ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse81922020.pdf
https://doi.org/10.30534/ijatcse/2020/81922020

Genetic Optimization in Hybrid Level Sentiment Analysis for Opinion Classification

Gatta Sravya[1], **Marriboyina Sreedevi**[2]
[1]M.Tech Student, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Guntur District, Andhra Pradesh-522502, India. sravya.gatta@gmail.com.
[2]Professor, Koneru Lakshmaiah Education Foundation (KLEF), Vaddeswaram, Guntur District, Andhra Pradesh-522502, India .msreedevi_27@kluniversity.in.

ABSTRACT

Now-a-days the major task in Natural Language Processing (NLP) is Sentiment Analysis (SA).Identifying the polarity of the textual data has been a complex problem which is applied in the vast range of areas from product feedback analysis to user statement understanding. Yet many approaches and models were developed for this problem, classification problem still remained. To overcome this problem and to train the machine to perfection, we introduced a feature optimization algorithm with CNN (Convolutional Neural Network) as a classification algorithm. The main aim of our project is to attain more accuracy than the proposed capsule-based model. Our experimented results showed huge outcomes for this model.

Key Words Crossover, Mutation, Polarity, Sentiment classification.

1. INTRODUCTION

With the growth of technology era, many social media platforms, third parties, e-commerce websites etc., were evolved thus leading in abundance of data. Sentiment Analysis (SA) in Natural Language Processing (NLP) helps identifying the polarity of textual data. In Business Analytics, this is crucial in enhancing their business to profits by understanding the user or customer opinion on different products and predicting the trends in the market. Likewise, it is also applicable to various fields in which movie reviews is one. Thus, we experimented on the IMDb movie reviews dataset and obtained greater results than the studied experiments. This is obtained by using feature optimization algorithm that results the best elements in the total population of the input variables by using the objective function. Usually, in genetic optimization best chromosomes from the total population are selected and by the genetic operators, they are crossed over and mutated to produce new population. The process repeats by initializing with the new population until best outcomes (children) are produced. This algorithm was introduced at the classification problem in our project so as to obtain the best results and train the machine in generating the best out of any data.
The main intension of this project is to attain greater accuracy than the previous stated methods.

2. RELATED WORK

[1] A capsule based hybrid neural network has been introduced to obtain effective semantic information, where the Bidirectional Gated Recurrent Unit (BGRU) was used in order to attain interdependent features with long distance. The introduced hybrid model has the leverage in simple work structure and less training time in contrast to attention-based model that merges self-attention mechanisms and CNN. Two short review datasets for used to evaluate the performance. By using the machine learning approaches like semantic-orientation approach and support vector machines, [2] developed an hybrid concept in neural networks for enhancing the conduct measures of Egyptian dialect sentence-level sentiment analysis. The result obtained shows the compelling advancements in terms of the accuracy, precision, recall and F-measure, denoting their developed hybrid approach was better in sentence-level sentiment classification.[3] Introduced a model for extended vocabulary and CNN for the Weibo comment texts and wiki Chinese data set by expanding train word embeddings and the vocabulary for sentiment classification in sentence level. Using k-Max pooling method they captured more features for classification accuracy. In overcoming the challenging task, quantitative feedback system [4] used machine learning and deep learning algorithms in sentiment analysis for solving issues at earlier stages and in enhancing productivity and sales. In overcoming the classification problem based on the subject of the review [5], opted Navie Bayes algorithm while observing the importance of data cleaning for Fintech Industries of Indonesia in order to challenge the competitors by knowing the user opinion in real time. Explaining the importance of the

overshoot percentage and settling time as performance parameters [6] opted Genetic Algorithm in Optimization to achieve steady state condition in shortest time for ITE error criterion in PI controller. For extracting the sentiments of tweets based on their polarity and subjectivity, visualizing and classifying their results graphically, [7] proposed an approach that encompasses a hybrid learning method for the given partially labelled training data by classifying the tweets based on Bayesian probabilistic method for sentence level models. The uniqueness of this research belongs to the hybrid learning method for sentiment analysis.[8] Proposed an versatile sentiment analysis approach for analyzing post in social media and extract users' opinion in real-time by collecting the data related to a set of selected hashtags into a productive dictionary of words based on polarity and conducted a case study for 2016 US election for guessing favorite candidate. For reducing memory consumption and increasing classification performance [9] introduced a Distribution Feature Selection (DFS) using Symmetrical Uncertainty (SU) and Multi-Layer Perception (MLP) which is applied on 7 well high dimensional and 1 low dimensional dataset. [10] Focuses on classification problems in high dimensional datasets in the medical field. A novel M-cluster feature selection (Mcfs) based on Symmetrical Uncertainty (SU) using Attribute Evaluator was proposed for clusters. KNN- lazy learner, Naïve Bayes (NB) classifier, J48 - rule-based learner, J48 - tree-based learners have also experimented for the proposed method.[11] Explains the graph problems in Natural Language Processing (NLP) and their applications by comparing a collection of texts (corpus) to different clustering methods applied in Pharma Covigilance terms and identify the semantic relations in the biomedical terms.[12] Presented an approach using Genetic Algorithm called Hybrid-based Single-document Extractive Arabic Text Summarization that calculates accuracy by using EASC corpus, and ROUGE evaluation method. A novel approach that is based on the tagging technique parts-of-speech (POS), lexicon-based approaches, word position algorithm and bench mark sentiment datasets was implemented in [13], whose experimented results showed that Improved Word Vectors(IWV) are very impressive for sentiment analysis. Some new features were proposed and evaluated which were accomplished through lightweight method of discourse analysis and were used in hybrid lexicon and machine learning based classifier[14].Co-relating with other traditional machine learning approaches, their results showed improvement in classification accuracy with their new features.

To extract implicit as well as explicit aspects, [15] proposed rule-based hybrid technique that uses sequential patterns and normalized Google distance (NGD) and groups synonyms by proposing a Google similarity distance in conjunction with *particle swarm optimization (PSO).* Focusing on less aspect of Sentiment Analysis, which is lexicon-based SA for Arabic language, [16] experimented and compared results of three different lexicon construction techniques and an Arabic tool is designed and implemented for taking advantage of the effectively constructed lexicons. The proposed SA tool possesses many novel features such as the way negation and intensification are handled. [17] For classifying sentiment on Meta- level features based on Twitter sentiment analysis where many methods and lexical resources were proposed previously for extracting sentiment indicators from both syntactic and semantic levels in Natural Language. They also addressed the different approaches for subjectivity and dimensional problems. [18] Proposed a latest approach to solve Arabic Word Sense Disambiguation (AWSD), where WSD is the complication for assigning a meaning for the given text. This experiment was performed using GA and tests their algorithm performance on Arabic sample text and Naïve Bayes classifier. [19] Presented an evaluation method for schema theory in designing genetic coding for NN topology optimization. They also determined excellent averages among different Evolutionary operators based on essence of coding pattern and testes on 2 GA-NN hybrid systems for NLP and Robot navigation.

3. PROPOSED SYSTEM

A combined repetitive, global max-pooling and convolutional layer for pre-trained Word2Vecmodel was performed in our hybrid model. In our experiment, to abstract the local features of input text and to take advantage of the concepts of LSTM, a very wide architecture of convolutional layers is used. It grabs the long-term dependencies among the progression of words. For experimenting IMDB movie reviews dataset by Cornell University was used in which, the dataset is break down into a train set and a test set.

A brief explanation of the work in this study is as follows:

1. An unsupervised model for word embeddings using Word2Vec is performed which is trained on immense number of words. Semantics of words is captured by this model.
2. LSTM model is used to notice deeper semantics of words for capturing sentiment polarity from texts. The long-term dependencies amidst the word strings in long texts are learned by the proposed model efficiently.
3. For further clarifications in embeddings for supervised dataset a wide CNN model is used. We use many weight matrices with different window lengths to generate number of features.
4. CNN model is advantageous by LSTM model that extracts the long distance dependencies and local

features which were combined into a single hybrid CNN-LSTM model. Our Experimented result shows that our model achieves competent results.

3.1. Research and Consequences

3.1.1 Datasets

The experiment is conducted by using IMDB Movie Reviews Dataset,[1]published by Cornell University, was given as the sentiment analysis inputcorpus.The data set is divided into 1780 samples for training and 445 samples for testing. To test the introduced model it was also experimented on IMDb dataset of 50k movie review by dataset that has 25000 test samples and 25000 train samples Shown below, is the proposed system's block diagram

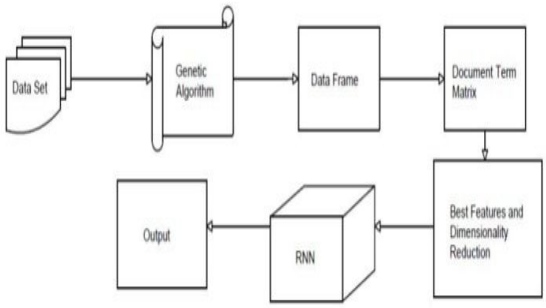

Figure 1: Proposed model for genetic optimization in hybrid level sentiment analysis for opinion classification

The details of our proposed model are shown in the above shown diagram, which has convolutional and recurrent neural networks. The input word embeddings were feed into the convolution layer to extract features which translate text into Word2Vec. Classification layer is applied in the end where the output of the Convolutional layer which is given to LSTM. LSTM learns long-term dependencies amidst the sequence of the words.

As the deep learning method does not understand the human text directly, a technique, Word2Vec is used. It takes a bunch of sentences as input and transforms them into their corresponding vector values. These vector values of different input words were then computed with neighboring vector values thus calculating the distance between them. The input of our CNN model are these vector values.

3.1.2 The presented Model HYBRID CNN-LSTM

The pre- processing stage which takes the input corpus data carries out sentence segmentation, tokenization, stop word removal and stemming tasks. High level features were extracted by the Convolutional layer and dependencies between the words are detected by LSTM layer. A sigmoid function is used in the classification layer. Essentially, text data is converted to numerical data by Word2Vec and is applied to CNN for training numerical values. In this procedure, we use *three* pooling layers, *three* convolutional layers and *a* fully connected layer. For experiments in numerical computations, *tensor flow* an open source python library was used. In CNN for increasing accuracy, pooling layers, convolutional layers, dropout layers and RELU are used. To overcome over-fitting problems in machine learning, *dropout techniques* are used by skipping the neurons that do not contribute in back propagation and co-adaption.0.5 probability is the output for each hidden neuron. We discuss the proposed CNN model of different layers in the following sub sections.

3.1.3 CNN Initialization

Pre-Processing In this paper, for word embeddings a pre-trained Word2Vec is used.

Convolution Layer There are 7 layers in CNN model which are three convolutional layers, three pooling layers and one fully connected layer. The input words in the form of sentences are passed from embedding layer to convolutional layer. The input is convolved by the Convolutional layer using pooling layers helps in reducing the portrayal of input parameter, input sentences, controls overfitting and computation in the network.

Global Max-Pooling is a pooling operation which is used to calculate highest or maximum values in each sector of a feature map which results are sampled to get or highlight most present or used feature map. /this is used to retrieve the best global results from the total network by applying it at different convolutional layers at the end of each network layers.

Activation Function establishes the output of a deep learning model. For this trail, we used ReLU (Rectified Linear Unit), a non-linear activation function that gives zero for negative values and it increases with positive values

Dense Layers or fully connected layers which applies scaling, rotation and translation transform for the vector is used to in convolutional layers for implementing classification on the extracted features. Every prevailing neuron in this network associates with every input neuron in the approaching layer of the same network and continues.

Softmax calculates the median for the arbitrary results in the concluding layer of the neural network from 1 and 0. Fig.1shows the proposed model using CNN.

3.1.4 LSTM In the Model

Embedding Layer random weights are initialized to the words by embedding layer and all the words in are embedded in the training dataset using Word2Vec model.

LSTM Layer After embedding, to handle the flow of information in a network, LSTM is used in the RNN layers having three types of gates and cells which are forget gate, input gate and output gate

Dropout Techniques These techniques dump the redundant data from the network that does not help in further processing which enhances the effectiveness of our design by preventing overfitting. At training session, with each iteration we drop a neuron with probability *a* that disables all the inputs and outputs of it temporarily Until the completion of that iteration. These disabled neurons will be active in the next iteration. This technique is used as to reduce the usage of all the neurons to operate independently and to prevent the dependency on the network which leads in off handling the code for autoencoder.
In the convolutional layer, the pooling layer has the prevalent factors except 1) a filter or a kernel having no value; 2) whose output is the maximum (generally) or median value at the range it stops, and 3)Throughout the pooling layer, the spatial size of input is not changed. If the maximum value is returned, then he type of the pooling layer is max, otherwise mean. Therefore, the parameters used are the kernel size, stride size i.e., the number of shifts by the pixel in the input matrix and the pooling type.

Genetic Algorithms Primarily, we initialize the random population of individuals and calculate its fitness using a deterministic function called *fitness function* for each individual based on the context of the problem to be optimized. Decision variables that are encoded in the individuals are the input for this function. The *selection* operator chooses the individual with best fitness, to achieve off-springs with best fitness. Respectively, the *crossover* operator produces the off-springs by exchanging or modifying the bits of the *chromosome* and produces off-springs which are reproduced by the *mutation* operator. Again these off-springs are treated as new population and the process repeats where the initial population is called *parent population* and the derived off-springs are called *children or new population of off-springs.* We can find the expected optimal solution from the population by repeating these series of operations by the end of GA. Note that, the classic principle of GA is to terminate the predefined maximal generation number 10, in this experiment.

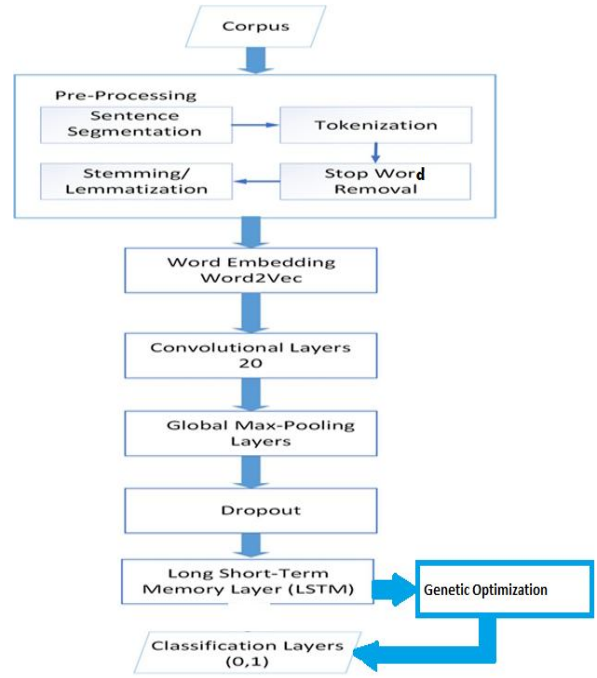

Figure 2: Methodology of the proposed system

3.1.5 Proposed Algorithm

Input: Text Sentiment File Feature as Token Value

Output: Optimized token value Initialize the

GA with their operating functions –

Population size define fitness function by GA for Feature Selection

Fit_Function=True;

If fs>ft

False; else

// to check fitness of each data in feature list if data fulfill the condition, then those are considered in feature list else replace with objective value
Calculating Feature size in the terms of rows and columns (r, c)
for a=1 to r
for b=1 to c
Fs=Feature (a,b) // to select one by one feature
Ft=Threshold (a,b) // mean of text token values
Fit_func_GA =Call Fit_func_GA (Fs, Ft)
No. of Variables=1
Fitdata=GA (Fit_func_GA, No. of Variable, GA functions)
// apply GA on feature data with fitness function of system
End
End'

Returns: Best Fit data as an optimized token value of the text file
// return optimized feature set which helps to train the CNN
End

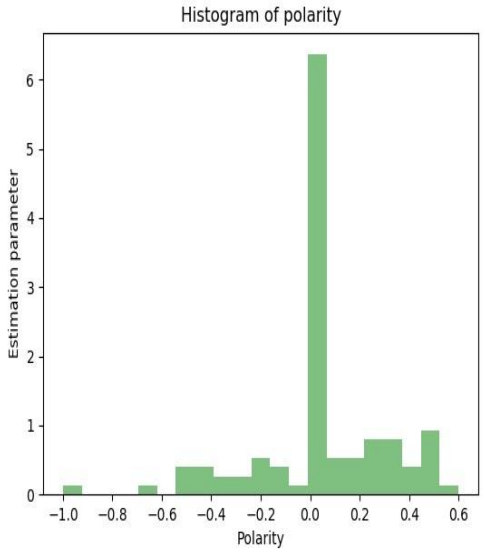

Figure 3: Polarity distribution for IMDb movie reviews dataset.

From the above bar chart, we say that the considered dataset i.e., IMDB movie reviews dataset has more neutral reviews than the positive and negative reviews. (The graph varies with the sentiment of the data considered. It may differ for other datasets). In this experiment, the IMDb movie review dataset has 1780 train samples and 445 test samples of total 2225 samples and the IMDB 50k dataset of movie reviews has 25000 train samples and 25000 test samples. The calculated polarity accuracy for each generation by genetic algorithm considering for 10 iterations for both datasets is tabulated below.

In our experiment, the considered probabilities for crossover and mutation are 0.5 and 0.2 respectively for both datasets. In the model, we defined the sentiment by comparing the polarity with 0.1 as probability. The below graph shows the distribution of polarities in the Movie review dataset by Cornell University.

Table 1: Comparative results of accuracy in percentage for the experimented datasets.

Epoch no.	Accuracy in percentage for IMDb movie review dataset	Accuracy in percentage for IMDb 50k dataset of Movie reviews
1/10	33.54	31.80
2/20	33.60	36.52
3/10	37.53	39.66
4/10	55.84	58.88
5/10	73.31	72.98
6/10	85.22	86.12
7/10	91.52	92.08
8/10	93.76	94.44
9/10	95.00	94.83
10/10	95.22	95.51

Our presented model shows an accuracy of 95%, a best performance on both datasets because of the usage of genetic algorithm.

4. CONCLUSION

In analyzing peoples' attitude and behavior, Sentiment Analysis (SA) has proven its efficiency. Depending on the data from the social media on specific topics, a unique approach was designed in order to train the machine efficiently and to attain best results. For implementing this activity,70% of dataset was 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 95%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. Using genetic algorithm in classification with CNN is an

added advantage as the process undergoes choosing the best chromosomes in the population that yields best results which again leads to the new best chromosomes for the next generation.

REFERENCES

[1] Yongping Du, Xiaozheng Zhao,Meng He, AndWenyangGuo **"A Novel Capsule Based Hybrid Neural Network for Sentiment Classification"**, IEEE. Vol. 7, Page(s): 185001 – 18501, 2019. https://doi.org/10.1109/ACCESS.2019.2906398

[2] Rezaeinia, S. M., Rahmani, R., Ghodsi, A., &Veisi, H, **"Sentiment analysis based on improved pre-trained word embeddings."** Expert Systems with Applications, vol.117, 139-147, 2019. https://doi.org/10.1016/j.eswa.2018.08.044

[3] Yang, X., Xu, S., Wu, H., &Bie, R, **"Sentiment Analysis of Weibo Comment Texts Based on Extended Vocabulary and Convolutional Neural Network."**Procedia Computer Science, 147, 361-368, 2019 https://doi.org/10.1016/j.procs.2019.01.239

[4] Maganti Syamala, N.J. Nalini **"A Deep Analysis on Aspect Based Sentiment Text Classification Approaches."** International Journal of Advanced Trends in Computer Science and Engineering, vol.8, no.5, 2019. https://doi.org/10.30534/ijatcse/2019/01852019

[5] Rein Rachman Putra , Monika Evelin Johan, Emil Robert Kaburuan, **"A Naïve Bayes Sentiment Analysis for Fintech MobileApplication User Review in Indonesia"** International Journal of Advanced Trends in Computer Science and Engineering, vol.8, no.5, 2019.

[6] A.A.M Zahir, S.S.N Alhady, W.A.F.W Othman ,Zhiling Low , A.A.A Wahab **"Genetic Algorithm Optimization and Implementation of Velocity Control PI Controller for Cart Follower Application."** International Journal of Advanced Trends in Computer Science and Engineering, vol.8, no.5, 2019. https://doi.org/10.30534/ijatcse/2019/12852019

[7] Ketaki Gandhe, Aparna S. ,Vardevardea, Xu Du **"Sentiment Analysis of Twitter Data with Hybrid Learning for Recommender Applications",** 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2018.

[8] El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A., & Kobi, A, **"A novel adaptable approach for sentiment analysis on big social data."** Journal of Big Data, 5(1), 12, 2018.

[9] Pothuraju S.P, Sreedevi .M, **"Distributed feature selection (DFS) strategy for microarray gene expression data to improve classification performance."** In Clinical Epidemology and Global Health Volume 7, Issue 2, 171-176, 2018.

[10]. Sai Prasad Potharaju, M.Sreedevi, **"A novel m-cluster of feature selection approach based on symmetrical uncertainty for increasing classification accuracy of medical datasets."** Journal of Engineering Science and Technology Review 10 (6) 154 – 162, 2017. https://doi.org/10.25103/jestr.106.20

[11] Engström, Christopher, **"PageRank in Evolving Networks and Applications of Graphs in Natural Language Processing and Biology"** Mathematics/Applied Mathematics ORCID iD: 0000-0002-1624-5147, 2016.

[12] Yazan A. Jaradat ; Ahmad T. Al-Taani, **"Hybrid-based Arabic single-document text summarization approach using genatic algorithm"** IEEE, 2016 7th International Conference on Information and Communication Systems (ICICS), 2016.

[13] Amira Shoukry, Ahmed Rafea**"A Hybrid Approach for Sentiment Classification of Egyptian Dialect Tweets"** 2015 First International Conference on Arabic Computational Linguistics (ACLing). IEEE 2016. https://doi.org/10.1109/ACLing.2015.18

[14] Daniel Ansari SenticNet **"Sentiment Polarity Classification using Structural Features"** IEEE International Conference on Data Mining Workshop (ICDMW), 2015. https://doi.org/10.1109/ICDMW.2015.57

[15] Toqir Ahmad, Yu-N Cheah **"Hybrid Rule-Based Approach for Aspect Extraction and Categorization from Customer Reviews"** 9th International Conference on IT in Asia (CITA), 2015.

[16] Nawaf Abdulla, Roa'sMajdalawi, Salwa Mohammad, Mahmoud Al-ayyoub and Mohammad Al-kabi, **"Automatic Lexical Construction For Arabic Sentiment Analysis"**, International Conference on Future Internet of Things and Cloud, 2014. https://doi.org/10.1109/FiCloud.2014.95

[17] Bravo-Marquez, F., Mendoza, M., & Poblete. B, **"Meta-level sentiment models for big social data analysis."** Knowledge-Based Systems, 69, 86-99, 2014. https://doi.org/10.1016/j.knosys.2014.05.016

[18] Mohamed El BachirMenai; WojdanAlsaeedan, **"Genetic Algorithm for Arabic Word Sense Disambiguation"** IEEE, 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2014

[19] J. Davila, **"A new metric for evaluating genetic optimization of neural networks,"** IEEE Symposium on Combinations of Evolutionary Computation and Neural Networks. Proceedings of the First IEEE Symposium on Combinations of Evolutionary Computation and Neural Networks, 2000.