An Optimized Bayesian Probabilistic Neural Network Model for Sentiment Analysis

Chaturbhujachari Chaturvedi[1], Prof. Preetish Kshirsagar[2]

***Abstract*— Of late, big data and big data analytics has fund applications in diverse fields. Social media and allied applications is one such domain for research, where Artificial Intelligence has shown unprecedented impact. In this paper a mechanism has been proposed which can classify text data into classes of different sentiments. Data in the form of tweets has been used in this case. Pre-processing of raw data has been done prior to using it to train a neural network. A Neural Network is then trained using the categories of the data which are tweets that correspond to happy, neutral and sad moods of the Twitter users. The Bayesian Regularization (BR) algorithm has been used for training the artificial neural network. It has been observed that this proposed technique achieves an MAE of 0.6 which is significantly lesser compared to previously existing work.**

Keywords—Sentiment Analysis, Machine Learning, Deep Nets,Bayesian Regularization, Mean Square Error (MSE), Mean Absolute Error (MAE).

I. INTRODUCTION

The advent of data analytics has been enormous and text mining and opinion mining has garnered huge importance because of its broad range of applications in a variety of domains like the social media, analytics of data, business applications etc. Sentiment analysis can be defined as a study that is based on a computational analysis and determination of textual opinions, emotions, behaviour and the attitude exhibited towards any entity [1] Sentiment analysis tries to find out the attitude or an opinion of the user based user's textual data. It also aids in making decisions. Sentiment Analysis helps in determining whether the piece of tweet or any piece of writing is positive, negative or neutral. It analyses the sentiment behind the text of any user, hence it helps companies for product reviews and enhance business prospects [2]. It has got a broad range of applications today, especially in the areas where outcomes are dependent on human sentiments and opinions. It can also be considered as opinion mining. To be able to analyse and implement such tasks, Artificial Intelligence is used. In this context, the concept of data mining is utilized which a knowledge based procedure which is based on extraction of skilled patterns and information [3]. The extracted data is then used in visualization of applications and creation of real time programs for the process of decision making. The applications can be diverse such as marketing and finance, advertising, opinion polls, social media, product reviews just to name a few. The following diagram illustrates the mechanism [4].

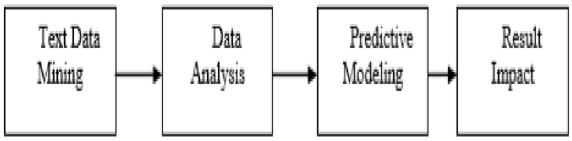

Fig.1 Text Mining model for sentiment analysis

While several data sources are available on the internet to be mined, yet a judicious use of web mining is to be done prior to any system design model is to be used. The critical factor is also the feature selection from the raw data to be included in the analysis of the data as a whole. The unstructured text mining approach is often used and the text is to be replaced with suitable tokens or numerical counterparts prior to training any designed mechanism for the classification of the text data [5]. While data as a whole can consist of more than textual data, hence pre-processing of the data is of topmost priority. The automated classification of sentiment based classification can be leveraged in several applications which need an automated mechanism for sentiment classification. The major challenge in this section is the proper training of the automated system as the training accuracy would yield high classification accuracy later [6].

II. CONTEXTUAL ANALYSIS AND DEEP LEARNING

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 09 ISSUE: 02 | FEB - 2025 SJIF RATING: 8.448 ISSN: 2582-3930

One of the major challenges in sentiment analysis is the contextual analysis of data. The different aspects are discussed subsequently [7].

2.1 Contextual Analysis

It is often difficult to estimate the context in which the statements are made. Words in textual data such as tweets can be used in different contexts leading to completely divergent meaning [8].

2.2 Frequency Analysis

Often words in textual data (for example tweets) are repeated such as

##I feel so so so happy today!!

In this case, the repetition of the word is used to emphasize upon the importance of the word. In other words, it increases to its weight. However, such rules are not explicit and do not follow any regular mathematical formulation because of which it is often difficult to get to the actuality of the tweet [9].

2.3 Converting textual data into numerically weighted data

The biggest challenge in using an ANN based classifier is the fact that the any ANN structure with a training algorithm doesn't work upon textual data directly to find some pattern. It needs to be fed with numerical substitutes [7]. Hence it becomes mandatory to replace the textual information with numerical information so as to facilitate the learning process of the neural network [10]

the machine or artificial intelligence system requires training for the given categories [11]. Subsequently, the neural network model needs to act as an effective classifier. The major challenges here the fact that sentiment relevant data vary significantly in their parameter values due to the fact that the parameters for each building is different and hence it becomes extremely difficult for the designed neural network to find a relation among such highly fluctuating parameters. Generally, the Artificial Neural Networks model's accuracy depends on the training phase to solve new problems, since the Artificial Neural Networks is an information processing paradigm that learns from its environment to adjust its weights through an iterative process [12].

Deep learning models do have the capability to extract meaning form large and verbose datasets by finding patterns between the inputs and targets. Since neural nets directly process numeric data sets, the processing of data is done prior to training a neural network [13]. The texts are first split into training and testing data samples in the ratio of 70:30 for training and testing. Further, a data vector containing known and commonly repeated spam and ham words is prepared [4]. Text normalization is followed by removal of special characters and punctuation marks.

Subsequently the data set structuring and preparation is performed based on the feature selection. The deep learning structure is depicted in figure 2 [15].

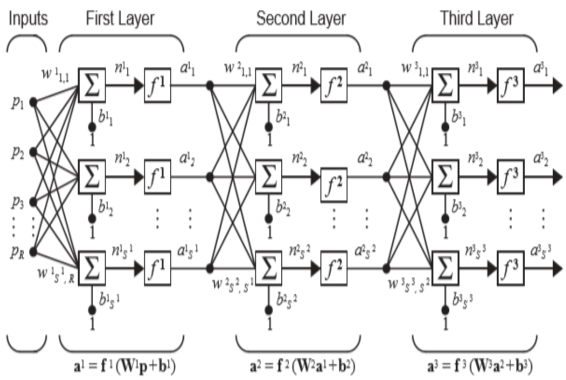

Fig.2 The deep learning structure

The deep learning structure is depicted in figure 2 and it is basically a cascade of stacked neural networks [14]. Multiple hidden layers facilitate the analysis of complex data. The cascading weight updating can be understood as [15]:

$$a^n = \varphi_n(\varphi_{n-1} \dots \dots \varphi_1\{wp + b\}) \qquad (1)$$

Here,
W is the weight
b is the bias
a is the input to the final nth layer
ϕ is the activation function

III. PROPOSED ALGORITHM

The proposed approach is mathematically modelled as:

The prepared data vector for training is used for training wherein the weights are initialized randomly. A stepwise implementation is done as [16]:

1. Prepare two arrays, one is input and hidden unit and the second is output unit.

Here, a two dimensional array W_{ij} is used as the weigt updating vector and output is a one dimensional array Y_i.

3. Original weights are random values put inside the arrays after that the output [17].

$$x_j = \sum_{i=0} y_i W_{ij} \qquad (2)$$

Where,

y_i is the activity level of the j[th] unit in the previous layer and
W_{ij} is the weightof the connection between the i[th] and the j[th] unit.
4. Next, activation is invoked by the sigmoid function applied to the total weighted input [18].

$$y_i = \left[\frac{e^x - e^{-x}}{e^x + e^{-x}}\right] \quad (3)$$

Summing all the output units have been determined, the network calculates the error (E).

$$E = \frac{1}{2}\sum_i (y_i - d_i)^2 \quad (4)$$

Where, y_i is the event level of the j[th] unit in the top layer and d_i is the preferred output of the j_i unit [19].

A. Implementing Back Prop:

Calculation of error for the back propagation algorithm is as follows:
Error Derivative (EA_j) is the modification among the real and desired target:

$$EA_j = \frac{\partial E}{\partial y_j} = y_j - d_j \quad (5)$$

Here,
E represents the error
y represents the Target vector
d represents the predicted output

Error Variations is total input received by an output changed given by:

$$EI_j = \frac{\partial E}{\partial X_j} = \frac{\partial E}{\partial y_j} X \frac{dy_j}{dx_j} = EA_j y_j (1 - y_i) \quad (6)$$

Here,
E is the error vector
X is the input vector for training the neural network
In Error Fluctuations calculation connection into output unit is computed as [20]:

$$EW_{ij} = \frac{\partial E}{\partial W_{ij}} = \frac{\partial E}{\partial X_j} = \frac{\partial X_j}{\partial W_{ij}} = EI_j y_i \quad (7)$$

Here,
W represents the weights
I represents the Identity matrix
I and j represent the two dimensional weight vector indices
Overall Influence of the error:

$$EA_i = \frac{\partial E}{\partial y_i} = \sum_j \frac{\partial E}{\partial x_j} X \frac{\partial x_j}{\partial y_i} = \sum_j EI_j W_{ij} \quad (8)$$

The partial derivative of the Error with respect to the weight represents the error swing for the system while training. The gradient is computed as:

$$g = \frac{\partial e}{\partial w} \quad (9)$$

Here,
g represents the gradient
e represents the error of each iteration
w represents the weights.

The gradient is considered as the objective function to be reduced in each iteration. A probabilistic classification using the Bayes theorem of conditional probability is given by [21]:

$$P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right)P(H)}{P(X)} \quad (10)$$

Here,
Posterior Probability [P (H/X)] is the probability of occurrence of event H when X has already occurred
Prior Probability [P (H)] is the individual probability of event H
X is termed as the tuple and H is is termed as the hypothesis.
Here, [P (H/X)] denotes the probability of occurrence of event X when H has already occurred.
The final classification accuracy is computed as:

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

Here.
TP represents true positive
TN represents true negative
FP represents false positive
FN represents false negative

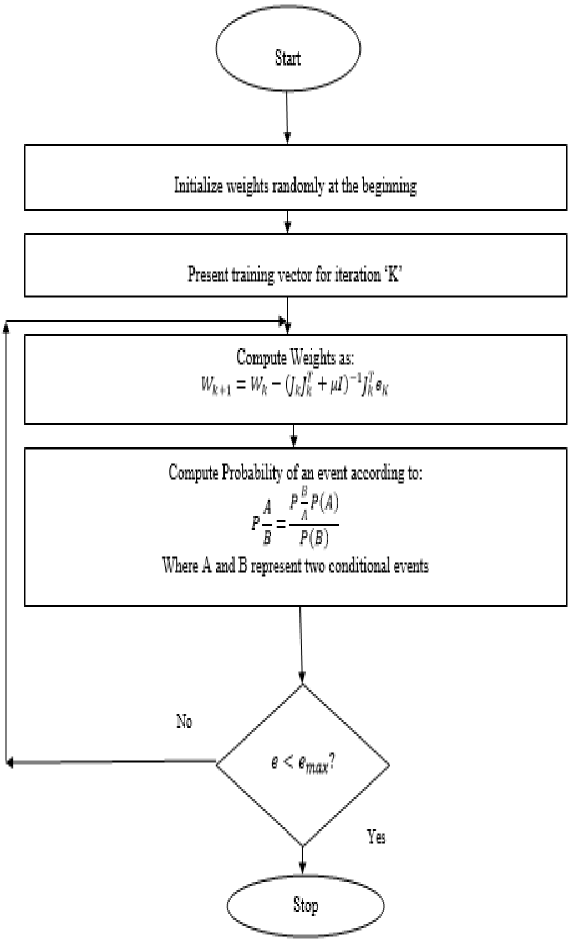

Fig.3 Flowchart for training

IV. RESULTS

The proposed system utilizes the textual data in the form of tweets to be analyzed based on positive, negative and neutral tokens to be represented by -1, 0 and 1 respectively. Subsequently, the number of tokens with polarity is also fed to the neural network as a training parameter.

09877

1	can wait me I'm ground trying get gate after were moved crap
2	hate Time Warner So wish had Vios Cant watch fricken Mets game w/o buffering feel like im watching free internet porn
3	Oh sure it's not planned but occurs absolutely consistently it's usually only flight that's Cancelled Flightled daily
4	Tom Shanahan's latest column Baseball Regional
5	Found self driving car
6	arrived YYZ take our flight Taiwan Reservation missing our ticket numbers Slow agent Sukhdeep caused us miss our flt
7	Driverless cars ? What's point
8	how can not love Obama? makes jokes about himself
9	Safeway very rock n roll tonight
10	RT Ultimate jQuery List
11	saw Night Museum Battle Swithsonian today okay Your typical [kids] Ben Stiller movie
12	History exam studying ugh
13	Missed this each newer generation' I'd start allegra go claritin zyrtec don't envy you
14	being fucked by time warner cable didnt know modems could explode Susan Boyle sucks too
15	hope girl work buys my
16	good luck
17	needs someone explain lambda calculus him
18	yeah looks like only fucking me yeah my
19	Loves twitter
20	really dont want phone servicethey suck when comes having signal
21	Thank Margo Houston's Bush Intercontinental getting me home earlier
22	don want either RT might get pilotless planes before driverless cars
23	Super cool
24	DITTO not good Nirvana Sandwiches
25	waiting line safeway
26	OMG would died actually no take back I keep updated version my Xdrive it's all good
27	There's google self-driving car parked next me Shall ask ride?

Fig.4 Sentiment Data

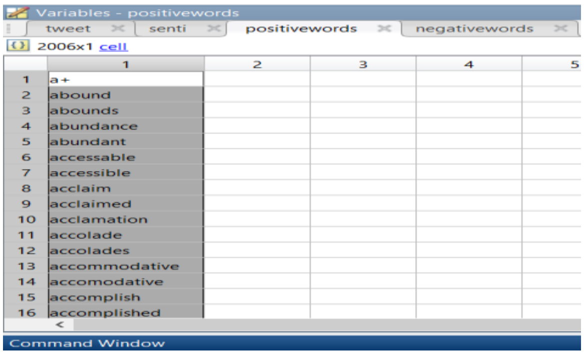

Fig. 5 Positive Tokens

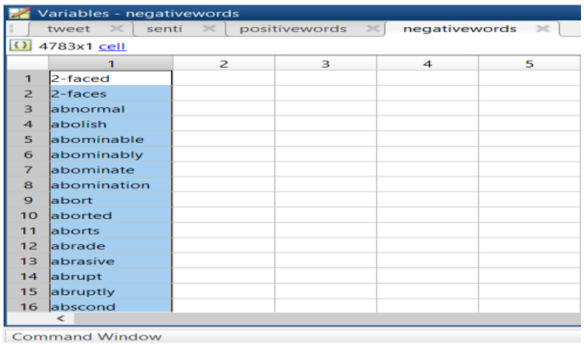

Fig. 6 Negative Tokens

Figure 5 and 6 depict the positive and negative tokens to train the Bayesian Model presented next.

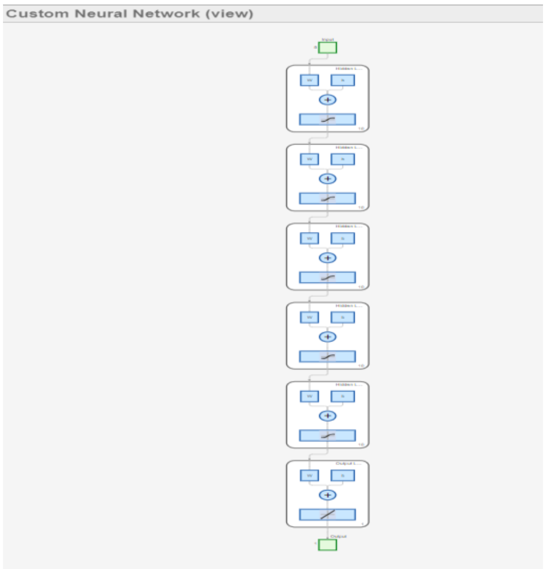

Fig.7 Deep Net Parameters

Fig.8 MSE Variation

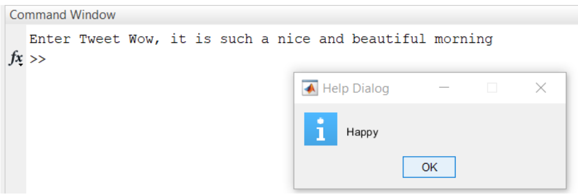

Fig.9 GUI for classification (happy)

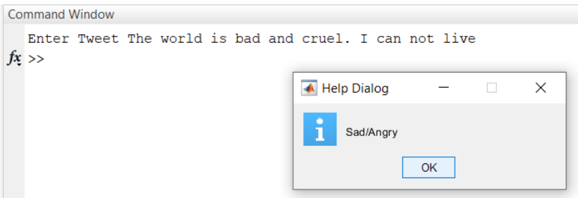

Fig.10 GUI for classification (sad)

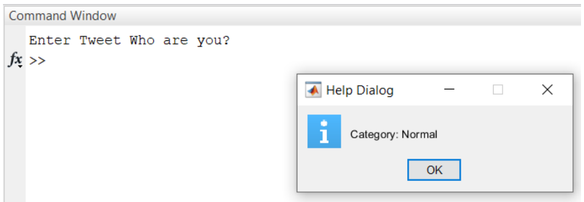

Fig.9 GUI for classification (neutral/normal)

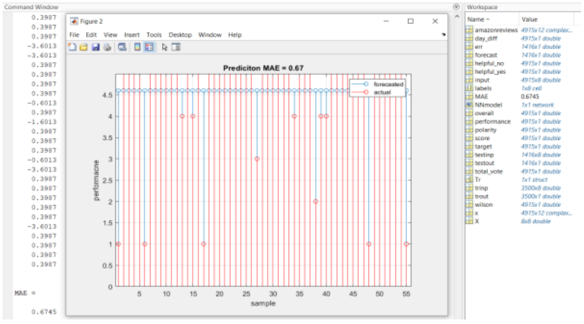

Fig.11 Obtained MAE

The proposed system parameters can be summarized in table 1.

Table 1. Summary of Results

Parameter	**Value**
ML category	Bayesian Net
No. of hidden layers	5
Iterations	11
MAE	0.67
Accuracy (Proposed Work)	99.3% (APPROX)
Accuracy (Previous Work, [1])	93.5%

CONCLUSION

Sentiment analysis has a wide range of applications in information systems, including classifying reviews, summarizing review and other real time applications. There are promising possibilities to use sentiment analysis in real time business models. The present work focuses on sentient analysis by classification of tweets from social media (twitter) data. In this case, a deep Bayes net has been employed for training and subsequent testing of tweets. The Bayesian Regularization (BR) algorithm has been used and their respective results show variation in outcomes because of its probabilistic classification. It has been shown that the proposed algorithm attains an MAE value of 0.67 which is substantially lesser than previously existing method.

REFERENCES

[1] Y Zhao, M Mamat, A Aysa, K Ubul, "Multimodal sentiment system and method based on CRNN-SVM", Neural Computing and Applications, Springer, 2023, pp.1-13.
[2] M Dhyani, GS Kushwaha, S Kumar, "A novel intuitionistic fuzzy inference system for sentiment analysis", International Journal of Information Technology, Springer 2022, vol.14., pp. 3193–3200.
[3] A Vohra, R Garg, "Deep learning based sentiment analysis of public perception of working from home through tweets", Journal of Intelligent Information Systems, Springer 2022, vol.60, pp. 255–274.
[4] H. T. Phan, N. T. Nguyen and D. Hwang, "Aspect-Level Sentiment Analysis Using CNN Over BERT-GCN," in IEEE Access, 2022, vol. 10, pp. 110402-110409.
[5] R. Obiedat R. Qaddoura, A. Al-Zoubi, L. Al-Qaisi, O. Harfoushi, M. Alrefai, H. Faris., "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," in IEEE Access, vol. 10, pp. 22260-22273, 2022.
[6] S Vashishtha, S Susan, "Neuro-fuzzy network incorporating multiple lexicons for social sentiment analysis", Applications in computing, Springer 2022, vol.26, pp. 487–4507.
[7] A. Saha, A. A. Marouf and R. Hossain, "Sentiment Analysis from Depression-Related User-Generated Contents from Social Media," 2021 8th Intern, "ational Conference on Computer and Communication Engineering (ICCCE), Kuala Lumpur, Malaysia, 2021, pp. 259-264
[8] MLB Estrada, RZ Cabada, RO Bustillos, "Opinion mining and emotion recognition applied to learning environments", Journal of Expert Systems, Elsevier 2020, vol. 150., 113265
[9] A. M. Rahat, A. Kahir and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2019, pp. 266-270.
[10] H. Hasanli and S. Rustamov, "Sentiment Analysis of Azerbaijani twits Using Logistic Regression, Naive Bayes and SVM," 2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, 2019, pp. 1-7.
[11] R. B. Shamantha, S. M. Shetty and P. Rai, "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 21-25.
[12] M. Yasen and S. Tedmori, "Movies Reviews Sentiment Analysis and Classification," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 860-865.
[13] P. Karthika, R. Murugeswari and R. Manoranjithem, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm," 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 2019, pp. 1-5.
[14] L Zheng, H Wang, S Gao," Sentimental feature selection for sentiment analysis of Chinese online reviews", International journal of machine learning and cybernetics, Springer, 2018, vol.9, pp. 75–84.
[15] R. D. Desai, "Sentiment Analysis of Twitter Data," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 114-117.
[16] A. Bayhaqy, S. Sfenrianto, K. Nainggolan and E. R. Kaburuan, "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes," 2018 International Conference on Orange Technologies (ICOT), Nusa Dua, Bali, Indonesia, 2018, pp. 1-6.
[17] L. Wang, M. Han, X. Li, N. Zhang and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," in IEEE Access, vol. 9, pp. 64606-64628, 2021.
[18] G. Karatas, O. Demir and O. K. Sahingoz, "Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset," in IEEE Access, 2020, vol. 8, pp. 32150-32162.
[19] D. Dablain, B. Krawczyk and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," in IEEE Transactions on Neural Networks and Learning Systems, 2023, vol. 34, no. 9, pp. 6390-6404.
[20] M. I. Zul, F. Yulia and D. Nurmalasari, "Social Media Sentiment Analysis Using K-Means and Naïve Bayes Algorithm," 2018 2nd International Conference on Electrical Engineering and Informatics (ICon EEI), Batam, Indonesia, 2018, pp. 24-29.
[21] M. I. Zul, F. Yulia and D. Nurmalasari, "Social Media Sentiment Analysis Using K-Means and Naïve Bayes Algorithm," 2018 2nd International Conference on Electrical Engineering and Informatics (ICon EEI), Batam, Indonesia, 2018, pp. 24-29.