International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-10 Issue-8, June 2021

# Sentiment Analysis using Deep Belief Network for User Rating Classification

**Ravi Chandra, Basavaraj Vaddatti**

***Abstract:*** ***People's attitudes, opinions, feelings and sentiments which are usually expressed in the written languages are studied by using a well known concept called the sentiment analysis. The emotions are expressed at various different levels like document, sentence and phrase level are studied by using the sentiment analysis approach. The sentiment analysis combined with the Deep learning methodologies achieves the greater classification in a larger dataset. The proposed approach and methods are Sentiment Analysis and deep belief networks, these are used to process the user reviews and to give rise to a possible classification for recommendations system for the user. The user assessment classification can be progressed by applying noise reduction or pre-processing to the system dataset. Further by the input nodes the system uses an exploration of user's sentiments to build a feature vector. Finally, the data learning is achieved for the suggestions; by using deep belief network. The prototypical achieves superior precision and accuracy when compared with the LSTM and SVM algorithms.***

***Keywords: Bag of Words, Deep Belief Network, Restricted Boltzmann machine, Term Frequency/Inverse Document Frequency, Word2Vector, Support Vector Machine, and Long Short-Term Memory.***

## I. INTRODUCTION

Individuals communicate their thoughts and feelings regarding a wide range of issues on microblogging websites every day, including items, media, organizations, and so on. For the recommender system users, reviews or comments are much important because they contain the various types of data related to emotions that may influence the precision or correctness of the recommendation [01]. Sentiment Analysis is a type of artificial intelligence that uses text mining to investigate user opinions, behavior, and emotions. Sentiment analysis is useful not just for product reviews, but also for other areas such as stock markets, news articles, and so on. The sentiment analysis has sought greater attention with the success of deep learning [02]. Deep learning may be able to build deep models of complicated multivariate patterns in big datasets. Deep learning has been applied across many fields due to the availability of neural networks, which this lead to significant successes in many applications [02]. Due to the complexity in human languages, vocabularies have such a large scale dimension which is mostly noisy or redundant. DBN overcomes this problem by using various hidden layers to acquire meaningful information from the input corpus.

User comments are used to categorize user ratings in the proposed method. The numerical value that people assign to a product or service determines the user rating categorization. Ordinarily, Generative and discriminative classifiers are the two main types of classifiers. Discriminative classifiers are the one which uses a function that relates to class label in the input set [01]. "The proposed model is a discriminative classifier which combines both the DBN and Sentiment analysis" [01]. Then uses deep learning models like Restricted Boltzmann machine (RBM) to train a model and to guess user reviews classifications. The feature extraction is done by using the word2vector in turn it combines the Term Frequency/Inverse Document Frequency (TF-IDF), and the Bag of Words (BOW). The train of a classification model is very much faster and also gets better performance when compared with the existing Long Short Term Memory (LSTM) and Support Vector Machine (SVM) algorithms.

The remaining portion of this paper includes the following sections: Section II, Section III, Section IV and lastly the Section V. Whereas the Section II is all about the literature review and Section III about the proposed system and Section IV gives the brief description about the experimental results and final one is Section V gives quick glance about the conclusion and future work.

## II. LITERATURE SURVEY

The authors of [01] presented sentiment analysis via the Deep Belief network with noise reduction. The rating prediction is from the user comments and is implemented in three steps firstly the noise reduction as pre-processing. Secondly, the model uses Sentiword for feature extractions. Thirdly it uses a deep learning algorithm like DBN. The system achieves an accuracy rate of 72.83 compared high with the baseline models MLP, CNN & LibSVM.

The author of [03] concentrates on the two major halal products i.e. halal tourism and halal cosmetics. These products reviews are collected in tweets. Deep learning algorithms were used to calculate and evaluate tweets or emotions in the experiments. To enhance the accuracy and build prediction models, CNN and LSTM were used.

Manuscript received on June 16, 2021.
Revised Manuscript received on June 23, 2021.
Manuscript published on June 30, 2021.
* Correspondence Author
**Ravi Chandra***, Student, Department of Computer Science and Engineering, Shri Dharmasthala Manjunatheshwara College of Engineering and Technology, Dharwad (Karnataka), India. E-mail: ravikajenkar@gmail.com
**Basavaraj Vaddatti**, Assistant Professor, Department of Computer Science and Engineering, Shri Dharmasthala Manjunatheshwara College of Engineering and Technology, Dharwad (Karnataka), India. E-mail: basavaraj.vaddatti@gmail.com

With a stacking of CNN and LSTM techniques, along with the word2vec extracting features obtained the highest accuracy of 93.78 percent. In [04], the author focuses mainly on the semi-supervised learning algorithm for the DBN called DBN with Feature selection. Where some of the irrelevant features are filtered making the model more efficient.

The result shows that the model achieves higher classification accuracy and speedups the training time.

In [05], the authors introduce two methods "the first method presented to find these aspect classes, the unsupervised technique uses association rule mining on the co-occurrence frequency data which are retrieved from a corpus and second uses the supervised method which is straightforward co-occurrence method where the conditional probabilities are calculated using the co-occurrence regularity amid annotated aspect classes and dependencies and both lemmas" [05]. Evaluation results show on the officialSemEval-2014 where the test set has a high F1-score value of 83%.

In [06], the authors described how to tackle the sentiment classification task in imbalanced datasets using the transfer learning. It incorporates fine-tuning of under-sampling and pre-trained model of text information built on transfer learning [06]. The method's performance is based on real-world sentiment detection dataset. Although the model only employs transfer learning based under sampling for sentiment analysis and it functions post verification, it is uncertain whether this will function for various text categorization.

## III. PROPOSED SYSTEM

The system starts by gathering customer feedback from product reviews. The collected data are pre-processed to reduce the noise at the initial stages. After pre-processing the dataset feature extraction word2vec is applied with the combination of both the BOW and TF-ID algorithms. Once the feature vector is generated it is sent as the input to the DBN training model.

The DBN models are used to learn each rating with a deep learning training model. To learn the important and relevant features in the input node we use the Restricted Boltzmann Machine (RBM) and secondly, it takes the output from the above layers of the RBM and process has the input to the next immediate layer this process is done in the loop manner till the final result is achieved in the output layer. The complete process is achieved through the Deep Belief Network algorithm.

Fig 01 depicts the proposed system's architecture. The suggested module is divided into four stages, as shown below.

### A. Collecting Data

Collection of data for training and testing the classifiers is the initial step in the implementation phase. The data received from Amazon reviews. There are two versions of the dataset one in the form of CSV file and another in SQLite Database. The SQLite dataset is selected to load the data because it allows for more effective data querying and visualization. The dataset is initially filtered to retrieve only the

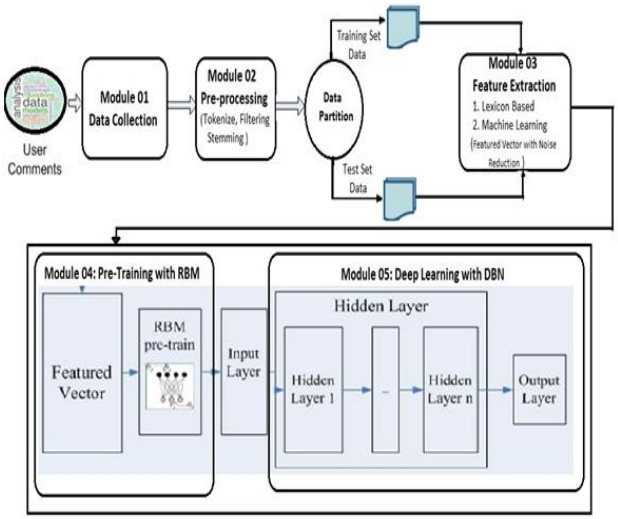


**Fig 01: Overall System Architecture**

Positive reviews and negative reviews just for implementing the SVM and LSTM algorithm but in the case of DBN all three categories of reviews are considered i.e. positive, negative, and neutral reviews. Each review includes many columns and few important columns are shown in Table-I.

**Table- I: Important fields in the dataset.**

| | |
|---|---|
| ProductId | Product ID is given by ASIN, bar-codes(EAN,UPC ISBN etc) |
| UserId | The user's id. |
| ProfileName | The customer name |
| HelpfulnessNumerator | This tells how many people found this review useful. |
| HelpfulnessDenominator | This tells how many people found this review useful and not useful. |
| Score | Rating of product from 1 to 5 |
| Time | The reviews date and time |
| Summary | The summary of the review |
| Text | The text specified in the review |

Exploratory Data Analysis approach that reveals the important characteristic of the dataset which are quite normally not visible through traditional close reading or it just summarizes the data. The data cleaning strategy entails preparing the data needed with a simple python code that removes the unnecessary elements. These approaches are applied before the data begins for the feature extraction.

### B. Pre-Processing Data

This pre-processing is done on both the review text as well as the review summary based on the requirement this is as follows.

- Tokenization is done at sentence level this is known as converting text into tokens before transforming it into vectors.
- Eliminate the HTML tags.

- Special characters like,. # $ & * etc and also any punctuations are removed.
- Checking that the word is not alpha-numeric and is made up of English letters.
- A word's size must be greater than 02 characters (There are no adjectives in two-letter words, according to studies).
- Reduce the word's capitalization to lowercase ones.
- Remove away of stop words and add some of the extra stop words removal.
- Lemmatization identifies the lemma, which is the base or dictionary form of the term.
- Snowball stemming the word it was found to be superior to porter stemming.
- Words that are shortened by dropping the letters and by replacing them with an apostrophe, this type of words are decontracted to their original full form.
- At last, the Substitution is done by doing the Decoding and UTF-8 Character Bytes.

### C. Feature Extraction

The main goal is to create the new features in the existing dataset by reducing the number of features in a dataset and then discarding the original features [07]. The feature extraction is done in a general way which gives better performance results over the classification algorithms. The Word2Vec method uses training data to build a vocabulary.

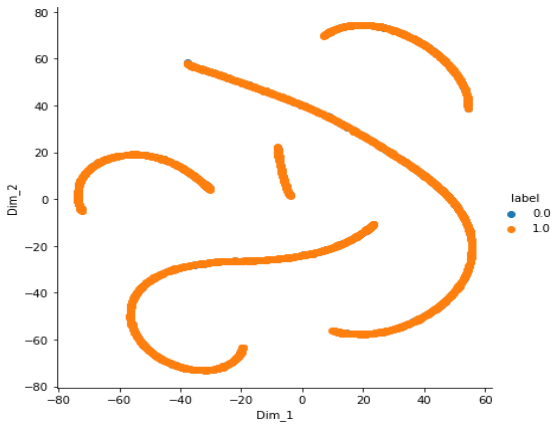


**Fig 02: TSNE applied on the TF-IDF weighted word2vec.**

Then it learns as well as identifies each word's vector representation. Word2vec is a similar meaning to have the vector representation. Now, for the better result the BOW and TF-IDF are calculated on the average word2vec this is done just for achieving good performance for the classification algorithms. Once the TF-IDF is done on the word2vec, then the TSNE (t-Distributed Stochastic Neighbor Embedding) is applied to showcase the positive and negative reviews distributions by extracting the main feature. TSNE applied on the TF-IDF weighted word2vec for the pre-processed dataset the result is shown in Fig 02.

### D. Classifications Models

This section discusses the classifier models that have been used to classify reviews into positive, negative, and neutral categories during the training and testing phase. We have used LSTM, SVM, and DBN. The DBN outperforms the competition, with a 95% accuracy rate.

**LSTM:** It is a Recurrent Neural Network unit. Such networks are highly adapted for classifying, processing, and making predictions placed on time series data. In our study LSTM is implemented with one layer as well as two layers. The only difference observed is increasing in the classification accuracy.

**SVM:** It is the most widely used algorithm available in the supervised machine-learning category. This is a classification method for the linear and non - linear information [08]. In our study the SVM is implemented by using BOW and TF-IDF. Where SVM using BOW performs better compared to SVM using TF-IDF.

**DBN:** It is the most extensively used and fast unsupervised algorithm available in machine-learning. It is a graphical model built of numerous layers of hidden variables that is used to abstract higher representations from raw inputs, and it is one of the most significant deep learning models. A basic DBN has one RBM at the top two adjacent hidden layers of the network, in which the units between two layers are bidirectional and there are no connections within the same layer.

## IV. EXPERIMENTAL OBSERVATIONS

### A. Dataset and Evaluation Metrics

In the initial stage, the data preparation and pre-processing work are done to achieve better performance. The total reviews available in our dataset are around 525814 after pre-processed the available data for the feature extraction and classification is 70% of the original dataset. The 90% of available dataset is used to train the model, with the remaining 10% being used in the testing phase. We used the two specific evaluation indicators to assess the recommended method's performance:

**Accuracy**: The accuracy of a model is calculated as the proportion of number of accurately predicted data samples to the total of sample data.

$$Accuracy = \frac{Number\ of\ correct\ predicitions}{Total\ number\ of\ data\ samples} \quad (1)$$

**Precision:** The precision of a model is calculated as the value obtained by number of correctly anticipated positive sample data over the total number of positive sample data.

$$Precision = \frac{Number\ of\ correct\ positive\ prediction}{Total\ number\ of\ positive\ sample\ data} \quad (2)$$

### B. Experiment Results

The model uses the dataset consisting the positive reviews of 443766, negative reviews of 82007, and neutral reviews of 42638. The complete distribution of these reviews among the three classes is shown in Fig 03.


Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

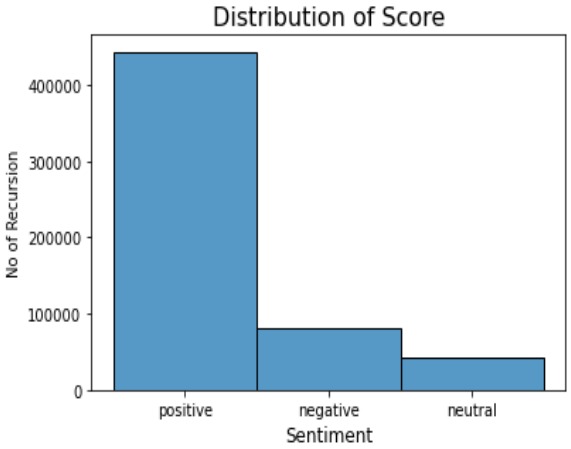


**Fig 03: Distribution of sentiments.**

The rating distribution from one to five stars in the available dataset is as shown in Fig 04. These ratings consist of positive, negative, and neutral reviews.

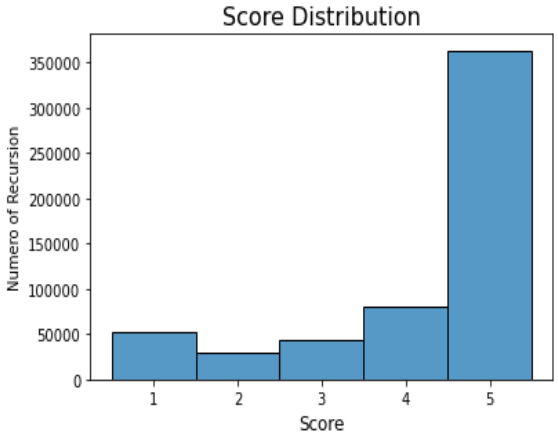


**Fig 04: Distribution of score in the reviews.**

Following Table-II indicates the accuracy and precision of the five models where LSTM with one and two layer achieves accuracy up to 92-93% with 10 epochs. The SVM using BOW obtains up to 89% of accuracy and SVM using TF-IDF obtains up to 88% of accuracy. Lastly proposed DBN model using the word2vec achieves accuracy up to 95% with 15 epochs. The accuracy and precision of five models is shown in the Fig 05. In this figure, the blue portion indicates model's accuracy where as red portion indicates model's precision. Experimental results show that the DBN model performs better. We have achieved accuracy up to 95%, which is higher than the accuracy rate compared with other models.

**Table- II: Accuracy and precision of model comparison.**

| Model | Accuracy | Precision |
|---|---|---|
| LSTM-01 layer | 92% | 75% |
| LSTM-02 layer | 93% | 78% |
| SVM-BOW | 89% | 88% |
| SVM-TFID | 88% | 90% |
| DBN | 95% | 94% |

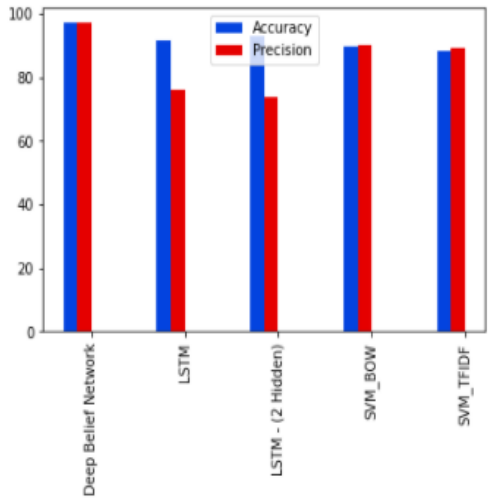


**Fig 05: comparison result of five models with accuracy and precision.**

## V. CONCLUSION

Using a DBN and sentiment analysis, we give better classification technique for user ratings from user reviews. We collected features from the reviews, constructed a dataset, trained five distinct classifiers, and afterwards chosen the best model based on the accuracy and precision in this study. In future the existing model can be enhanced to include sarcasm and the emotion classification, not just positive, negative, and neutral.

## REFERENCES

1. R. Chen and Hendry, "User Rating Classification via Deep Belief Network Learning and Sentiment Analysis," in IEEE Transactions on Computational Social Systems, vol. 6, no. 3, pp. 535-546, June 2019, doi: 10.1109/TCSS.2019.2915543.
2. S. Seo, C. Kim, H. Kim, K. Mo and P. Kang, "Comparative Study of Deep Learning-Based Sentiment Classification," in IEEE Access, vol. 8, pp. 6861-6875, 2020, doi: 10.1109/ACCESS.2019.2963426.
3. A. Feizollah, S. Ainin, N. B. Anuar, N. A. B. Abdullah and M. Hazim, "Halal Products on Twitter: Data Extraction and Sentiment Analysis Using Stack of Deep Learning Algorithms," in IEEE Access, vol. 7, pp. 83354-83362, 2019, doi: 10.1109/ACCESS.2019.2923275.
4. P. Ruangkanokmas, T. Achalakul and K. Akkarajitsakul, "Deep Belief Networks with Feature Selection for Sentiment Classification," 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), 2016, pp. 9-14, doi: 10.1109/ISMS.2016.9.
5. K. Schouten, O. van der Weijde, F. Frasincar and R. Dekker, "Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis with Co-occurrence Data," in IEEE Transactions on Cybernetics, vol. 48, no. 4, pp. 1263-1275, April 2018, doi: 10.1109/TCYB.2017.2688801.
6. Z. Xiao, L. Wang and J. Y. Du, "Improving the Performance of Sentiment Classification on Imbalanced Datasets With Transfer Learning," in IEEE Access, vol. 7, pp. 28281-28290, 2019, doi: 10.1109/ACCESS.2019.2892094.
7. Dang , N.C.; Moreno-García , M.N. ; De la Prieta , F. " Sentiment Analysis Based on Deep Learning : A Comparative Study ". Electronics 2020 , 9 , 483. https://doi.org/10.3390/electronics9030483
8. S. Khan, K. Chopra and P. Sharma, "Brand Review Prediction using User Sentiments: Machine Learning Algorithm," 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-8, doi: 10.1109/IDEA49133.2020.9170730.
9. Machine Learning in Python, https://scikit-learn.org.
10. Datasets Repository, https://www.kaggle.com.

## AUTHORS PROFILE

**Ravi Chandra,** M.Tech Learner, from the Dept. Of CSE, SDMCET college, Dhavalagiri, Dharwad, Karnataka, India-580002. He has done his BE in Dr. Ambedkar Institute of Technology, Near, Gnana Bharathi Outer Ring Road, Bengaluru, Karnataka 560056. He has a teaching experience of 09 years in the Govt. Polytechnic Lingasugur, Karnataka 584122.

**Prof. Basavaraj Vaddatti,** Assistant Professor in the Dept of CSE at SDMCET, Dharwad, Karnataka, India. He completed his BE (CSE) in 2000 from **UVCE** and M.Tech (CSE) in 2008 from Dr. Ambedkar Institute of Technology, Bengaluru. Under his guidance around **10** numbers of PG and UG students have successfully completed. He published two papers at international journals/conferences. He actively conducted/Participated in about fifteen numbers of Workshops and conferences for students. His special areas of interests are image processing and language processing. His has 10 years of teaching experiences and a life member of ISTE.