An optimized CNN based robust sentiment analysis system on big social data using text polarity feature

Komalpreet Kaur, Chitender Kaur, Tarandeep Kaur Bhatia

Abstract: People express their personal views and emotions by reacting to the event, product and individuals. These reactions are very vulnerable, therefore comprehending such data and then processing it effectively, have been the content of research in various areas such as business and politics. Sentiment analysis in social media, in which the task of extracting subjective information from the e-commerce site, social sites have drawn great recognition from the Web mining community. Social media gives an invaluable insight not only into human ideas but also in the challenges incurred due to huge amounts of big data. These issues comprises of the processing of massive amounts of streaming data, as well as automatically identifying human expressions within short text messages. To solve this problem, an intelligent sentiment analysis approach is presented, which is used to extract the opinions of people from social media. Initially, a lexicon dictionary is created that comprises of positive, negative and neutral words. Then, pre-processing (normalization, punctuation removal, stop word removal and tokenization) is applied to the test data. Finally, Feature extraction, feature optimization and classification algorithms are applied to the pre-processed data. PSO is used as a feature optimization algorithm with Convolutional neural network (CNN) as a classification algorithm. CNN is trained as per the optimized features. During testing phase, sentiments such as positive, negative is identified by matching the uploaded text with the saved data. The presented model helps to identify people feeling in terms of business, comments, and reviews written on the social sites. The detection accuracy up to 98.32 % is obtained.

Keywords: Big data, convolutional neural network, particle swarm optimization, sentiment analysis, social media.

I. INTRODUCTION

Social media and related applications provides millions of users with the opportunity of expressing and sharing their views on a particular topic, moreover, the participants can share their relationship towards the content by hitting the like or dislike icon [1]. These continuing series of actions on social media high volume, high speed, high volatility data known as big social data. Generally, this kind of information leads to a broader exchange of views that are helpful in identifying the trends in the digital world [2]. A number of researchers have come up with the keen interest in the development of big data for describing, identifying and predicting human behavior in various areas [3].

This type of development covers various research tools, especially text analysis. In fact about 80% of Internet content is text, thus making text analysis, a fundamental element in public sentiments. Sentiment analysis is sometimes termed as opinion mining, which is used to identify the feeling of people on the basis of posts and actions displayed on social media. The post is categorized on the basis of polarity, that is, positive, negative or neutral [4].

Sentiment analysis can be classified as: Lexicon analysis and Machine learning.

Lexicon analysis determines the polarity of the text using the semantic sense of the words and phrases contained in the document. This research work is carried out using the lexicon dictionary [5].

Machine Learning (ML) covers building models from a tagged training database (texts and sentences) to find out the document orientation. The machine learning algorithms like CNN and PSO have been used in this research [6].

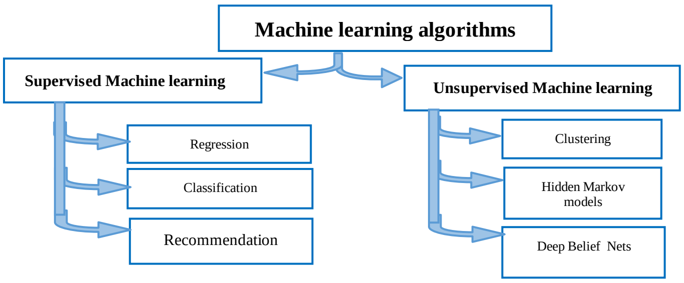

Fig.1 Machine learning algorithms

Fig. 1 demonstrates the different machine learning algorithms that can be used by various researchers to train the system. The categorization of the algorithms is mainly done in two types named as supervised and unsupervised [7]. In supervised learning process, the system learns as per the set of provided data features, whereas, in un-supervised learning, the system discovers the hidden structure without the dataset. In this research, a supervised learning approach named as CNN is utilized as a classification algorithm [8].

Revised Manuscript Received on April 07, 2019.

Komalpreet Kaur, M.Tech Research scholar, Computer Science and Engineering, Chandigarh Engineering College, Landran, Punjab, India.

Chitender Kaur, Assistant Professor, Computer Science and Engineering, Chandigarh Engineering College, Landran, Punjab, India.

Tarandeep Kaur Bhatia, Assistant Professor, Computer Science and Engineering, Chitkara University, Rajpura, Punjab, India.

II. RELATED WORK

This section demonstrates the existing work done by several authors in the area of sentiment analysis for big social data.

The work presented by various authors along with techniques and dataset used and their outcomes is provided in Table. 1.

Table. 1 State-of art

Authors and year	Proposed work	Techniques used/ Dataset	Outcomes
Bravo-Marquez et al. [9]	Presented a new method for classifying the sentiments on the basis of meta-level features.	Naive Bayes, SVM, decision trees are used as a supervised learning approach.	The maximum accuracy up to 83.3 % has been achieved.
El Alaoui et al. [10]	Demonstrated a sentiment analysis technique that examined the posts on social media and inferred people opinions in real scenario.	Tweets related to the 2016US election have been used for training. Naïve Bayes technique has been used as a classification algorithm	The average accuracy up to 90.21 % has been achieved.
Glorot, X. et al. [11]	Proposed a deep learning scheme to resolve the problem of domain adaptation of ‘sentiment analysis.	The dataset has been collected from Amazon site comprising of 40k reviews with 22 product types in which reviews have been labelled as positive or negative. SVM has been used as a classification technique.	Stacked Denoising Autoencoder has performed better than SVM.
Pak, A et al. [12]	Presented a linguistic analysis of the Twitter data for sentiment analysis.	The data set has been made by gathering 3 lakh of text posts from the Twitter site. Two types of emotions such as happy and sad have been collected. N-gram method has been used as feature extraction and SVM has been used as a classification algorithm.	The detection accuracy using bi-gram is found to be better than the other uni-gram and trigram approaches which is approximately equal to 68.18%.
Alarifi, A et al. [13]	Designed a sentiment analysis system for Amazon Web site that comprises of product information using data acquisition and preprocessing method (normalization).	Cat swarm optimization with neural network has been used to optimize and classify the extracted features. Optimization algorithm has minimized error level of available features.	The computed parameters such as precision, recall, accuracy and error rate have been measured. The maximum accuracy up to 96.89% has been obtained.
Chiong, et al. [14]	Proposed a sentiment analysis approach based on news disclosures for predicting the behavior of financial markets.	PSO and SVM have been used as a deep learning approach for enhancing the accuracy of the proposed work. 13,135 regulated German ad hoc announcements in English has been considered as a dataset for the experiment. The work has been performed in Python	57.8% accuracy has been achieved using SVM and PSO technique.

III. PROPOSED METHOD

The research has been conducted mainly in four phases: (i) Pre-processing (ii) Feature extraction (iii) Optimization (iv) Classification.

Initially, a dictionary has been created known as **Lexicon dictionary** as shown in Fig. 2 The Lexicon-based approach is based on sentences or words and a collection of known emotions. In short, the dictionary is divided into, dictionary based scheme and corpus-based scheme. The dictionary-based scheme, discovers opinions in the document, and then finds their origin of semantic in the dictionary. The available Lexicon dictionaries are SentiWordNet, WordStat Sentiment Dictionary, SenticNet and many more [15]. However, lexicon dictionary can be created manually. In this research, the created dictionary is represented as below.

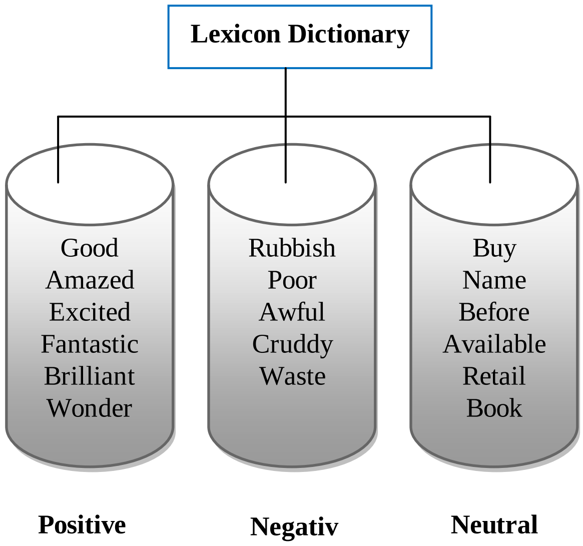

Fig. 2 Lexicon Dictionary

A. *Preprocessing*

In this step, the high dimension text features are converted into the low dimension text features that contain accurate data. In this research, the data goes through different phases such as:

i. Normalization

The words that are written in a different way but have same meaning are required to be processed in a better way. Normalization processes make sure that these words are treated equally [16]. During this process, the text is converted into lower case letters. For an example, suppose the text is:

"My name is Komalpreet"

When normalization process is applied to this sentence, all the letters are converted into lowercase and the text will be changed to:

"my name is komalpreet".

ii. Punctuation Removal

During this process, the punctuation from the text is removed and the text will be obtained like:

my name is komalpreet

iii. Stopword removal

During this course of action, the stop words present in the sentence are removed and the data will be appeared as shown:

name komalpreet

In our example, the stop words, "my" and "is" are removed.

iv. Tokenization algorithm

Tokenization method is used to extract the token values of the available text. In this example, the token value of *name komalpreet is (417, 1076).*

B. *Feature Extraction*

In this step, the features of the uploaded text such as (positive, negative or neutral) are determined by using the lexicon dictionary. The token values of the text, after applying the tokenization algorithm, are compared with the values stored in the Lexicon Dictionary. In the above stated example, the values obtained, that is, 417 and 1076, shown in Fig. 3 are compared with the values stored in the lexicon dictionary. By comparing these values, the decision is taken [17].

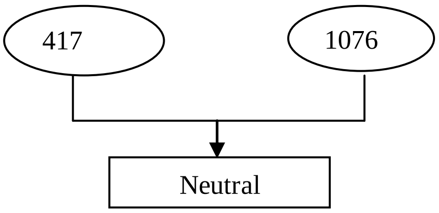

Fig. 3 Feature extraction

C. *Feature Optimization*

In this work, PSO is used as a feature optimization algorithm. It is a population-based scheme that is used to resolve the continuous as well as discrete optimization issues. The particle is used as a software agent that travels in the search area by varying their velocity until proper solution to the existing problem is achieved. In PSO, a set of random particles is initialized with a solution and then particles search for best results by modifying generations. In every generation, the agent is updated by using two optimal values. The initial value is updated as per the fitness function and this value is stored known as pbest. The other best value that is obtained by searching an optimal position of the agent is known as global best and represented by gbest.

As these two values are obtained by the agent, then the particles update their velocity as well as position by using the following equation.

$$v_p = v_p + k_1 \times rand() \times (pbest - present) + k_2 \times rand() \times (gbest - present)$$

$$Present = present + v_p$$

In the above equation, v_p denotes the particle velocity, present () represents the present particle solution, rand () signify the random number lies between (0, 1), k_1 and k_2 are the learning factors [18]. The algorithm is defined as below.

Algorithm: PSO Algorithm
Input: Feature of Text Sentiment File as Token Value
Output: Optimized token value
Initialize the PSO with their operating functions –Swarm Size
Define fitness function of PSO for feature selection
Fit_function_PSO = True; if $f_s > f_t$
False; otherwise
// to check fitness of each data in feature list if data fulfill the condition, then those are consider in feature list else replace with objective value
Calculate size of Feature in terms of row and columns (R, C)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

```
For i=1 to R
    For j=1 to C
        Fs=Feature (i, j) // to select one by one feature
        Ft=Threshold (i, j) // mean of text token values
        Fit_function_PSO =Call Fit_function_PSO (Fs, Ft)
        No. of Variable=1
        Fitdata=PSO (Fit_function_PSO, No. of Variable, PSO functions) // apply PSO on feature data with fitness function of system
    End
End
Returns: Fit data as an optimized token value of text file // return optimized feature set which helps to train the CNN
End
```

D. Convolutional Neural Network (CNN)

CNN is a machine learning algorithm, like neural networks, consisting of neurons that are used to train the system on the basis of weights biases. Each neuron accepts a few entries, applies weighted and bias function, passes those entries through the activation function, and responds with an output [19]. In CNN, the input, as well as the output data is three dimensional. The optimized data is passed through the input layer which again forwards it to the convolutional layer. In the convolutional layer, the dot product of weight and input is added. In the pool layer, the volume of the data is reduced to increase the processing speed of CNN. On the other side, the size of memory is decreased to protect the system from over lifting. After this, the data is passed to the fully connected layer, in which the class score is determined [20]. The CNN algorithm used in the present work is described as follows:

Algorithm: CNN Algorithm

```
Input: Optimized token value as a Training Data (T), Target (G) and Neurons (N) // all are the inputs of CNN
Output: Text Polarity as Positive, Negative or Neutral // CNN return text polarity in terms of sentiments types
Initialize CNN with parameters    – Epochs (E)
                                  – Neurons (N)
                                  – Performance parameters: Cross Entropy, Gradient, Mutation and Validation // these are the CNN basic parameters
                                  – Training Techniques: Scaled Conjugate Gradient (Trainscg) // CNN training algorithm
                                  – Data Division: Random // selection of data based on the random
For each set of T // Loop for all data in training array
If Training Data ε Positive // if data belongs to 1st category
    $Group(1) = Categories\ of\ Trainingdata$
Else if Training Data ε Negative // if data belongs to 2nd category
    $Group(2) = Categories\ of\ Trainingdata$
Else if Training Data ε Neutral // if data belongs to 3rd category
    $Group(3) = Categories\ of\ Trainingdata$
End
Initialized the CNN using Training data and Group
Net = patternnet ($N$) // for initialization of CNN
Set the training parameters according to the requirements and train the system
Net = Train ($Net, Training\ data\ , Group$) // to train the system
Classification Results = simulate (Net, Optimized Current Text Token Value) // correlate test data with training database
If Classification Results = True // if matched with database
Show classified results in terms of the their polarity
Calculate the performance parameters
End
Return: Classified Results
End
```

The architecture of CNN is as shown below in Fig. 4

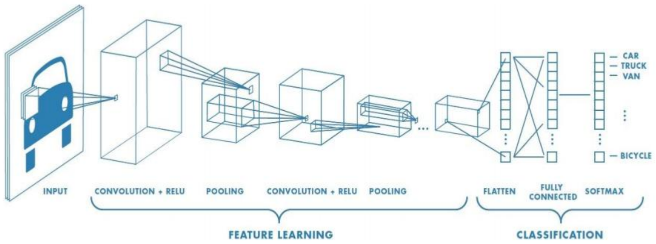

Fig. 4 Structure of CNN

IV. EXPERIMENT RESULTS

The experiment is conducted in MATLAB tool and the efficiency of the proposed model is determined by considering four types of database that are created manually. The detail description of these databases is given below.

A. Preprocessing steps

This process is applied in the pre-processing stage to remove or filter out the words such as ‘the’, ‘an’, ‘a’, ‘is’, ‘there’ etc. The system is trained in such a manner so that when these words exist in the sentence, then these words are ignored. The stop words list is shown in Table. 2. Also, the positive, negative and neutral words used in this research work are shown in Table. 3, 4, and 5 respectively.

Table.2 Stop words list list

1	a
2	about
3	above
4	across
5	after
6	again
7	all
8	almost
9	along

Table.3 Positive words list

1	good
2	amazed
3	fantastic
4	really
5	brilliant
6	wow
7	better
8	amazing
9	wonderful

Table.4 Negative words list

1	rubbish
2	disappointed
3	poor
4	waste of time
5	shocking
6	bad
7	expense
8	awful
9	cruddy

Table. 5 Neutral word

1	buy
2	book
3	retail
4	product
5	horror
6	online purchasing
7	discount
8	now
9	available

B. Result and analysis

The work is carried out in two phases: training and testing. In the training phase, the data goes through different steps such as uploading test data, pre-processing, feature extraction, optimization and storing data through CNN algorithm. During testing phase, same processes are repeated and at last classification is performed by using classification algorithm. CNN model is used to check the polarity of the sentence and also the performance parameters are evaluated to determine the efficiency of the design model. The process is depicted in Table. 6 and evaluation parameters calculated are depicted in Table. 7.

Table. 6 Computational Result

Sr. no.	Upload data	Pre-processing	Token value	Classification results	Evaluation Accuracy (%)
1	Got this for my son and aparantly it was so good he won't use anything else now	got son apparently so good wont use	226 330 333 336 425 456 974	Positive	97.96
2	Bought this for my son and he tells me it's the best thing ever	bought son tells best things	336 430 548 649 653	Positive	99.41
3	Don't buy this product its rubbish!	dont buy product rubbish	336 437 751 769	Negative	98.93
4	I think this is a shoddy design	think shoddy design	542 634 651	Negative	99.82
5	Buy before the stocks run out	buy stocks run	336 341 663	Neutral	97.29
6	Get 20% of what you buy 2 of these	20 buy 2	50 98 336	Neutral	97.25

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table. 7 Evaluation parameters

Number of test data	Error (%)	Execution Time	Precision	Recall	F-measure	Accuracy
1	2.7001	0.08938	0.9529	0.15	0.181	97.29
2	1.3599	0.074023	0.9529	0.14	0.181	98.64
3	0.8725	0.10676	0.9429	0.17	0.290	99.12
4	1.3599	0.8384	0.9529	0.13	0.182	95.64
5	1.2826	1.1454	0.9429	017	0.290	98.71
6	0.8400	0.1462	0.9429	0.17	0.290	99.15
7	0.298	0.29243	0.8885	0.34	0.496	99.70

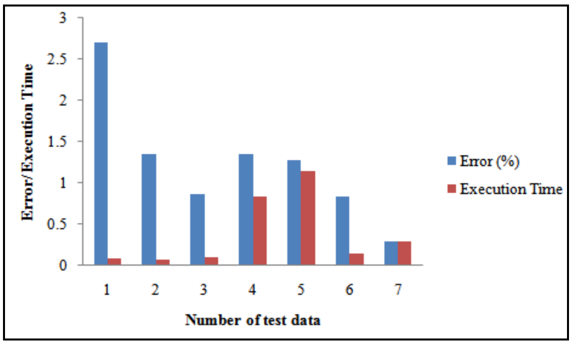

Fig. 5 Error and execution time

Fig. 5 represents the graph plotted between computed parameters (Error & execution time) with respect to number of test data uploaded (positive, negative or neutral). The average value of Error and execution time measured for the proposed work is 1.24 and 0.38 respectively.

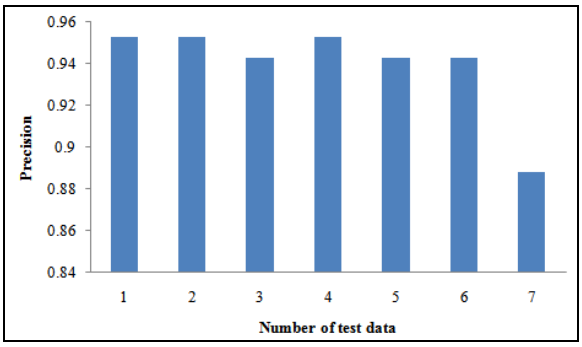

Fig. 6 Precision

Fig. 6 represents Precision values, measured for seven test data are plotted in the fig 6. The average value of precision calculated after classifying the uploaded test data is 0.937.

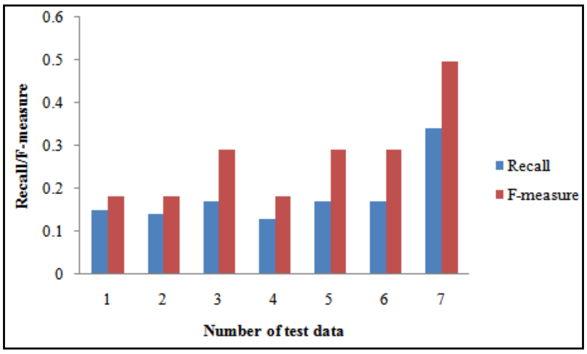

Fig. 7 Recall and F-measure

The average value of Recall and F-measure computed after analyzing the sentiments of the uploaded data are 0.181 and 0.272 respectively are shown in Fig. 7.

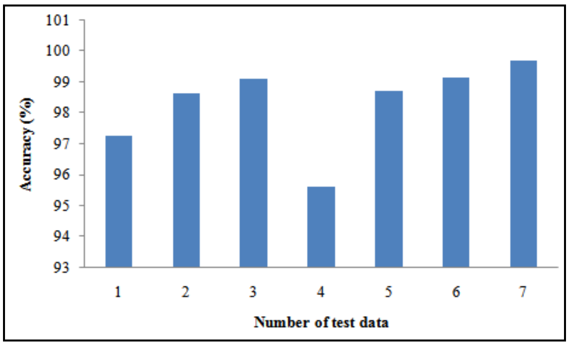

Fig. 8 Accuracy

The accuracy of the designed sentiment analysis system is shown in Fig. 8. The average accuracy of the proposed system is up to 98.32%.

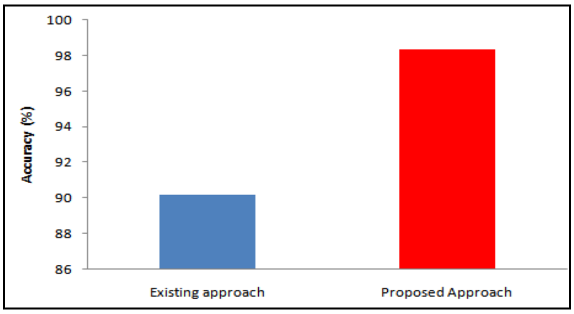

Fig. 9 Comparison of accuracy

The mean accuracy of the proposed work and existing work are obtained up to 98.32 % and 90.21% respectively and the graphical representation is shown in Fig. 9. From the above figure it has been observed that the accuracy of the proposed work has been increased by 8.99% from the existing work.

V. CONCLUSION

Sentiment analysis has proven to be effective while analyzing people's attitudes by examining large social data. In this, a novel approach is designed to extract people's opinions on specific topics by relying on social media content. 70% of dataset is used for training whereas remaining dataset is used during testing. It has been determined that the proposed classifier, classifies the positive, negative and neutral sentiments with an accuracy of 98.32% From the experiment, it has been analyzed that the detection accuracy of sentiments has been increased by 8.99 % from the existing work. The main advantage of this work is that a stop word panel is added in to the GUI, so that a user can add or remove the stop words as per the need.

VI. FUTURE SCOPE

The domain of sentiment analysis has become an electrifying research field due to involvement of a wide range of real-world applications where uncovering people’s sentiments helps in enhanced decision making. The field of sentiment analysis is not just influencing socially and economically in a positive way, but is also having an impact on everyone’s way of thinking for one’s benefit. In future, the sentiment analysis’ field is going to be unshakeable as the detection accuracy can be improvised by hybridizing the classification approach with more relevant feature optimization techniques like grasshopper algorithm.

REFERENCES

1. Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 42(24), 9603-9611.
2. Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., & By, T. (2012, August). Sentiment analysis on social media. In Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on (pp. 919-926). IEEE.
3. Balahur, A. (2013). Sentiment analysis in social media texts. In Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 120-128).
4. Anandarajan, M., Hill, C., & Nolan, T. (2019). Sentiment Analysis of Movie Reviews Using R. In Practical Text Analytics(pp. 193-220). Springer, Cham.
5. Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In Cognitive Informatics and Soft Computing (pp. 639-647). Springer, Singapore.
6. Jiang, D., Tao, Q., Wang, Z., & Dong, L. (2019). An Intelligent Logistic Regression Approach for Verb Expression’s Sentiment Analysis. In Recent Developments in Intelligent Computing, Communication and Devices (pp. 173-181). Springer, Singapore.
7. Stine, R. A. (2019). Sentiment Analysis. Annual Review of Statistics and Its Application.
8. Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. Expert Systems with Applications, 117, 139-147.
9. Bravo-Marquez, F., Mendoza, M., & Poblete, B. (2014). Meta-level sentiment models for big social data analysis. Knowledge-Based Systems, 69, 86-99
10. El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A., & Kobi, A. (2018). A novel adaptable approach for sentiment analysis on big social data. Journal of Big Data, 5(1), 12.
11. Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In Proceedings of the 28th international conference on machine learning (ICML-11) (pp. 513-520).
12. Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).
13. Alarifi, A., Tolba, A., Al-Makhadmeh, Z., & Said, W. (2018). A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks. The Journal of Supercomputing, 1-16.
14. Chiong, R., Fan, Z., Hu, Z., Adam, M. T., Lutz, B., & Neumann, D. (2018, July). A sentiment analysis-based machine learning approach for financial market prediction via news disclosures. In Proceedings of the Genetic and Evolutionary Computation Conference Companion (pp. 278-279). ACM.
15. Araque, O., Zhu, G., & Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. Knowledge-Based Systems, 165, 346-359.
16. Yang, X., Xu, S., Wu, H., & Bie, R. (2019). Sentiment Analysis of Weibo Comment Texts Based on Extended Vocabulary and Convolutional Neural Network. Procedia Computer Science, 147, 361-368.
17. Lin, Q., Zhu, Y., Zhang, S., Shi, P., Guo, Q., & Niu, Z. (2019). Lexical based automated teaching evaluation via students’ short reviews. Computer Applications in Engineering Education, 27(1), 194-205.
18. Jiang, D., Tao, Q., Wang, Z., & Dong, L. (2019). An Intelligent Logistic Regression Approach for Verb Expression’s Sentiment Analysis. In Recent Developments in Intelligent Computing, Communication and Devices (pp. 173-181). Springer, Singapore.
19. Yang, X., Xu, S., Wu, H., & Bie, R. (2019). Sentiment Analysis of Weibo Comment Texts Based on Extended Vocabulary and Convolutional Neural Network. Procedia Computer Science, 147, 361-368.
20. Zhang, Y., Zhang, Z., Miao, D., & Wang, J. (2019). Three-way enhanced convolutional neural networks for sentence-level sentiment classification. Information Sciences, 477, 55-64.

AUTHORS PROFILE

Ms. Komalpreet Kaur is an M.Tech research scholar at Department of Computer Science and Engineering, Chandigarh Engineering College, Landran, Punjab. Her areas of expertise are Big data, Machine learning and Neural networks. She has worked on the project “Validating sentiment analysis system using PSO and CNN”.

Ms Chitender Kaur is an Assistant Professor at Department of Computer Science and Engineering, Chandigarh Engineering College, Landran, Punjab. Her areas of expertise are Networking and Big data. She has four publications in International journals.

Ms. Tarandeep Kaur Bhatia is a PhD research scholar and an Assistant Professor, at Department of Computer Science and Engineering, Chitkara University, Punjab. Her areas of expertise are Software engineering, Big data, Machine learning, Networking and VANETs. She is currently working with simulation tools in the field of VANETs.