Journal of Engineering and Applied Sciences
MEDWELL PUBLICATIONS

Machine Learning Approach to Classify the Sentiment Value of Natural Language Processing in Telugu Data

Palli Suryachandra, P. Venkata and Subba Reddy
Department of CSE, SVUCE, SVU, Tirupati, India

Key words: Natural language processing, Telugu, sentiment analysis, polarity, machine learning, optimization

Corresponding Author:
Palli Suryachandra
Department of CSE, SVUCE, SVU, Tirupati, India

Page No.: 3593-3598
Volume: 15, Issue 21, 2020
ISSN: 1816-949x
Journal of Engineering and Applied Sciences

Abstract: Natural Language Processing (NLP) is a computer software utilized in big information programs to specify the consumer review additionally, it's a major a part of AI as a result NLP strategy is processed with numerous languages because the language is differing from kingdom to country. Furthermore, sentiment analysis in NLP is superior in lots of programs and various languages to evaluate the sentiment but part of speech specification for distinctive language is simply too tough. to overcome this trouble the contemporary studies aimed to expand a singular Evolving C4.5 machine mastering with Spider Monkey Optimization (EC4.5-ML-SMO) to categorise the sentiment evaluation in Telugu language effectively. Furthermore, the fitness feature of SMO more desirable the accuracy of sentiment class in Telugu dataset. Subsequently, the effectiveness of proposed module is evaluated with current existing works and attained better end result via. reaching better class accuracy rate.

INTRODUCTION

These days, nicely-informed selections are required in all specialists area, for that reason the whole expertise is wanted to evaluate consumer reviews for all huge records applications. It's far increasingly difficult venture whilst it done manually[1]. To lessen this difficultness NLP is delivered, now the NLP can rule the massive facts industry for lots functions along with, question and answering system, semantic analysis, sentiment evaluation and so on[2] in Fig. 1 additionally, the NLP strategy is applicable for all languages it may feature the process with the help of gadget studying version[3]. Beside this entire element the net business are run effectively with the purchaser satisfactions[4] for that reason the categorization of sentiment fee for each and each customer assessment is extra vital[5]. Further, processing huge amount of facts became a chance in NLP[6] due to the fact the most of collected information are unstructured consisting of news article; software opinions web blogs and so forth[7].

Moreover, the NLP completed nicely in marketing evaluation, competitive analysis and locating unsuccessful gossip for threat control[8] in massive records surroundings. Sentiment analysis in Natural Language Processing (NLP)[9] is a complicated undertaking that distribute with unstructured textual content and classifies it as both a wonderful, terrible or impartial sentiment[10]. Sentiment evaluation is the portion of text mining that tries to give an explanation for the opinions[11], feelings and attitudes present in a text or a fixed of textual content[12]. Many researchers find numerous system

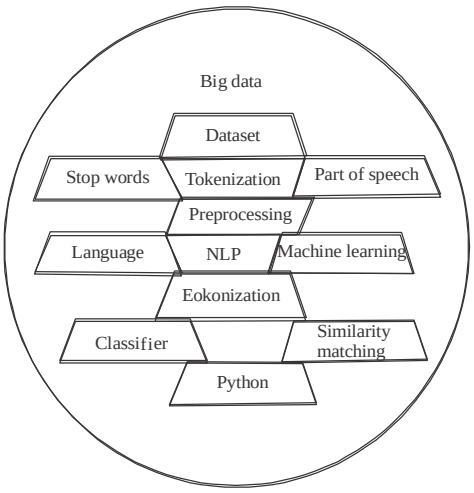

Fig. 1: System model

studying models till the specification of sentiment price is a query mark[13] because of one-of-a-kind unstructured dataset with distinctive languages[14]. Accordingly the improvement of hybrid gadget studying mechanism with a few gamming approach can improve the sentiment class method. The rest this research approach is organized as follows: module 2 exact a few latest related literatures of NLP module 3 describes trouble assertion, module 4 elaborated the proposed approach and module 5 certain the final results of the proposed method and its contrast and module 6 concludes this studies.

Literature review: A number of the recent literatures associated with sentiment evaluation in NLP is summarized underneath, a prime studies scheme certain as like one-on-one interviews represent an expensively applied approach according with reap meaningful perceptions and make whole conclusions as is subordinate with the aid of the influx over technology, client demand is skilled thru one-on-one interviews take after at the crucial product traits, start strategies and pricing as a result Manojkumar Parmar etc. proposed a method in conformity to create sentiment evaluation and execute different quantitative strategies. This technique is expanded in imitation of carry out the question-sensible complete evaluation due to the fact better perception. The authors have the idea in conformity with make bigger the amount on the information elements or additionally execute the weighted average analysis. The approach generated may be utilized according to discover outlier interviews in imitation of broaden learning due to researchers amongst quite efficient manner.

Word embeddings suggests the words in a vocabulary as real-valued vectors in a multidimensional area. They are educated utilizing a big set of unlabeled facts and formulated as real-valued vectors primarily based at the word look contexts word embeddings can capture syntactic and semantic information with out using categorized records and as a result they are usefully carried out in lots of natural language processing tasks which includes information retrieval, statistics extraction text class, sentiment analysis, query answering and device translation. Consequently, Duc-Hong Pham and Le[15] proposed a technique the way to integrate diverse representations of enter for the hassle of aspect-based sentiment evaluation. Recollect that this pattern may be helpful for several sentiment analysis troubles which includes issue ratings detection. in addition, this pattern will be implemented successfully in conformity with languages ignoble than English.

With an upsurge between communal media usage and online disclosure approximately ethnical reviews or evaluations, the issue on SA has turn out to be the point of interest of NLP researchers everywhere in the world. as a result Khurana and Sahu *et al*.[16] proposed a approach in accordance with enforce a supervised discipline approach in line with function sentiment analysis. The research receives input from a broadly utilized micro-running a blog website: Twitter as serves as a suited database due to the fact the project handy. It determined that precision may be extraordinarily evolved if the amount of sentiment set is reduced according to entirely: tremendous and poor.

In the preceding bit decades, there has been excessive call for from special businesses and agencies in conformity to get admission for applicable data extra flexibly as like mining such facts beside more than one disported sources has been a very best vicinity on evaluation and problem. Once the answer approach in conformity with this issue has been textual content extraction where in records can be labeled based concerning concord homes, therefore, Chandra *et al*.[17] proposed an technique in conformity to stumble on the comparable text via. natural call processing techniques by way of making use of textual content mining methods, text blocks can be condensed to separate the set of files by is evaluated thru processing difficulty of text files.

Some summarizer creates summaries by manner of the calculating devices, maintaining its crucial abilities and factors is referred to as computerized summarizer. subsequently Mandal *et al*.[18] proposed a method makes a speciality of the approach due to retrieving the facts inside compact shape or summarizes form. The simple wondering is in conformity with choose the deserving cluster afterwards great range and insurance constraints arranging the sentences within the cluster among honor in accordance with sentiment score in reducing order.

Problem statement: Commonly, the sentiment analysis in natural language processing is done over big facts dataset inclusive of Facebook, Twitter and so on moreover, sentiment evaluation for Telugu language is a few more difficult as because of its complexity and part of speech classification.

In addition, the sentence which contains high quality phrases may also end with bad sentence. Also, the type of opinion in huge volume of dataset is just too difficult. accordingly, the category of sentiment degree is more critical. This inspire this research to find the clinical approach to enhance big information analytics the use of sentiment evaluation in Telugu natural language processing to reduce all types of problems. The system model of NLP in big data is shown in Fig. 1.

MATERIALS AND METHODS

Proposed methodology: Sentiment evaluation for Telugu language in NLP is the important assignment due to its a part of Speech (PoS) type. So, this research introduces the unconventional evolving C4.5 system getting to know (EC4.5ML) algorithm to make the class process easier by using decreasing the similarity of sentence or phrases and blunders.

Moreover, the sentiment analysis is done in the manner of neutral, advantageous and bad type. subsequently, the accuracy of category is stepped forward by means of Spider Monkey Optimization (SMO) set of rules. The process of proposed methodology is shown in Fig. 2.

Evolving C4.5-SMO machine learning: To process the sentiment analysis, initially the data is trained to the system consequently the error is removed. The training error is represented as na. The error removing or preprocessing function in machine learning can enhance the classification accuracy. Thus the data preprocessing function is performed by Eq. 1. Here, E is the dataset:

$$E = \sum_{a=1}^{m} \text{-na} \tag{1}$$

After the error removing process, the data is specified as tree like structure to proceed the further process. Thus the tree nodes are determined as b = 1, 2,..., m in Eq. 2:

$$Db = \text{sum}\left(\text{pqr}\left(GX\text{-}v1(b)\right)\right)/2\text{sum}\left(V1(b)\right) \tag{2}$$

Here, the membership function of node is determined as GX also radbas Db evaluate the maximum sentiment value of reviews.

One of the recent inspired algorithm is known as Spider Monkey Optimization Which is characterized by

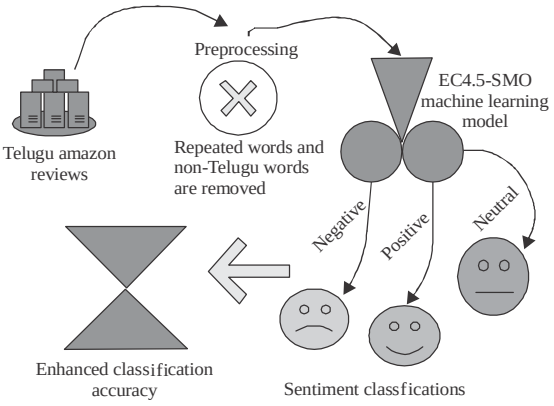

Fig. 2: Proposed methodology

Swarm intelligence strategies. In this current research this procedure is used to enhance the accuracy of sentiment classification. Moreover, the bio inspired model has its own fitness based on its behavior. That fitness function is utilized in machine learning classification layer, finally, the improved classification accuracy is obtained.

The monitoring phase of SMO Model is processed based on several groups, for each group a specific leader is elected. Here, this function is utilized to choose the aspect terms. Where, the error removed dataset Telugu Amazon review is represented as mm, then p, q, c are the aspect terms also the sentiment classification parameter is denoted as C. Moreover, '0' represent neutral sentiment value, '1' represent positive sentiment value and '-1' represent negative sentiment value. Thus the preprocessed data set is trained in the form of Eq. 3:

$$\begin{aligned} mn_{new}q = &\left\{mn_{pq} + Db(0,1)\times\left(ll_t\text{-}mn_{pq}\right) + Db(\text{-}1,1)\times\left(mn_{pq}\text{-}mn_{pq}\right)\right. \\ &\text{if } DbC(0,1) \geq ic \qquad mn_{pq} \qquad \text{Otherwise} \end{aligned} \tag{3}$$

To enhance the sentiment classification, the fitness function of spider monkey is upgraded in machine learning classification model. Based on the probability of aspect terms the sentiment classification is performed, the validation aspect terms probability is detailed in Eq. 4:

$$\text{Probability}_p = \left(\frac{\text{Sentence}}{\text{aspect terms}}\right) \tag{4}$$

$$\begin{aligned} mn_{newp} = &\, mn_{pq} + Db(0,1)\times\left(fl_q\text{-}mn_{pq}\right) + \\ &Db(\text{-}1,-1)\times\left(mn_{cq}\text{-}mn_{pq}\right) \end{aligned} \tag{5}$$

Where, the specified sentence is determined as fl the sentiment value for each review is estimated using (Eq. 5):

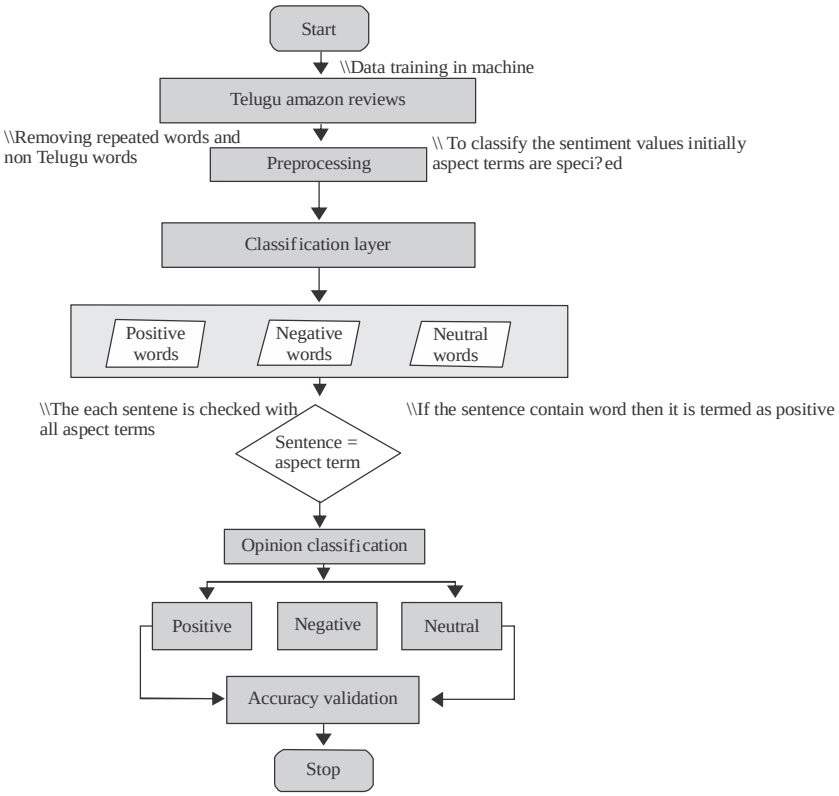

Fig. 3: Function of proposed methodology

The working process of proposed methodology is shown in Fig. 3, initially the set of reviews are trained to the system. Consequently the errors are removed in the filtering module of machine learning. Then, the sentiment aspect terms are specified and stored in the classification layer. Subsequently, the classification is done based on the aspect terms.

RESULTS AND DISCUSSION

The proposed approach is elaborated in python going for walks in windows 10 platform. The technique of sentiment evaluation is the specification of humans opinion which is found in on-line offerings. Moreover, the sentiment categorization is completed using a few set of words that carries the sentiment price as impartial, fine and terrible right, here, the dataset evaluated in this paintings are Amazon critiques, for that reason the opinions specification is based on the polarity classification which is fantastic poor and neutral. to begin with general set of words are educate to the device the sentence are split in to words like decision tree order meaning it has root node and branches eventually unwanted branches are pruned to make the polarity specification manner simpler.

Case study: In this proposed approach, Telugu language for Amazon reviews is taken for implementation; some samples are shown in Table 1. Initially, the Telugu reviews are trained to the system.

The process and feature of sentiment category is elaborated in Fig. 3. The gadget can't understand human language, so, its schooling and process is functioned within the manner of zero's and 1's.

Performance metrics: To validate the effectiveness of the proposed system some of the recent research works are adopted such as Term Frequency-Inverse Document Frequency with Support Vector Machine (TF-IDF-SVM)[19] and cluster Named Entities (NEs) extracted from Telugu corpus based on semantic similarity[20] (CNES).

Accuracy: The overall performance validation of system getting to know technique is finished by using comparing the classification accuracy based on actual advantageous, real negative, false high-quality and fake negative. The evaluation validation of accuracy for sentiment type is shown in Table 2:

$$\text{Accuracy} = \frac{(TN+TP)}{TN+TP+FN+FP}$$

Table 1: Telugu sentences and its polarity

1	మీరుఆడిందిఆటలయితే, దానిసేపథ్యసంగీతంనాకునచ్చదుపాటలన్నీ ఒకసందేహంతోకూడిఉంటాయి.	If you play the game, I don't like its background music.	-	-	-1
2	కథఅద్భుతంగాఉంది, పుస్తకంలోవచ్చినమార్పులను చూసినాకుచాలాఆశ్చర్యంవేసింది. ఈపుస్తకంఎంతఅద్భుతమైనాగానీ, నామిత్రులందరికీఅదిచదవమనిచెప్పాను.	The story is amazing, and I was very surprised to see the changes in the book. No matter how wonderful this book is, I told all my friends to read it	+1		-

Table 2: Accuracy comparison

Reviews taken	CNES	TF-IDF-SVM	RNN	Proposed
60	83	79	84	98.5
120	84	76	84	98.0
180	81	73	82	97.0
240	79	71	81	96.0

Table 3: Precision comparison

Precision

Reviews taken	CNES	TF-IDF-SVD	RNN	Proposed
60	85	79	85.0	98.0
120	83	76	84.0	97.9
180	81	72	82.1	98.2
240	80	71	81.0	97.2

Precision: The precision of processed data is estimated as the number of accurate specific sentiment predictions by the total number of sentiment sentences. Here, the precision rate is calculated for each set of reviews:

$$\text{Precision} = \frac{TP}{TP+FP}$$

The evaluation validation of precision for sentiment type is shown in Table 3.

Recall: The recall is calculated as the number of exact positive values divided by the whole number of true positives and false negatives. Recall sentiment evaluation in NLP is evaluated as whole document intersection of separated sentence divided by polarity values (positive, negative and neutral):

$$\text{Recall}(T) = \frac{\text{Telugu sentence amazon reviews}}{\text{Polarity value}}$$

The evaluation validation of recall for sentiment type is shown in Table 4.

F-measure: The F-measure is validated to verify the mean average for precision and recall, thus the comparison of F-measure:

Table 4: Recall comparison with existing approaches

Recall

Reviews taken	CNES	TF-IDF-SVD	RNN	Proposed
60	85	79	85.0	99.0
120	83	75	84.0	98.7
180	80	74	81.0	98.0
240	79	71	81.7	97.0

Table 5: Comparison of f-measure

F-measure

Reviews taken	CNES	TF-IDF-SVD	RNN	Proposed
60	85.0	79	85	99.0
120	82.4	75	84	98.4
180	80.0	74	81	98.2
240	79.0	71	81	97.0

$$\text{F-measure} = 2\times\frac{\text{Percision}\times\text{recall}}{\text{Percision}\times\text{recall}}$$

The average of F-measure is calculated by taking between the average of accuracy and precision. To verify the accuracy of classification F-measure is evaluated. High accuracy and precision yields better F-measure rate. The proposed strategy attained 97% as F-measure rate for 200 reviews simultaneously, the existing approach TF-IDF-SVM gained 70%, CNES achieved the F-measure rate as 81.5% and RNN attained 80% of F-measure rate which is defined in Table 5.

CONCLUSION

In massive data region, system studying method is one of the trending area consequently the opinion or sentiment class is one of the essential tasks in NLP that is generally helpful for on-line services. So, the present work advanced a unique hybrid system gaining knowledge of model to validate the customer overview in Telugu dataset. Furthermore, the health version of optimization enables to improve the sentiment classification charge consequently, the attained sentiment category accuracy the use of system learning and heuristic model is 97%. Furthermore, the assessment results proved the efficiency of the proposed paintings. Therefore, the evolved model is relevant for on line offerings to classify the reviews of each clients also it enables to improve the web services.

REFERENCES

01. Carvalho, A., A. Levitt, S. Levitt, E. Khaddam and J. Benamati, 2019. Off-the-shelf artificial intelligence technologies for sentiment and emotion analysis: A tutorial on using IBM natural language processing. Commun. Assoc. Inf. Syst., Vol. 44, 10.17705/1CAIS.04443.
02. Marie-Sainte, S.L., N. Alalyani, S. Alotaibi, S. Ghouzali and I. Abunadi, 2018. Arabic natural language processing and machine learning-based systems. IEEE. Access, 7: 7011-7020.

03. Yang, H., L. Luo, L.P. Chueng, D. Ling and F. Chin, 2019a. Deep Learning and its Applications to Natural Language Processing. In: Deep Learning: Fundamentals, Theory and Applications, Huang, K., A. Hussain, Q.F. Wang and R. Zhang (Eds.). Springer, Switzerland, pp: 89-109.
04. Nasukawa, T. and J. Yi, 2003. Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the 2nd International Conference on Knowledge Capture, October 23-25, 2003, Sanibel Island, FL., USA., pp: 70-77.
05. Alayba, A.M., V. Palade, M. England and R. Iqbal, 2018. Improving sentiment analysis in Arabic using word representation. Proceedings of the 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), March 12-14, 2018, IEEE, London, UK., pp: 13-18.
06. Al Amrani, Y., M. Lazaar and K.E. El Kadiri, 2018. Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Comput. Sci., 127: 511-520.
07. Ma, Y., H. Peng, T. Khan, E. Cambria and A. Hussain, 2018. Sentic LSTM: A hybrid network for targeted aspect-based sentiment analysis. Cognit. Comput., 10: 639-650.
08. Ahuja, R., A. Chug, S. Kohli, S. Gupta and P. Ahuja, 2019. The impact of features extraction on the sentiment analysis. Procedia Comput. Sci., 152: 341-348.
09. Hasan, M., I. Islam and K.A. Hasan, 2019. Sentiment analysis using out of core learning. Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), February 7-9, 2019, IEEE, Bangladesh, pp: 1-6.
10. Young, T., D. Hazarika, S. Poria and E. Cambria, 2018. Recent trends in deep learning based natural language processing. IEEE. Comput. Intell. Mag., 13: 55-75.
11. Zhang, L., S. Wang and B. Liu, 2018. Deep learning for sentiment analysis: A survey. Wiley Interdiscip. Rev. Data Mining Knowl. Discovery, Vol. 8, No. 4. 10.1002/widm.1253.
12. Zhang, Y., D. Miao, J. Wang and Z. Zhang, 2019. A cost-sensitive three-way combination technique for ensemble learning in sentiment classification. Int. J. Approximate Reasoning, 105: 85-97.
13. Yang, C., H. Zhang, B. Jiang and K. Li, 2019b. Aspect-based sentiment analysis with alternating coattention networks. Inf. Process. Manage., 56: 463-478.
14. Chiranjeevi, P., D.T. Santosh and B. Vishnuvardhan, 2019. Survey on Sentiment Analysis Methods for Reputation Evaluation. In: Cognitive Informatics and Soft Computing, Mallick, P., V. Balas, A. Bhoi and A. Zobaa (Eds.). Springer, Singapore, pp: 53-66.
15. Pham, D.H. and A.C. Le, 2018. Exploiting multiple word embeddings and one-hot character vectors for aspect-based sentiment analysis. Int. J. Approximate Reasoning, 103: 1-10.
16. Khurana, H. and S.K. Sahu, 2018. Bat Inspired Sentiment Analysis of Twitter Data. In: Progress in Advanced Computing and Intelligent Engineering, Saeed, K., N. Chaki, B. Pati, S. Bakshi and D. Mohapatra (Eds.). Springer, Singapore, pp: 639-650.
17. Chandra, N., S.K. Khatri and S. Som, 2019. Natural Language Processing Approach to Identify Analogous Data in Offline Data Repository. In: System Performance and Management Analytics, Kapur, P., Y. Klochkov, A. Verma and G. Singh (Eds.). Springer, Singapore, pp: 65-76.
18. Mandal, S., G.K. Singh and A. Pal, 2019. PSO-Based Text Summarization Approach Using Sentiment Analysis. In: Computing, Communication and Signal Processing, Iyer, B., S. Nalbalwar and N. Pathak (Eds.). Springer, Singapore, pp: 845-854.
19. Reddy, D.A., M.A. Kumar and K.P. Soman, 2019. Paraphrase Identification in Telugu Using Machine Learning. In: Advances in Big Data and Cloud Computing, Peter, J., A. Alavi and B. Javadi (Eds.). Springer, Singapore, pp: 499-508.
20. Gorla, S., A. Chandrashekhar, N.B. Murthy and A. Malapati, 2019. TelNEClus: Telugu Named Entity Clustering Using Semantic Similarity. In: Computational Intelligence: Theories, Applications and Future Directions-Volume II, Verma, N. and A. Ghosh (Eds.). Springer, Singapore, pp: 39-52.