Text Reviews Sentiment Mining by Invasive Weed Optimization and Deep Learning

Rajesh Sisodiya[*], Dr. Praveen Kumar Mannepalli[**]

School of Computer Science and Technology, LNCT University Bhopal

Abstract— Any kind of business depends on customer feedback. Best resource of such feedback is online platforms where user share its experience. Analysis of such reviews play an important role for the product, service improvement. This paper has developed a sentiment class identification as per user text reviews. In order to improve classification feature set were optimized by use of weed optimization algorithm. As per dynamic nature of algorithm no prior training or knowledge need to be used for feature optimization. Selected feature set were transform into numeric form. Transformed feature set were used for the training of linear regression model. Experiment was done on real dataset of user text reviews. Result shows that proposed weed and deep learning model has increased the sentiment detection model accuracy.

Keywords:- Sentiment mining, Digital review analysis, Text mining, Feature Extraction.

I. INTRODUCTION

An algorithmic formulation is used to recognise opinionated information and label it as 'positive, negative, or neutral' polarity. For most of us, "what other people think" has always been a crucial piece of information when making decisions [1]. Users' opinions not only assist individuals in making well-informed decisions, but they also assist businesses in discovering client attitudes and opinions about products and services. The new user-centric, participatory Web allows a massive number of people to express themselves on nearly any issue, from movie reviews to product reviews to social and political events. However, in the lack of automated ways to extract meaningful and comprehensive information from the vast amount of data available on the Web (including multiple social media platforms), information overload occurs [2].

Sentiment analysis, often known as opinion mining, is the study of people's attitudes, sentiments, assessments, and emotions regarding entities and their properties as expressed in written language [3]. Products, services, organisations, personalities, events, causes, and topics are all examples of entities. The field encompasses a vast problem domain. Sentiment analysis currently encompasses a wide range of related terms and tasks, including sentiment analysis, opinion mining, opinion analysis, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, and review mining.

In this sector, the data sets employed in SA are a major concern. Product reviews are the primary source of information. These reviews are significant to business owners because they can make business decisions based on the outcomes of the analysis of customer opinions about their items. The majority of the reviews come from review sites. SA can be used in a variety of settings, including product reviews [4], financial markets [5], news stories [6], and political arguments [7]. For example, in political debates, we could learn about people's views on specific election candidates or political parties. Political positions can also be used to forecast election outcomes. People openly express and discuss their thoughts about a topic on social networking and microblogging sites, which are regarded excellent sources of knowledge. In the SA process, they are also used as data sources. Social networking sites can quickly supply all of the information needed to make a certain decision, such as purchasing any item A from a shopping site. Social media is a fantastic platform for expressing oneself to the rest of the world. The amount of data accessible to mine for opinions is incredible. Many research projects have been conducted based on the study of sentiments posted on social media. Sentiment analysis presents a new set of challenges for extracting information from natural language text. Sentiment analysis is a field that seeks to categorise and analyse all of the opinions stated in natural language. Sentiment analysis is conducted on review sites and social media platforms such as Twitter, where tweets provide us with more accurate and diverse opinions from individuals all around the world on topics such as the latest cellular phone, the iPhone6. The buyer's selection will undoubtedly be influenced by product reviews.

II. RELATED WORK

Pecar et al. [6] suggest using a model ensemble to improve sentiment analysis in Slovak customer evaluations and provide references to more recent work on sentiment classification in Slavic languages.

Tsakalidis et al. [8], which includes the POS tag, subjectivity, polarity, and intensity of the six primary emotions for each phrase.

Most supervised learning techniques, according to Saif, He, Fernandez, and Alani, [8] 2016, rely on training classifiers, such as Nave Bayes (NB), Maximum Entropy (MaxEnt), and Support Vector Machines (SVMs), built from various functionalities-word n-grams, Part-Of-Speech (POS) tags, with or without: words' earlier sentiment, words' semantic concept, sentiment topic features, semantic examples, and tweets syntax features These tactics have had generally positive results, with accuracy rates ranging from 80 to 84 percent.

Rani and Kumar [9] (2017) report on sentiment analysis that used a supervised half and hybrid method that combined hidden Markov models with support vector machines to determine the polarity of e-learning sites and discussion platforms (SVMs).

To take use of numerous modalities, the authors of [10] proposed a joint type visual-textual sentiment analysis. Furthermore, a cross-modality mechanism was proposed, as well as a semantic embedding technique based on bi-directional neural networks. The major goal was to create a model that focuses on textual and visual features that contribute to sentiment classification. The goal of this approach was to show that the image and text do not contribute equally to the sentimental classification findings since textual and visual information differs in giving sentiment analysis results.

In [11], Yue Feng et al. The sentiment analysis was done using MCNN-MA, or multi-head attention mechanism, in conjunction with a multi-channel convolution neural network. This model evaluates parameters like position features, speech features, dependence syntax characteristics, and gives extra features by combining them, before sending the results to a multi-channel neural network. This concept also included a multi-headed attendance mechanism to obtain the essential feelings.

III. Proposed Methodology

This section of the paper covered sentiment mining using a raw natural language dataset. A hybrid model of soft computing techniques was proposed in this paper. The job was broken down into three sections, the first of which was data processing and feature extraction. The invasive weed optimization approach was used to reduce the number of features. Deep learning with convolutional, maxpooling operator and linear regression learning came in third. Figure 1 depicts the operation of the first and second modules, while Figure 2 depicts the operation of the third module.

Module 1: *Pre-processing and Feature Extraction*

Enter the content in its raw form. RT is a one-sentence blog written by a random user. R may have one or more sentences to process in each review. As a result, the initial step in this module is to use the stopwords removal technique to eliminate noisy words from the material as stopwords. Sw are frequently used in text to frame a statement [17]. Words that do not appear in the Sw list are filtered and gathered in a B vector bag. Each R has a B of its own. The entire set of B is listed in a single word vector called U. The words that were classified had to fall into two categories: positive and negative.

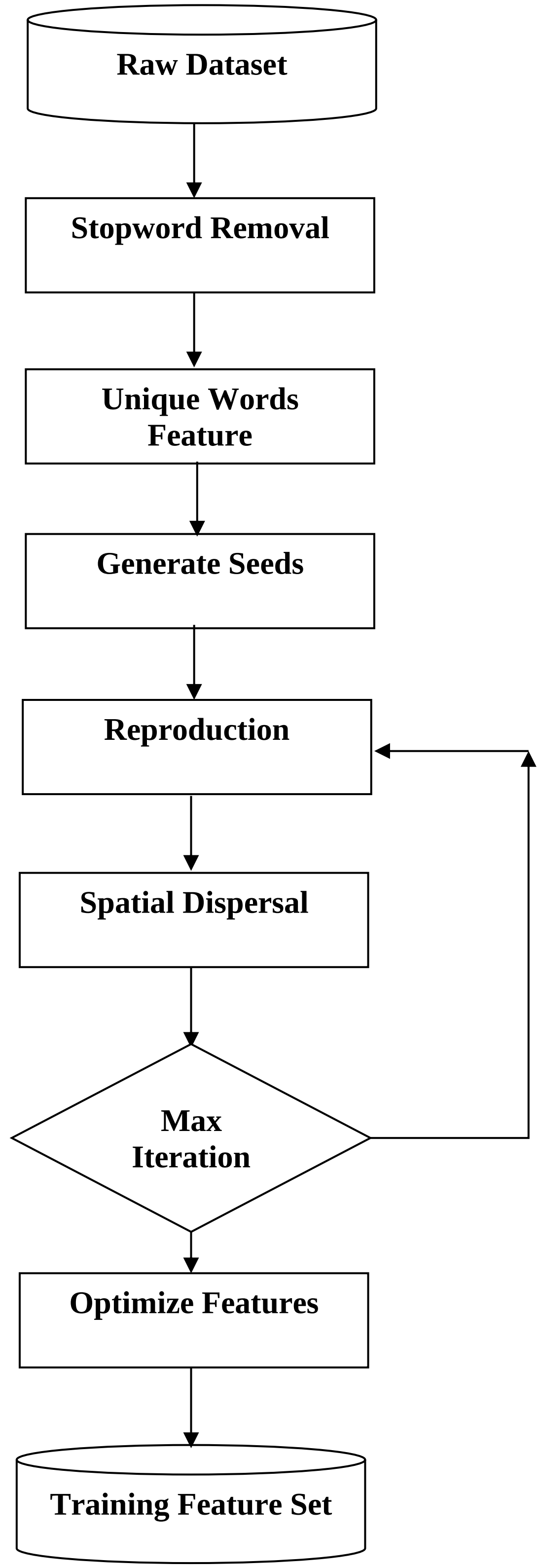

Fig. 1 Review content feature optimization.

The methods above provide you with a one-of-a-kind set of text terms for the job. Furthermore, each word in U was transformed into a numeric value for term frequency that was normalised by the total number of words counted in B.

Module 2: *Invasive Weed Optimization Algorithm*
Some of the words in U should be deleted in order to minimise the feature set for learning. Invasive Weed optimization was responsible for these feature upgrades (IWOA). IWOA was proposed in [14] with the goal of improving seed set through balanced weed optimization. This seed and weed optimization approach is used in the paper to filter phrases found in the U.S.

Generate Seeds
Produce Seeds
In the algorithm, each element in U acts as a seed element. In the population, chromosomes are seeds. Seed is a vector of 0 and 1, with 1 indicating that the term is present in the seed and 0 indicating that the term was not selected. The size of a seed is |U| let t terms, and the population Sp has s seeds.

SP←Gaussian_Distribution(t, s) --------Eq. 1

Reproduction
In this step fitness of the seed was checked by bowing in a area and check its production. Production of seed was test by fitness function. Chromosome having good fitness value are promote while other are replace or transform into other type of seed.

$F_s = \sum_{i=1}^{t} Tf_i \times S_{Pi,s}$--------Eq. 2

Spatial Dispersal
Fitness value of seed helps to find the better seed set. To improve the solution seed quality some changes need to be done randomly in low quality seeds. This operation act as crossover operation done in genetic algorithm. As solution need improvement in each iteration hence seed element get change from 1 to 0 or 0 to 1.

Optimize Features
After m number of max iteration best fitted seed act as final or optimize term set for the training of neural network. Seed vector having position element value 1 act as selected term and other or element value 0 acts as rejected terms (Weed). Terms were select as final training feature set Sf. This paper has found that use of IWOA has increased the work efficiency of the sentiment mining.

Review processed terms in B are further select or reject as per final seed vector Sf to get optimized feature set Of.

```
Input: B, Sf
Output: Of
Loop 1:RT
        Loop 1:t
                If Sf[t] is Nonzero
                        Of←B[t]
                EndIf
        EndLoop
EndLoop
```

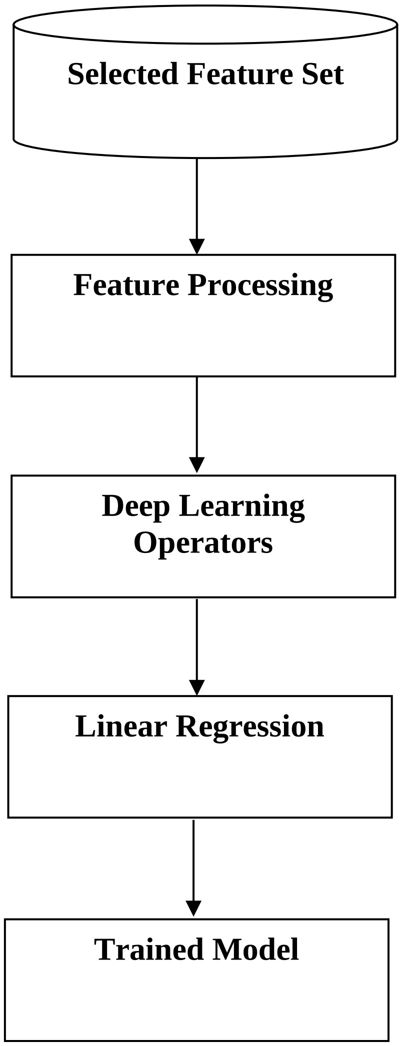

Fig. 2 Module 3 block diagram.

Module 3: *Deep Linear Regression*
This module learned text features because each selected word has its own class (positive/negative), hence the required output is known. This section of the paper processes the input text terms in order to train the linear regression model.

Feature Processing
Input selected words from BOW were transformed into an ASCII number sequence based on character presence. This is understandable. If the word "JOY" is used, the ASCII number sequence for it is 74, 79, 89. Each number is converted to an eight-digit binary sequence: 01001010, 01001111, 01011001. Because the number of letters in a word varies, the input feature vector was a binary sequence of 100 bits. This 100-bit vector is transformed into a 10-by-10 matrix that can be used to apply learning operations. Operators for Deep Learning

Deep Learning Operators
To improve the efficiency of this effort, deep learning operators were used to minimise the dimension of the input feature set [18]. High-value feature values are filtered using the convolutional operator. A hxh fixed size filter moves from left

to right and top to bottom in a 10x10 matrix. The filter has a range of 0 and 1 values. The maxpooling operator reduces the dimension of the 10x10 input feature, and the pooling window pxp applies left to right and top to bottom according to the reduction ratio. P has a certain amount of seeds.

Linear Regression Learning|

Two dimension feature set obtained from deep operato was reset into single dimension vector. As work known the class of the word positive / negative so its representative output number is 1/0. Each input feature pass in the model to get desired output. Once regression over then work will get trainind Regression model [19]. This model takes processed word vector as input and predict its class (positive and negative).

IV. Evaluation Parameter

Implementation of proposed IWOLR model was done on hardware having 4GB RAM, I5 processor configuration. Experimental was performed on MATLAB software. Input dataset was taken from reviews of AMAZON website having [14, 15]. Comparison of models was done by MFLRSA model proposed in [16] and MCNN-MA model proposed in [12].

V. Results

Table 1 User review based sentiment class identification precision values comparison.

Dataset	MFLRSA	MCNN-MA	IWO-LR
200	1	0.5055	0.9677
300	0.5328	0.5214	0.9787
500	0.9878	0.4772	0.9874
1000	0.9588	0.4744	0.9916
1500	0.9574	0.4902	0.9946

Precision value table shows that proposed IWOLR model has increases the parameter values by 9.82% as compared to previous model proposed in MFLRSA [16]. Use of linear regression model fro learning of selected feature gives better output for the work. It was also shown from the able that with increase of review testing dataset parameter is always close to 96 or above.

Table 2 User review based sentiment class identification recall values comparison.

Dataset	MFLRSA	MCNN-MA	IWD-LR
200	0.8462	0.38	0.7895
300	0.8488	0.3989	0.8415
500	0.88	0.3925	0.8906
1000	0.9012	0.41	0.9147
1500	0.909	0.4227	0.9304

Table 2 shows sentiment class identification recall value comparison. Use of term based text feature for the sentiment identification in IWOLR model has increases the recall parameter value. As feature has increases the classification efficiency of invasive weed optimization work. Use of invasive algorithm in the model for feature optimization indirectly enhance the learning accuracy of the work.

Table 3 User review based sentiment class identification f-measure values comparison.

Dataset	MFLRSA	MCNN-MA	IWO-LR
200	0.9167	0.434	0.8696
300	0.6547	0.452	0.9049
500	0.9308	0.43	0.9365
1000	0.9291	0.4399	0.9516
1500	0.9326	0.454	0.9614

F-measure value table shows that proposed IWOLR model has increases the parameter values by 5.62% as compared to previous model proposed in MFLRSA [16]. Use of linear regression model fro learning of selected feature gives better output for the work. It was also shown from the able that with increase of review testing dataset parameter is always close to 92or above.

Table 4 User review based sentiment class identification accuracy values comparison.

Dataset	MFLRSA	MCNN-MA	IWO-LR
200	0.91	0.4	0.865
300	0.4867	0.408	0.9033
500	0.928	0.392	0.936
1000	0.929	0.4099	0.952
1500	0.9307	0.4581	0.9606

Table 4 shows sentiment class identification recall value comparison. Use of term based text feature for the sentiment identification in IWOLR model has increases the recall parameter value 9.36% as compared to [16].. As feature has increases the classification efficiency of invasive weed optimization work. Use of invasive algorithm in the model for feature optimization indirectly enhance the learning accuracy of the work.

Table 5 User review based sentiment class identification testing time (seconds) values comparison.

Dataset	MFLRSA	MCNN-MA	IWO-LR
200	0.1489	0.17535	0.1233
300	0.12904	0.16299	0.1109
500	0.154786	0.196046	0.1438
1000	0.158287	0.254256	0.1352
2000	0.165351	0.265634	0.1487

Linear regression models takes less execution time for predicting the class of a review as compared to other exsiting models proposed in the work. It was obtained that neural network takes more time and convolution filter have high time as compared to regression models.

VI. Conclusion

Review collection and analysis is great use for the online platforms as this directly impacts to sales of any product or services. This paper has proposed a sentiment class identification model that can finds the class of review without any background knowledge of the product. It was obtained from the survey of different scholars that learning model increase the work efficiency. This work proposed a feature selection algorithm by implementing invasive weed optimization algorithm. Selected features were further process for improving the learning efficiency. Linear regression model was used in the work for learning and texting of proposed model. Result shows that proposed IWOLR has increases the precision parameter by 9.82%, while accuracy of sentiment class identification was improved by 9.36% as compared to other existing models. In future scholars can implement same model on other language review contents.

References

[1] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1–135.

[2] Liang-Chih Yu, Jheng-Long Wu, Pei-Chann Chang, Hsuan-Shou Chu Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news Knowl-Based Syst, 41 (2013), pp. 89-97

[3] Xu Tao, Qinke Peng, Yinzhao Cheng Identifying the semantic orientation of terms using S-HAL for sentiment analysis Knowl-Based Syst, 35 (2012), pp. 279-289

[4] Isa Maks, Piek Vossen A lexicon model for deep sentiment analysis and opinion mining applications Decis Support Syst, 53 (2012), pp. 680-688.

[5] Nikiforos, M.N.; Voutos, Y.; Drougani, A.; Mylonas, P.; Kermanidis, K.L. The Modern Greek Language on the Social Web: A Survey of Data Sets and Mining Applications. Data 2021, 6, 52.

[6] Pecar, S.; Šimko, M.; Bielikova, M. Improving Sentiment Classification in Slovak Language. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, Florence, Italy, 2 August 2019; pp. 114–119.

[7] Tsakalidis, A.; Papadopoulos, S.; Voskaki, R.; Ioannidou, K.; Boididou, C.; Cristea, A.I.; Liakata, M.; Kompatsiaris, Y. Building and evaluating resources for sentiment analysis in the Greek language. Lang. Resour. Eval. 2018, 52, 1021–1044.

[8] Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. Information Processing and Management, 52(1), 5–19.

[9] Rani, S., & Kumar, P. (2017). A Sentiment Analysis System to Improve Teaching and Learning. Computer, 50(5), 36–43.

[10] Zhu X., Cao B., and Xu S. et al: 'Joint visual-textual sentiment analysis based on cross-modality attention mechanism', Int. Conf. on Multimedia Modeling, Thessaloniki, Greece, 2019,

[11] Yue Feng And Yan Cheng. "Short Text Sentiment Analysis Based on Multi-Channel CNN With Multi-Head Attention Mechanism" School of Computer Information Engineering, February 3, 2021.

[12] A.R. Mehrabiana,b, C. Lucasc. "A novel numerical optimization algorithm inspired from weed colonization". Science Direct, cological Logistics 2006.

[13] Leila Mohammadpour, T.C. Ling, C.S. Liew, Alihossein Aryanfar, "A Mean Convolutional Layer for Intrusion Detection System", Security and Communication Networks, vol. 2020.

[14] https://ijsret.com/2017/12/14/computer-science/

[15] https://www.kaggle.com/bittlingmayer/amazonreviews

[16] Rajesh Sisodiya, Dr. Praveen Kumar Mannepalli. "Moth Flame Optimization and Deep Learning Based Digital Review Sentiment Mining". ISSN: 0011-9342, issue 7 2021.

[17] Rajesh Sisodiya, Dr. Praveen Kumar Mannepalli. "A Survey on Social Digital Data Based Sentiment Mining Techniques and Feature " has been " Accepted for Publication " in International Journal of Computer Trends & Technology - IJCTT , April 2021.

Authors

First Author – Rajesh Sisodiya, Research Scholar, School of Computer Science and Technology, LNCT University Bhopal MP India and sisodiya18jan@gmail.com

Second Author – Dr. Praveen Kumar Mannepalli, Ph.D, School of Computer Science and Technology, LNCT University Bhopal MP India and praveen.hawassa@gmail.com.