision tree, Random forest, SVM are used to classify the tweets in 2-class or 3-class.

Here the survey shows that sentiment analysis for low resource language is a challenge due to the unavailability of complete resources such as parser, tagger, WordNet etc. Knowledge bases such as WordNet, SentiWordNet, bilingual dictionary needs continuous improvement, updation and maintenance. The thought was to perform the sentiment analysis by training the deep learning model which is independent of the NLP tools and the knowledge bases.

The next session explains the reason for selecting LSTM model sentiment detection of Marathi news text.

3 Methodology

To understand the LSTM, we need to understand the Recurrent Neural Network (RNN). Traditional neural network does not have an ability to retain the previous event. Figure 1 shows the simple perceptron model which retain the information by repeatedly occurring. RNN is a

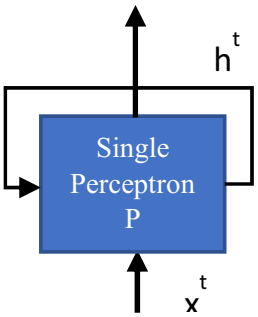


Fig. 1 Single perceptron with feedback loop

collection of such perceptron which allows the network to keep the information re-occurring.

Figure 2 shows the RNN model where every node passes the information to the next successive node. This helps in predicting the possibility of occurrence of the next event. Learning process is carried out using the back propagation in time and updating the weight values.

$$W_{\text{new}} = W_{\text{old}} - \text{learningrate} * \text{Gradient}$$

Gradient value computed here is the derivative of the predicted error. RNN model uses sigmoid activation function which is stated as,

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

Gradient descent value of sigmoid function is between 0 and 0.25. Gradient is the chain rule of differentiation. As the number of the layers in RNN increase, at one particular layer the new weight and old weight become same (because of the long series of multiplication of derivatives) and that causes no significant learning. This is called vanishing gradient problem.

The problem of long-term memory is overcome in LSTM. Figure 3 shows the LSTM node which consist of four different neural layers respectively: (1). forget gate f_t, (2) input gate i_t (3). Output gate O_t and (4) current cell state C_t. The cell state equation for LSTM node is given as:

$$c_t = c_{t-1} * f_t + \tilde{c}_t * i_t \tag{2}$$

The gradient of the LSTM node is in additive form which eliminates the problem of vanishing gradient [20]. Hence for the word sequence, language modeling, LSTM is used [21, 22].

LSTM node has four gates. As shown in Fig. 4, sentence (the Movie was awesome) is pre-processed, tokenized and each token in sentence is represented as a unique integer value. The word vector is of the uneven length and hence it is padded with zeros. This process is called as *Word-embedding*. Word Embedding is important as it reduces the dimension of the word representation. Such embedded vector, x_i, is considered for the further analysis.

Every node in LSTM receives the cell state $C_{t-1}$from previous LSTM node and the new word vector x_t and the output vector h_{t-1}. The forget gate is use to decide whether the information is to forget or keep in the cell state.

$$f_t = \sigma(W_f(x_t, h_{t-1}) + b_f) \tag{3}$$

σ represent the sigmoid function which produce the output in range of 0 to 1, where 1 represents to keep the information and 0 to forget. The next layer is input gate i_t which decided which the value to be consider next and $\tilde{c}_t$ for the vector value to be updated to the current state.

$$i_t = \sigma(W_i(x_t, h_{t-1}) + b_i) \tag{4}$$

$$\tilde{C}_t = tanh(W_c(x_t, h_{t-1}) + b_c) \tag{5}$$

Output of the LSTM node, h_t, shown in Eqs. (6) and (7), is the multiplication of outcomes of sigmoid layer and the tanh layer respectively.

$$O_t = \sigma(W_o(x_t, h_{t-1}) + b_o) \tag{6}$$

$$h_t = O_t * tanh(C_t) \tag{7}$$

Neural network's sequential model is used, where sequence data will flow from one layer to another layer in same sequence. Every hidden layer neuron uses *tanh* activation function. Output layer of the model uses softmax activation function. SoftMax is a form of logistic regression function which produces the output in the range of

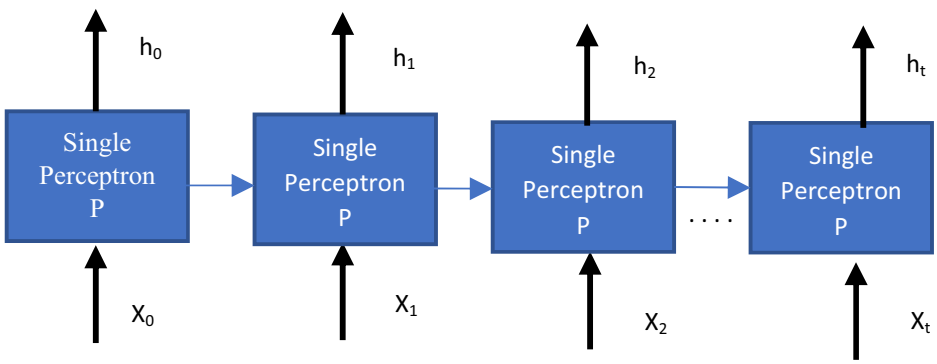
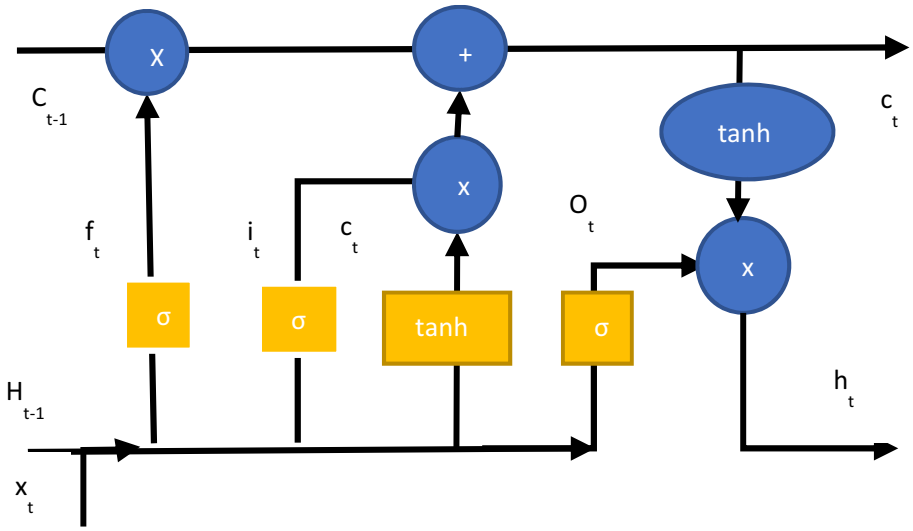


Fig. 2 Simple RNN

Fig. 3 The LSTM Node



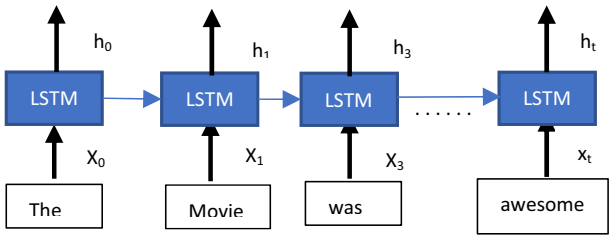


Fig. 4 LSTM layer

[0,1] and hence it is suitable for the multiclass classification [23].

4 Result analysis

Bidirectional LSTM deep learning model is setup as discussed in section III. For this pilot experiment 132 Marathi e-news sentences were collected from the online news website such as Loksatta, Maharashtra Times, Lokmat, Sakal etc. Punctuation marks, special symbols, English characters are removed from the sentences as a part of data pre-processing. With the help of 2 native speakers of Marathi language, the sentences were annotated with the sentiments positive, negative and neutral. Inter rater agreement between the annotators is computed using kappa statistics and found satisfactorily. Sentence with neutral sentiment (11 sentences) were eliminated from the experiment. Remaining 121 sentences become an input set to the embedded layer. The LSTM model is trained with 80% of data.

Standard evaluation matrices, precision recall and F-score measures are used to measure the performance of the model. For the evaluation, remaining 20% of data is use as a test data. The precision, recall and F-score measure for the test sample was recorded as 0.55, 1.00, 0.72 respectively. To check the performance of the model we compare our results with the *imdb* database provided by keras, and another researcher's work which is listed in Table 2.

Table 2 Result analysis

Data set	Language	Techniques	*P*	*R*	F1-measure
Imdb	English	LSTM	0.89	0.83	0.86
Marathi_news	Marathi	LSTM	0.55	1.00	**0.72**
Tweeter data [24]	English	SVM	0.413	0.88	0.855
Indian Tweets [8]	Hindi	SVM-Wordnet	–	–	**0.48**
	Bengali				**0.42**
Vietnamese Corpus [22]	Vietnamese	CNN-LSTM	0.91	0.86	**0.88**
[19]	Bengali	NB	–	–	**0.67**
	Hindi	LR			**0.81**
	Tamil	NB			**0.62**
Telugu news [3]	Telugu	SentiWord-Net	0.70	0.77	**0.73**

Bold are the comparative f-score measures and shows that Marathi_news data set with LSTM method performs better than other listed methods of Sentiment Analysis

It has been observed that, LSTM model for the *imdb* movie review dataset in English text shows better results than the *Marathi_news* sentiment analysis. The reason may be the size of training data set used is small compared to the English *imdb* movie dataset. It is also observed that, compared to SVM algorithm for sentiment classification for the Indian tweets in Hindi and Bengali [8], which gives the accuracy of the 0.42 and 0.48, proposed LSTM model performs better.

The proposed system is comparable with the Telugu news sentiment analysis having accuracy 73% [3]. Telugu e-news system had used Telugu SentiWordNet. In Knowledge based approach, result of analysis is purely based on the quality of lexical resources used for the annotation of the sentiments. A system for Vietnamese corpus [22] performs far better than the proposed system. The use of CNN- LSTM captures both global and local dependencies of the sentiment which improves the performance of model.

Table also shows that Machine learning algorithm such as Naïve bayes (NB) and Linear Regression (LR) along with lexical resources (WordNet) produces better result.

5 Conclusion

To conclude, proposed systems performance is comparable with the other Indian Language sentiment analysis system shown in the Table 2. Performance could be better if the input set sentences are diverse and rich in vocabulary. Model should be trained with such sentences which are challengeable to naïve experts predict to predict their correct sentiments. Machine learning approach is more independent approach for the sentiment analysis as underline dependencies on lexical resource updation and maintenance is eliminated.

Future research should work on improving the accuracy of the model which could be used as a recommender tool for suggesting positive e-news.

References

1. Sánchez-Rada JF, Iglesias CA (2019) Social context in sentiment analysis: formal definition, overview of current trends and framework for comparison. Inf Fusion 52:344–356. https://doi.org/10.1016/j.inffus.2019.05.003
2. Drus Z, Khalid H (2019) Sentiment analysis in social media and its application: systematic literature review. Proced Comput Sci 161:707–714. https://doi.org/10.1016/j.procs.2019.11.174
3. Naidu R, Bharti SK, Babu KS, Mohapatra RK (2017) Sentiment analysis using Telugu SentiWordNet. Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017, pp 666–670, 2018, doi: https://doi.org/10.1109/WiSPNET.2017.8299844
4. John Pavlopoulos I (2014) Aspect based sentiment analysis. Athens University of Economics and Business
5. Akhtar MS, Ekbal A, Bhattacharyya P (2016) Aspect based sentiment analysis in Hindi: resource creation and evaluation. Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. pp. 2703–2709
6. Moens M, Steedman M (1987) Temporal ontology in natural language, pp. 1–7. Doi: https://doi.org/10.3115/981175.981176.
7. Alqaryouti O, Siyam N, Monem AA, Shaalan K (2019) Aspect-based sentiment analysis using smart government review data. Appl Comput Inf. https://doi.org/10.1016/j.aci.2019.11.003
8. Kumar A, Kohail S, Ekbal A, Biemann C (2015) IIT-TUDA: System for sentiment analysis in Indian languages using lexical acquisition. Lect Notes Comput Sci (Incl Subser Lect Notes Artif Intell Lect Notes Bioinform) 9468:684–693. https://doi.org/10.1007/978-3-319-26832-3_65
9. Ainin S, Feizollah A, Anuar NB, Abdullah NA (2019) Sentiment analyses of multilingual tweets on halal tourism. Tour Manag Perspect. https://doi.org/10.1016/j.tmp.2020.100658
10. Denecke K (2008) Using SentiWordNet for multilingual sentiment analysis. Proc. - Int. Conf. Data Eng. pp. 507–512, doi: https://doi.org/10.1109/ICDEW.2008.4498370.
11. Boiy E, Moens MF (2009) A machine learning approach to sentiment analysis in multilingual web texts. Inf Retr Boston 12(5):526–558. https://doi.org/10.1007/s10791-008-9070-z
12 Kula S, Choraś M, Kozik R, Ksieniewicz P, Woźniak M (2020) Sentiment analysis for fake news detection by means of neural networks. Lect Notes Comput Sci (Incl Subser Lect Notes Artif Intell Lect Notes Bioinform). https://doi.org/10.1007/978-3-030-50423-6_49
13. Gopalakrishnan K, Salem FM (2020) Sentiment analysis using simplified long short-term memory recurrent neural networks. arXiv, pp. 1–6
14. Wang JH, Liu TW, Luo X, Wang L (2018) An LSTM approach to short text sentiment classification with word embeddings. Proc. 30th Conf. Comput. Linguist. Speech Process. ROCLING 2018, pp. 214–223
15. Turney PD (2002) Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Proc. 40th Annu. Meet. Assoc. Comput. Linguist., 3, pp. 417–424. Available: http://www.google.com.
16. Hu M, Liu B (2004) Mining opinion features in customer reviews. Proc. Natl. Conf. Artif. Intell., pp. 755–760
17 Alayba AM, Palade V, England M, Iqbal R (2018) A combined CNN and LSTM model for Arabic sentiment analysis. Lect Notes Comput Sci (Incl Subser Lect Notes Artif Intell Lect Notes Bioinform). https://doi.org/10.1007/978-3-319-99740-7_12
18 Akhtar S et al (2016) Aspect based sentiment analysis: category detection and sentiment classification for Hindi. Appl Comput Inf. https://doi.org/10.1016/j.aci.2019.11.003
19. Phani S, Lahiri S, Biswas A (2016) Sentiment analysis of tweets in three indian languages. Proc. 6th Work. South Southeast Asian Nat. Lang. Process., pp. 93–102. Available: http://fire.irsi.res.in/fire/static/.
20. Arbel N (2018) How LSTM networks solve the problem of vanishing gradients?, 2018. https://medium.com/datadriveninvestor/how-do-lstm-networks-solve-the-problem-of-vanishing-gradients-a6784971a577#: ~ :text=However%2CRNNs suffer from the,no real learning is done

21. Olah C (2015) Understanding LSTM Networks. 2015. https://colah.github.io/posts/2015-08-Understanding-LSTMs/.
22. Vo QH, Nguyen HT, Le B, Le Nguyen M (2017) Multi-channel LSTM-CNN model for Vietnamese sentiment analysis. Proc 2017 9th Int. Conf. Knowl. Syst. Eng. KSE 2017, vol. 2017-Jan, October, pp. 24–29. Doi: https://doi.org/10.1109/KSE.2017.8119429.
23. Arras L, Montavon G, Müller KR, Samek W (2017) Explaining recurrent neural network predictions in sentiment analysis. arXiv. Doi: https://doi.org/10.18653/v1/w17-5221.
24. Elbagir S, Yang J (2018) Sentiment analysis of twitter data using machine learning techniques and scikit-learn. ACM Int. Conf. Proceeding Ser., pp. 0–5. Doi: https://doi.org/10.1145/3302425.3302492.