STAR: rating of reviewS by exploiting variation in emoTions using trAnsfer leaRning framework

Aditya Vijayvergia and Krishan Kumar *Member, IEEE*
Department of Computer Science and Engineering
National Institute of Technology Uttarakhand, Srinagar Garhwal, India
{adi_1997.cse15, kkberwal}@nituk.ac.in

***Abstract*—In this digital era, it is a common practice to check reviews about a service or product before it buying. A five star rating scale system provides much easier interface to the consumers for checking the reviews about the corresponding service or product, instead of just classifying the reviews as good, neutral and bad. Moreover, it is very common for a single review which can praise the product and criticize it as well. Even if two reviews over all, show the same sentiment. However, the order in which they praise or criticize a product, can make their star rating quite different. We have considered such observations to deploy our proposed STAR model, which addresses the above concerns by involving the variation of sentiment in reviews, to greatly affect the star rating performance. This work highlights a two phases based novel approach using transfer learning framework to analyze the reviews by exploiting the variation in human being emotions. The experimental analysis shows that the STAR model outperforms the state-of-the-art models.**

***Index Terms*—Sentiment Analysis; Opinion Mining; Natural Learning Processing; Recurrent Neural Network; Star-scale; Transfer Learning**

I. Introduction

In this Bigdata era, the entire data may be useful for the user, hence, data needs to be mined, this is known as data mining. It is a challenging job for the consumers to select the best product or service from the huge variety available in the market. Thus a star rating concept, in this case, can be more useful to the consumers for providing the interface to see the rate in points or stars extracted from the sentiments, i.e. extracted information from the pieces of text as either negative, positive, or neutral. Star based sentiment analysis is more appropriate in the most of the marketing domains, such as movie reviews, product reviews, teaching reviews, hotel reviews, e-learning etc. Therefore, numerous applications on sentiment analysis are found in the literature such as Public opinion management, Business intelligence system, Web advertising, and Purchase planning etc. There are widely two approaches- Machine learning based and Lexicon-based to illustrate the sentiment analysis process. Machine learning based sentiment analysis models are classifying the text using classification techniques after the rigorous training on the training dataset, while the lexicon-based sentiment analysis techniques employ the orientation or sentiment of the phrases or words mentioned in a document.

In another way, data mining can be extended as opinion mining where the input data is the people's opinion pertaining to a particular topic, service, product etc. In many aspects, the process of mining, the human being opinion and the analysis of the sentiments are comparable with the behavior of the people pertaining to the product or service. Both the processes, play a critical role in decision making about the review of a product. For e.g.; in daily life, a buyer usually purchases or hires the products or services where every user wants to go for the best service or product or relatively better product; the opinion of a user in the field of sales and marketing is highly valuable, they assist the individuals or company for the popularity or judgment of the service or product as a success or a failure. Therefore, it is very interesting, but quite a difficult task to classify and summarize the sentiments or opinion as written or provided by the users. Many researches and experiments have been introduced in the field of opinion mining. However, dealing with the massive and lethal data is still an arduous job. Further extension of the opinion mining leads to the analysis of the sentiments to extract the hidden emotions or sentiments in an opinion. Therefore, a novel technique is immediately required to analyze the emotional or sentimental data efficiently and within a stipulated time.

This paper highlights a novel approach to generate a 5-star analysis on the human being opinions. The model is used a Machine learning technique (i.e. transfer learning) which is performing the exploitations of the variations in the human emotions in the given review about the product or service. The underlying concept of the model is to first mine the orientations of each sentence which is further useful in the review analysis on 0 to 1 continuous rating scale. Then, these captured ratings is used in the next phase to predict the star rating for the entire review on a particular product or service. The salient points of the proposed STAR technique follow as:

- The proposed STAR model employs the transfer learning framework to summarize and classify the product reviews. The model facilitates the service of 5-star scaling (1 to 5) which is an enhancement on the existing 3-scale concept (neutral, positive, and negative).
- Exploitation of the variations in the emotion is captured by the two-phased proposed model where each review is first segmented in phase I and the analysis is done in the second phase to predict the 5-star rating of the complete review. These two-phase make the proposed model fast.
- The experimental results showing that the proposed model outperforms the existing state-of-the-art models.

The paper is further summarized as follows: Section II includes an overview of the existing works of the opinion mining. Section III elaborates the proposed STAR model. Section IV gives an insight into the training technique and datasets used. Section V discussed the experimental setup and results on the sentimental analysis. The work is concluded in Section VI including the future perspectives.

II. Related work

Owing to the rapid growth of Internet and digital technologies, the sentiment analysis on reviews, has become a vital interest of research. During last decay, a lots of work has been performed for the analyzing the sentiments along with the facts including opinion mining in comparative sentences; classifying and summarizing the sentences as negative, positive, or neutral; identifying or detecting the subjective as well as objective and many more. A long and detailed survey is done [1], [2] which revealed numerous types of challenges as well as the applications of Sentiment Analysis (SA) and they also elaborated the details of SA techniques.

Jindal et al. [3] introduced a noel model to identify the comparative sentences from the reviews or forum or tweets posts. They considered the extraction of the entities, the comparative features and words which are employed for the comparison. However, this approach does not facilitate the concept of selecting the preferred entity. Then, Bing et al. [4] further addresses the above concerns and enhanced the accuracy of the previous one. The sentence is treated as a basic unit of information [4] and figure out the preferred entity from the comparative sentences. In order to identify the opinion and sentiment of different customers, Hu et al. [5] used about 6,800 words, a semi-supervised machine learning technique is employ to understand the polarity of these 6,800 words and categorized them into two semantic scales negative and positive (4,783 negative and 2,006 positive).

Esuli et al. [6] proposed a publicly available lexical resource to mine the human being opinions refer as SentiWordNet(SWN). There are two steps for building the SWN a) Semi-supervised learning step and b) random-walk step. Lexicons are widely utilized to categorize the words in two scales as positive and negative without worrying about their context. Dictionary base SA is proposed by LIWC [7] where dictionary compromises around 5,000 words which are classified into 76 classes. Out of these 5000 words, about 905 words are further classified in two main classes; 406 as positive and 499 as negative. However, this approach does not capable to detect the intensity differences among the sentiment, which affects the review analysis results greatly.

In order to address the above issues, an easily available lexicon is used which is refer as SenticNet(SCN) [8], for analysis of the sentiment at concept-level. SCN model uses around 14,244 common sense related ideas, for e.g., wrath, woe, etc. These concepts comprises the range value in between -1 to 1. Therefore, it fails to decide the polarity of the sentence and the exact semantics of the reviews.

In another work, Word-sense disambiguation (WSD) [9] proposed to identify the sense of a word where the word has multiple semantics. A set of normative emotional ratings is used as lexicon for around 1,034 English words [10](ANEW). Gilbert et al. [11] have introduced a Rule-based SA model, i.e., Valence Aware Dictionary for sentiment Reasoning (VADER) where its effectiveness is compared with 11 other existing state-of-the-art models. A freely available dictionary is used in VADER which contains sentiment emotions and words along with their representing/ valence values. The standard deviation is limitized with a value of 2.5, which is calculated based on the above valence values. VADER provides four scores as the result outcome on the given piece of text. The most useful metric is 'compound' score, if a single unidimensional measure of sentiment is required for a given sentence. Then this score is estimated by adding these valence scores as per each word in the lexicon, and the adjusted according per the rules. Finally normalized is done in between -1 (which is the extremely negative) and +1 (which is the extremely positive). It is further referred as the 'normalized, weighted composite score' is correct. The 'negative', 'positive', and 'neutral' scores are the ratios for proportions from the input text, that fall in each category. Therefore, these all should sum up to one, or much close to one with float precision operation. These matrices are the most beneficial if multidimensional measures of sentiment needs to be find for a given sentence.

Recently, Kumar et al. [25], proposed a novel Sentimentalizer to mine the review over Cloud using Docker utility. Moreover, LSRC model [27] discussed the Lexicon based Star Rating over the customer reviews, which is much accurate than the existing models. LSRC model is much fast owing to the uses of the Virtualized environment. The authors are well motivated by this work and decided to present a 5-rating scale model for the consumes on a nivel idea of catching and identifying the variations in the emotions in reviews, which is more difficult to capture as discussed above. A detailed discussion on the STAR model is elaborated in Section III.

III. Proposed Method

The proposed model is divided into two phases as generating the degree of goodness for each sentence and then generating the star rating on them. In the proposed STAR model, Phase 1 also uses GloVe: Global Vectors for Word Representation for word embeddings for a better understanding and prediction of reviews. Here, we use a combination of two different phases to classify reviews. The first phase uses a combination of two Long Short Term memory (LSTM) layers to classify each sentence in the review on a continuous scale of 0 to 1. Second phase uses a simpler model made of single LSTM layer with a fully connected output layer on top of it. The second phase uses the out produced by the first phase to produce the final rating or the review. Both the phase models has been trained on various datasets available online on kaggle[26] and other self-extracted product reviews. The major components of the proposed STAR model is shown in Figure 1.

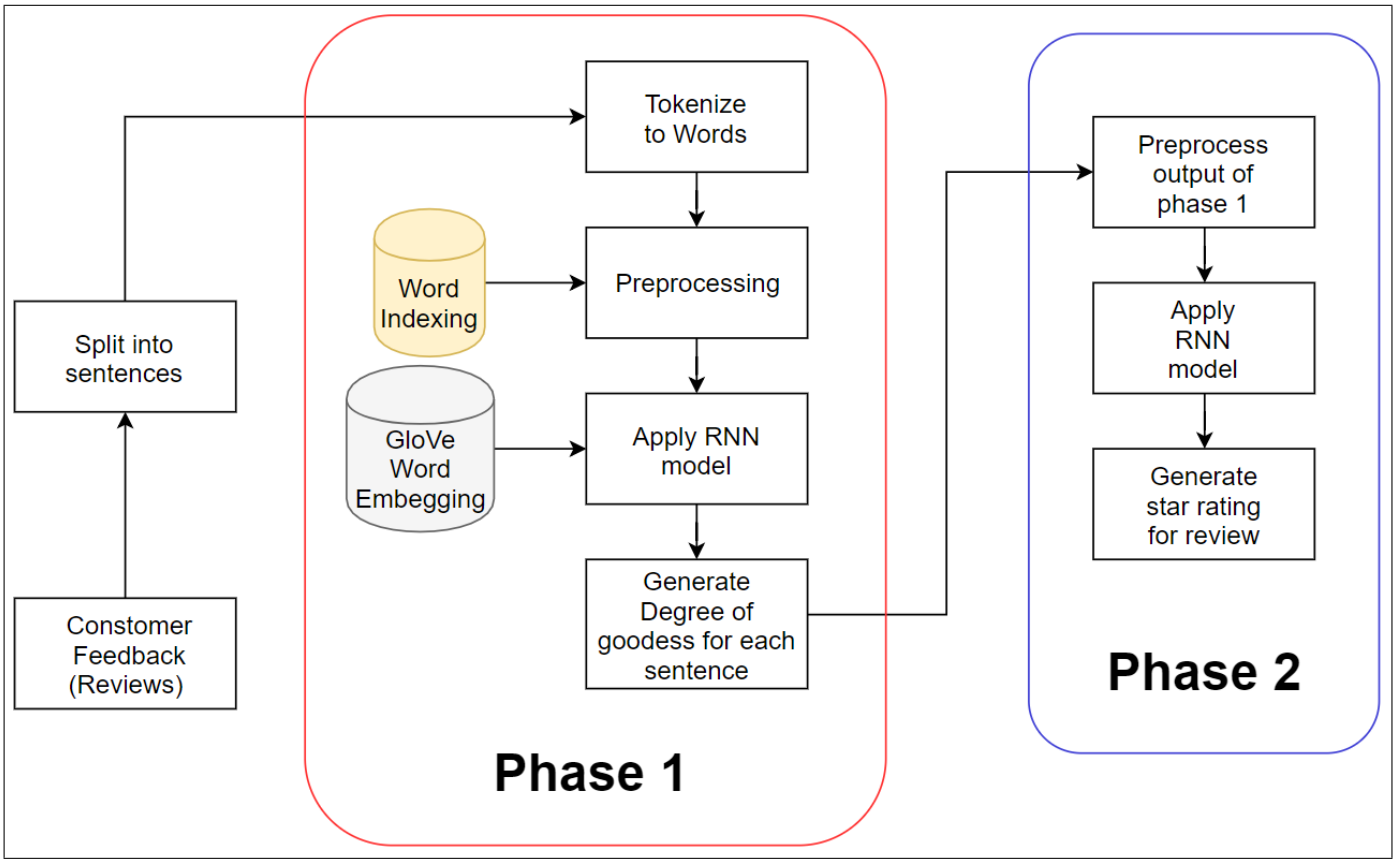

Fig. 1: The major elements of the proposed STAR model

Consider a situation where a person is giving review for a movie. He liked the movie and decides to give either a three of four star rating. He has to fill a star rating and write a review as well. Assume this hypothetical person starts on a good note. As he proceeds, he remembers some parts in the film that he did not like and end his review by criticizing the movie. Now as he was in the process of criticizing the movie, he is more likely to give 3 stars to the movie. Consider another movie, our hypothetical person disliked the movie and starts on a bad note but, similar to previous case, ends on a good note. This person may end up giving 3 star rating to this movie also. Hence the change in emotional state of the person while writing a review can change the star rating. This is one the many patterns that our model uses to predict the star rating of the reviews. The working of this model can also be understood as implemented in Algorithm 1.

Data: Review of the customers
Split review into list of sentences sen;
while *Sentences* **do**
  tokenize each sentence into list of words;
  if *word in Sentence is in wordIndex* **then**
    Store the index value of the word ;
  end
  Feed the produced list to model 1;
  Save the predicted rating of each sentence in a review to a new list;
end
feed the list of ratings to model 2;
Calculate overall Score of the Review;
Result: Rated Review on the scale of 5

Algorithm 1: Model Pseudo-code

A. *Phase 1 : For sentence-wise sentiment classification*

Instead of trying to predict the rating of entire review in a single go, we first split the review into sentences and then predict the rating of these sentences on a continuous scale of 0 to 1. We have used a continuous scale from 0 to 1 instead of a discrete scale of 0 or 1 in order to get a proper idea how good or bad a product or service is expressed to be in a particular sentence. Doing this has a huge impact on the accuracy of phase 2 which works on the sentence wise ratings produced by phase 1. Brief summary of phase 1 is shown in Figure 2.

B. *Phase 2 : For final star-rating prediction*

We can argue that all the sentences in a review have different degree of likeness and unlikeness about the product or service. Consider a movie review, there can be a sentence that appreciates the actors for their work and another sentence that shows dislike for the story line. We need to consider the

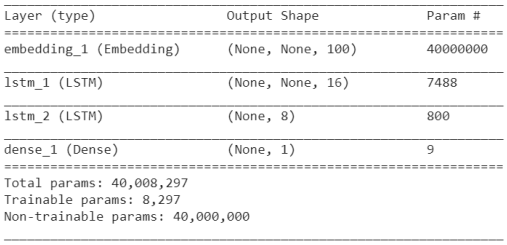

Fig. 2: Structure of phase 1 model.

connection between the sentiments in different sentences of a review in order to provide a better star rating to the review. We have used phase 1 to provide these ratings for each sentence, i.e., the degree of likeness in each sentence. A brief summary of phase 2 is shown in Figure 3.

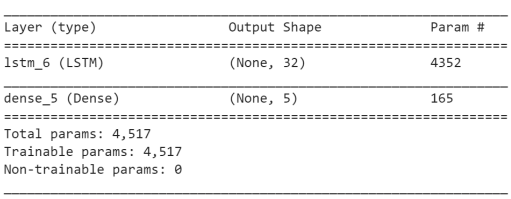

Fig. 3: Structure of phase 2 model.

The way we use these sentence-wise rating of a review can have a high impact on the star rating produced. We have used a recurrent neural network to predict the final rating in phase 2. It was observed that the order in which a consumer shows a like or dislike for a product or service in his review has a great affect on the rating given to product. For example, consider two reviews for the same movie below:

1) The movie is over all slow and dull. It completely lacks in sense and logic. But the movie has a good touch of humour. It picks up some pace only in the end. The ending of the movie was unexpectedly good compared to the rest of the movie.
2) The actors are good. The suspense in the last part of the movie was also good. But that is all the movie has to offer. It lacks a proper story line. There are countless parts in the movie that could have been much better. The senseless action scenes are over hyped.

Review 1 starts with a negative sense but ends in a positive sense where is Review 2 is just the opposite. Review 1 was rated 3 stars whereas Review 2 was rated only 2 stars. The two reviews are not much different in terms of lexical basis but differ a lot in the order of sentences in which the sentiment are expressed. Our phase 2 uses this concept of difference in order to predict the star rating for a review. The training of the both the phases are well discussed in Section IV. The five types of the training datasets are used to train the phases of the proposed model.

IV. Training of Phases

Phase 1 model was trained on a large dataset containing a mixture of positive and negative reviews of various products and services including books, movies, electronic appliances, music, video games, restaurants, and different types of shops. The dataset contained one line reviews which were labelled as either positive or negative. Neutral reviews were not used as predicting either one if the two classes positive and negative facilitates generating a single value between 0 and 1 as output. We can use the probability of output being 1 as the degree of goodness of the review.

Phase 2 model was trained using 2 different datasets using the concept of transfer learning. First it was trained input data containing discrete rating of individual sentences as 0 or 1, i.e., rated as good or bad and output as the final expected rating on a scale of 5. This allowed the model to learn different patterns related to change in degree of goodness in the review. Then the model was further trained on probability output produces by phase 1 which can be any value between 0 and 1.

V. Results and experiments

The qualitative analysis and quantitative analysis is carried out to determine the performance of the proposed model, and compared with the existing models. The following subsection holds the results obtained by applying the proposed model on various datasets and presents the accuracy obtained by our classifier. We have tested the performance of the proposed STAR model on a Computer system in Window 7 environment, DDRAM 4GB and 2.27GHz speed processor with anaconda (python 3).

A. Qualitative Analysis

TABLE I shows the accuracy of the model, applied to different product datasets [16] and Kaggle movie review dataset [17]. We have used four amazon product datasets, Kaggle movie reviews dataset and a kaggle dataset containing review of various products and services, and calculated the accuracy as shown.

TABLE I: Accuracies obtained for different datasets

Dataset	**Accuracy(%)**
Amazon Electronics	59.2
Amazon Apps for android	61.3
Amazon Digital music products	59.8
Amazon Books	60.8
Kaggle Movie reviews	57.1
Kaggle Combined reviews	64.6

B. Quantitative Analysis

TABLE II shows the comparison of our proposed model with other existing models. The heighest accuracy achieved among various model in every dataset used in comparision is shown in bold.

TABLE II: Detailed comparison with existing work

Model	Dataset	Accuracy(%)
WSD	Electronics products	43.2
	Apps	39.7
	Digital products	51.7
	Kaggle movie reviews	47.3
ANEW	Electronics products	32.1
	Apps	29.9
	Digital products	38.1
	Kaggle movie reviews	33.4
LIWC	Electronics products	30.6
	Apps	29.7
	Digital products	32.5
	Kaggle movie reviews	27.3
SCN	Electronics products	39.2
	Apps	41.9
	Digital products	47.3
	Kaggle movie reviews	38.2
VADER	Electronics products	57.3
	Apps	54.9
	Digital products	63.2
	Kaggle movie reviews	53.1
LSRC-NB	Electronics products	57.1
	Apps	56.8
	Digital products	**63.3**
	Kaggle movie reviews	54.1
LSRC-NN	Electronics products	58.9
	Apps	56.9
	Digital products	62.9
	Kaggle movie reviews	54.9
Proposed STAR Model	Electronics products	**59.2**
	Apps	**61.3**
	Digital products	59.8
	Kaggle movie reviews	**57.1**

The following observations are made from literature and table 2.

- Proposed model is able to achieve the heighest accuracy in analysing Electronic product reviews, app review and Kaggle movie review dataset. In Digital product reviews, the model gets a decent accuracy which is higher than most of the previous models. It was observerd that due to the presence of technical terms in the reviews, the phase 1 model had a comparatively lower accuracy which lead to overall lower accuracy of the model.
- Models like WSD, ANEW, LIWC, and SCN exploit different aspects of reviews but fail to achieve high accuracy for most of the datasets.
- LSRC-NN and LSRC-NB models, which use lexical based approach to classify model, work very well and produce an almost consistent accuracy across all the datasets.

Various imporvements can be made in the model to get a higher accuracy. The madel consists of a combination of two phases. Each phase uses its own machine learning model and hence we can work on both the models independently to improve the overall accuracy of the model.

Firstly, the accuracy of phase 1 model is near 75 to 80 precent for most of the datasets. Since phase 2 model uses the output produces by phase 1 model, if we can increase the accuracy of phase 1 model, it will improve the data available to phase two model and hence the accuracy of the model can be increased.

Secondly, phase 2 model works on ups and downs in degree of likeness in reviews to predict the final star rating. It was trained on two different datasets. First dataset has rating of one line reviews as good(rated 1) or bad(rated 0) and the second dataset was the output produced by the phase 1 model. Since ohase 1 model is not completely accurate, phase 2 model was not trained on 100 percent accurate data. If we can increase the accuracy of phase 1 model, it will directly affect the training of phase 2 model and hence increase its accuracy.

VI. Conclusion

In this work, we have proposed an approach based on combination of two phases. Phase 1 works on the review sentence by sentence and generates a value between 0 and 1 for each sentence to show the degree of likeness for product or service in the sentence. Phase 2 works on the output produced by phase 1 and based on ups and downs in the degree of likeness of sentence, it produces a final output star rating on a scale of 5. The experimental results shows that the proposed STAR model outperforms many of the existing models. In near future, the model can find their applications in the field of product review mining, sales and marketing, product reviews and movie reviews. The authors will work on the long dataset and will try to resolve the polarity issues in much ways in the near future work. The dataset associated to the reviews of the tweets will be used as online analysis for keeping in view of the real time applications.

References

[1] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis (Foundations and Trends (R) in Information Retrieval).

[2] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[3] Jindal, N., & Liu, B.(2006, July). Mining comparative sentences and relations. In AAAI (Vol. 22, pp. 1331-1336).

[4] Ganapathibholta, M., & Liu, B.(2008, August). Mining Opinions in comparative sentences. In proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 (pp.241-248). Association for Computational Linguistics.

[5] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

[6] Esuli, A., & Sebastiani, F. (2007). SentiWordNet: a high-coverage lexical resource for opinion mining. Evaluation, 1-26.

[7] www.liwc.net, 2017

[8] Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2. In Proc. AAAI IFAI RSC-12.
[9] Akkaya, C., Wiebe, J., & Mihalcea, R. (2009). Subjectivity word sense disambiguation. In Proc. EMNLP-09.
[10] Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings.
[11] Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.
[12] Jurek, A., Mulvenna, M.D., & Bi, Y.(2015). Improved lexicon-based sentiment analysis for social media analytics. Security Informatics 4. (1), 9.
[13] Murphy, K. P. (2006). Naive Bayes classifiers. University of British Columbia.
[14] Dos Santos, C. N., & Gatti, M. (2014, August). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In COLING (pp. 69-78).
[15] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).
[16] http://jmcauley.ucsd.edu/data/amazon/, 2017.
[17] https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data, 2017.
[18] Kumar, K., & Kurhekar, M. (2016, October). Economically Efficient Virtualization over Cloud Using Docker Containers. In Cloud Computing in Emerging Markets (CCEM), 2016 IEEE International Conference on (pp. 95-100). IEEE.
[19] Tan, S. S., & Na, J. C. (2017, September). Mining Semantic Patterns for Sentiment Analysis of Product Reviews. In International Conference on Theory and Practice of Digital Libraries (pp. 382-393). Springer, Cham.
[20] Harman Singh, et al., (2017). HDML: Habit Detection with Machine Learning, The 7th ACM International Conference on Computer and Communication Technology (ICCCT'17).
[21] Shikhar Sharma, et al., Computationally efficient ANN model for Small Scale Problems, Springer International Conference On Machine Intelligence And Signal Processing (MISP'17), Indore(India), 2017.
[22] Ray, P., & Chakrabarti, A. (2017, February). Twitter sentiment analysis for product review using lexicon method. In Data Management, Analytics and Innovation (ICDMAI), 2017 International Conference on (pp. 211-216). IEEE.
[23] Keshavarz, H., Abadeh, M. S., & Almasi, M. (2017, September). A new lexicon learning algorithm for sentiment analysis of big data. In Intelligent Systems and Informatics (SISY), 2017 IEEE 15th International Symposium on (pp. 000249-000254). IEEE.
[24] Sharma, S., Kumar, P., & Kumar, K. (2017, December). LEXER: LEXicon Based Emotion AnalyzeR. In International Conference on Pattern Recognition and Machine Intelligence (pp. 373-379). Springer, Cham.
[25] Krishan Kumar & Manish Kurhekar, (2017). Sentimentalizer: Docker container utility over Cloud, The 9th IEEE International Conference on Advances in Pattern Recognition (ICAPR'17).
[26] Oscar Tckstrm and Ryan McDonald (2011). Discovering fine-grained sentiment with latent variable structured prediction models. European Conference on Information Retrieval (ECIR 2011), Dublin, UK.
[27] Shubham Kumar, Krishan Kumar. LSRC: Lexicon Star Rating system over Cloud.