

A Data Driven Approach for Sentiment Analysis of Twitter Data

Priyanka Yadav¹ Prof. Ashish Tiwari²

Abstract— Of late, big data and big data analytics has found 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 accuracy of 96.90%, thereby outperforming existing work in the domain

Keywords—Sentiment Analysis, Twitter Data, Bayesian Regularization, Classification Accuracy.

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. [5] 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. 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. 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.



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 [3]. 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 [4].

II. CONTEXTUAL ANALYSIS AND DEEP LEARNING

One of the major challenges in sentiment analysis is the contextual analysis of data. The different aspects are discussed subsequently [5].

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 meanings.

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 [6].

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. Hence it becomes mandatory to replace the textual information with numerical information so as to facilitate the learning process of the neural network [7].

the machine or artificial intelligence system requires training for the given categories [8]. 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 [9].

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. 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. The SMS spam collection v.1 dataset is used as a dataset for the proposed work. 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.

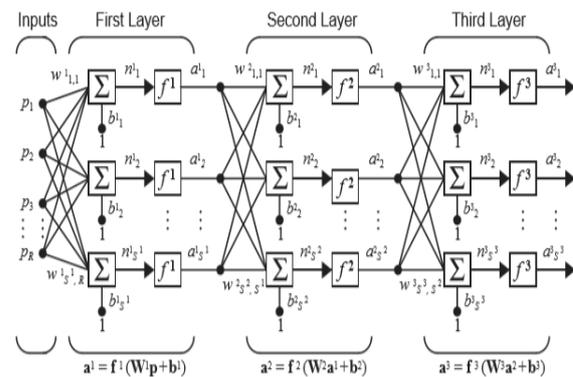


Fig.2 The deep learning structure

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

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

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

The output is given by:

$$y = f(\sum_{i=1}^n x_i w_i + \theta) \quad (2)$$

III. PROPOSED ALGORITHM

The proposed approach is mathematically modelled as:

Let the universal set of all data sample be represented by U. Thus any sample of the data class would be a subset of the universal set A given by:

$$A \in U \nabla U \quad (3)$$

Here,
A is the data sample set

U is the universal set

The overlapping group classification occurs in case of the data sample belonging to both data categories of A and B with some overlapping attributes. In such cases, the decision has to be made based on a probabilistic approach by the classifier and the classification is prone to errors. The decision becomes more complex with increasing number of attributes termed as features or parameters. The decision in such an overlapping attributed class can be done using the Bayes' theorem which is invoked by the Probabilistic Neural Network as:

$$P\left(\frac{C1}{C2}\right) = \frac{P\left(\frac{C2}{C1}\right)P(C1)}{P(C2)} \quad (4)$$

Here:

$P\left(\frac{C1}{C2}\right)$ represents the probability of occurrence of C1 given C2 is true

$Prob\left(\frac{N}{M}\right)$ represents the probability of occurrence of C2 given C1 is true

$Prob(C1)$ represents the individual occurrence probability of the event C1

$Prob(C2)$ represents the individual occurrence probability of the event C2

The weight updating strategy for the probabilistic classifier can follow the backpropagation based gradient descent approach given mathematically as:

$$W_{n+1} = W_n - \left\{ (J_n^T J_n + cI)^{-1} \right\} J_n^T e_n \quad (5)$$

Where,

W_n is weight of nth round or iteration of weight update for the network

W_{n+1} is weight of (n+1)st round or iteration of weight update for the network

e_n represents the error of the nth round or I represents an identity matrix

c represents the combination co-efficient given by $c = w_{n+1} - w_n$, which gives the magnitude of weight change per iteration at any given iteration n

J_n is the Jacobian matrix given by:

$$J_n = \begin{pmatrix} \frac{\partial^2 e_1}{\partial w_1^2} & \dots & \frac{\partial^2 e_n}{\partial w_1^2} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 e_m}{\partial w_n^2} & \dots & \frac{\partial^2 e_m}{\partial w_n^2} \end{pmatrix} \quad (6)$$

J_n^T is the transpose of the Jacobian Matrix.

The Jacobian matrix is essentially the second order rate of change of the errors of the network w.r.t. to the weights of the networks.

Typically, the classification is done based on the maximum posteriori probability of the data sample given by:

Let there be 'N' classes of data sets available in the sample space 'U'.

The conditional probabilities of N classes for the universal set U is given by:

$$P\left(\frac{A}{U}\right), P\left(\frac{B}{U}\right), \dots, P\left(\frac{N}{U}\right). \quad (7)$$

The maximum magnitude of the probability of a particular class is found as:

$$P(\max) = \begin{matrix} P\left(\frac{A}{U}\right) \\ P\left(\frac{B}{U}\right) \\ \vdots \\ \vdots \\ P\left(\frac{N}{U}\right) \end{matrix} \quad (8)$$

The maximum probability decides the class of the new data sample.

The final classification accuracy is computed as:

$$Ac = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

Here.

TP represents true positive

TN represents true negative

FP represents false positive

FN represents false negative

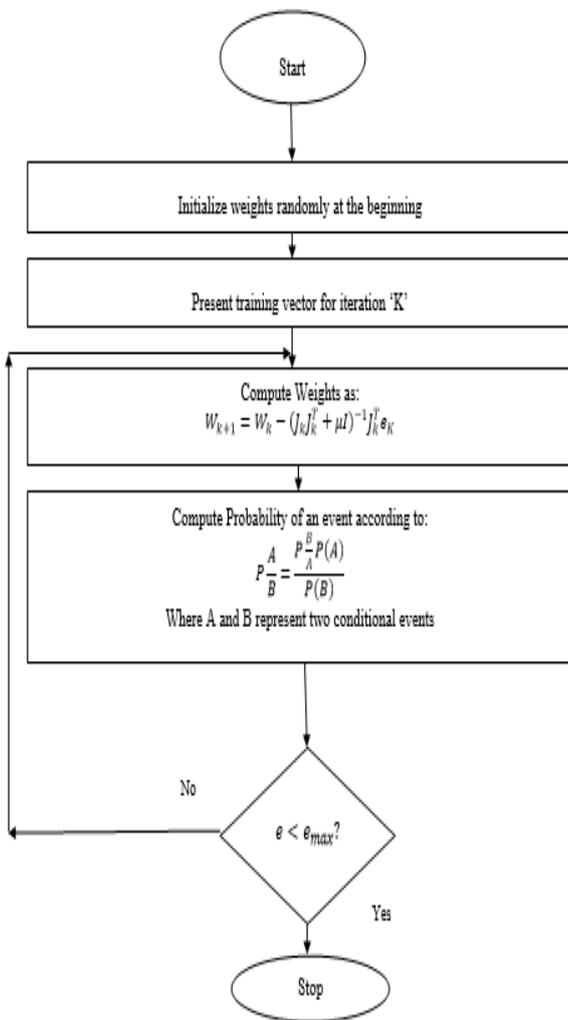


Fig.3 Flowchart for training

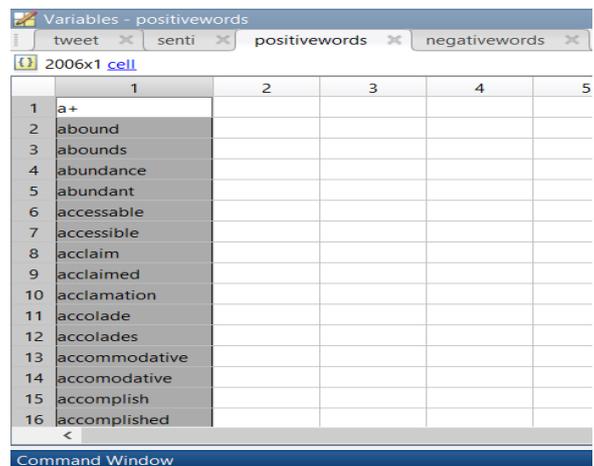
IV. RESULTS

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.

```

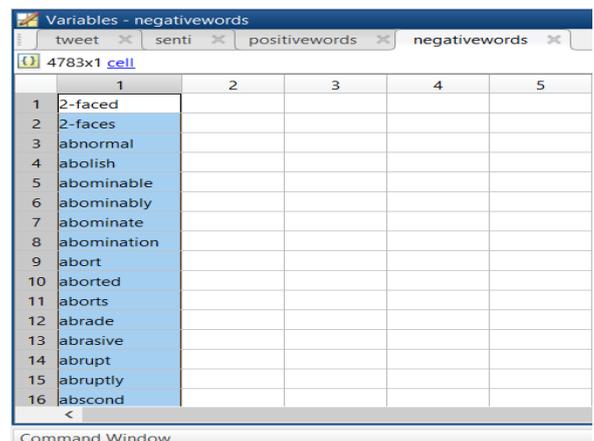
1 can wait me I'm ground trying get gate after were moved crap
2 hate Time Warner So wish had Vios Cant watch fricken Mets game w/o buffering feel like im watching free internet porn
3 Oh sure it's not planned but occurs absolutely consistently it's usually only flight that's Cancelled Flightlded daily
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 fit
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



1	2	3	4	5
1	a+			
2	abound			
3	abounds			
4	abundance			
5	abundant			
6	accessable			
7	accessible			
8	acclaim			
9	acclaimed			
10	acclamation			
11	accolade			
12	accolades			
13	accommodative			
14	acomodative			
15	accomplish			
16	accomplished			

Fig. 5 Positive Tokens



1	2	3	4	5
1	2-faced			
2	2-faces			
3	abnormal			
4	abolish			
5	abominable			
6	abominably			
7	abominate			
8	abomination			
9	abort			
10	aborted			
11	aborts			
12	abrade			
13	abrasive			
14	abrupt			
15	abruptly			
16	abscond			

Fig. 6 Negative Tokens

The Deep Net with 5 hidden layers is designed. The number of layers is limited to five to reduce the time complexity.

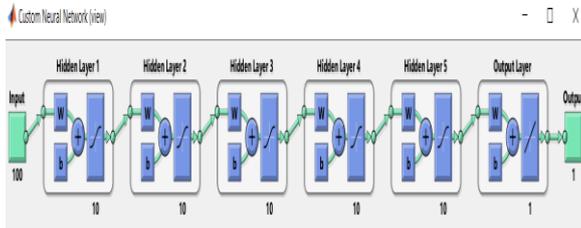


Fig. 7 Designed Deep Net

Neural Network

Algorithms

- Data Division: Random (dividerand)
- Training: Bayesian Regularization (trainbr)
- Performance: Mean Squared Error (mse)
- Calculations: MEX

Progress

Epoch: 0 18 iterations 1000

Time: 0:00:32

Performance: 6.55 3.10 0.00

Fig. 8 Deep Net Parameters

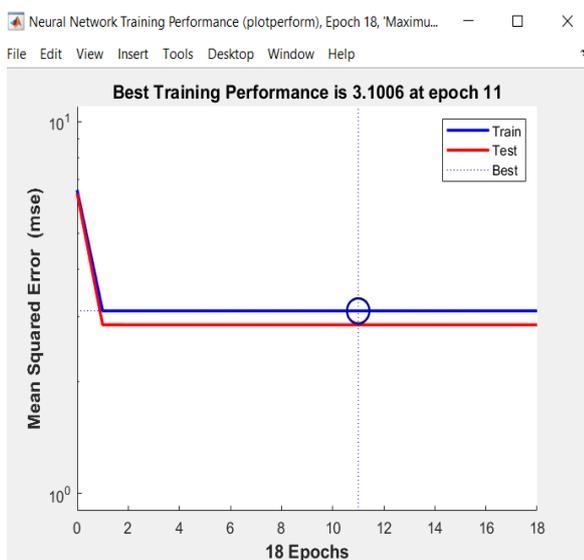


Fig. 9 MSE Variation

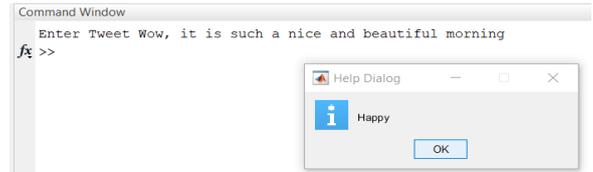


Fig. 10 GUI for classification (happy)



Fig. 11 GUI for classification (sad)



Fig. 12 GUI for classification (neutral/normal)

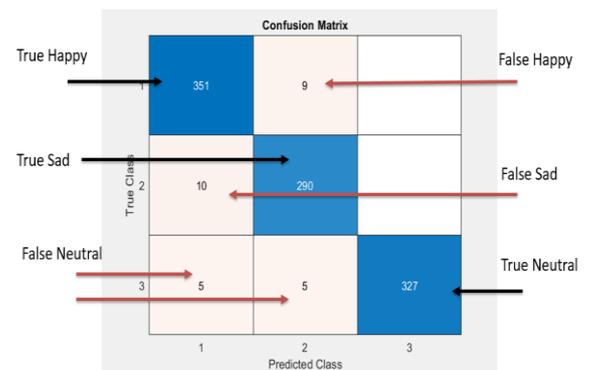


Fig. 13 Confusion Matrix

The proposed system parameters can be summarized in table 1.

Table 1. Summary of Results

Parameter	Value
ML category	Deep Supervised Learning
No. of hidden layers	5
Iterations	18
Error	3.1%
Accuracy (Proposed Work)	96.90
Accuracy (Previous Work)	89%

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 sentiment 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 a classification accuracy of 96.90% which is substantially higher than previously existing method.

REFERENCES

- [1] R. Obiedat et al., "Sentiment Analysis of Customers' Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," in *IEEE Access*, 2022, vol. 10, pp. 22260-22273.
- [2] H. A. Shehu et al., "Deep Sentiment Analysis: A Case Study on Stemmed Turkish Twitter Data," in *IEEE Access*, 2021, vol. 9, pp. 56836-56854.
- [3] MLB Estrada, RZ Cabada, RO Bustillos, "Opinion mining and emotion recognition applied to learning environments", *Journal of Expert Systems*, Elsevier 2020.
- [4] R. Wang, D. Zhou, M. Jiang, J. Si and Y. Yang, "A Survey on Opinion Mining: From Stance to Product Aspect," in *IEEE Access*, vol. 7, pp. 41101-41124, 2019
- [5] Lijuan Zheng, Hongwei Wang, Song Gao, " Sentimental feature selection for sentiment analysis of Chinese online reviews", *SPRINGER* 2018
- [6] Jitendra Kumar Rout, Kim-Kwang Raymond Choo, Amiya Kumar Dash, Sambit Bakshi, Sanjay Kumar Jena, Karen L. Williams, " A model for sentiment and emotion analysis of unstructured social media text", *SPRINGER* 2018
- [7] Asha S Manek, P Deepa Shenoy, M Chandra Mohan, Venugopal K R, "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier", *SPRINGER* 2017
- [8] Md Rakibul Islam ; Minhaz F. Zibran, , "Leveraging Automated Sentiment Analysis in Software Engineering", *IEEE* 2017
- [9] Hassan Saif, Yulan He, Miriam Fernandez, Harith Alani, "Contextual Semantics for Sentiment Analysis for Twitter", *Elsevier* 2016
- [10] Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, Ming Zhou, " Sentiment Embeddings with Applications to Sentiment Analysis", *IEEE* 2016
- [11] Xing Fang, Justin Zhan, " Sentiment analysis using product review data", *SPRINGER* 2015
- [12] Basant Agarwal, Soujanya Poria, Namita Mittal, Alexander Gelbukh, Amir Hussain, "Concept-Level Sentiment Analysis with Dependency-Based Semantic Parsing: A Novel Approach", *SPRINGER* 2015
- [13] Emitza Guzman ; Walid Maalej, " How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews", *IEEE* 2014
- [14] Geetika Gautam ; Divakar Yadav, " Sentiment analysis of twitter data using machine learning approaches and semantic analysis", *IEEE* 2014
- [15] Erik Cambria, Björn Schuller, Yunqing Xia, Catherine Havasi, " New Avenues in Opinion Mining and Sentiment Analysis", *IEEE* 2013
- [16] Erik Cambria, Björn Schuller, Bing Liu, Haixun Wang, Catherine Havasi, " Knowledge-Based Approaches to Concept-Level Sentiment Analysis", *IEEE* 2013
- [17] Hassan Saif, Yulan He, Harith Alani, " Semantic Sentiment Analysis of Twitter", *SPRINGER* 2012
- [18] Bing Liu, Lei Zhang, "A Survey of Opinion Mining and Sentiment Analysis", *SPRINGER* 2012
- [19] Alena Neviarouskaya, Helmut Prendinger, Mitsuru Ishizuka, "Secure SentiFul: A Lexicon for Sentiment Analysis", *IEEE* 2011
- [20] Jorge Carrillo de Albornoz, Laura Plaza, Pablo Gervás, Alberto Díaz, "A Joint Model of Feature Mining and Sentiment Analysis for Product Review Ratings", *SPRINGER* 2011
- [21] Gang Li, Fei Liu, "A clustering-based approach on sentiment analysis", *IEEE* 2010
- [22] Wei Wang, "Sentiment analysis of online product reviews with Semi-supervised topic sentiment mixture model", *IEEE* 2010
- [23] Erik Boiy, Marie-Francine Moens, "A machine learning approach to sentiment analysis in multilingual Web texts", *SPRINGER* 2009
- [24] Songbo Tan, Xueqi Cheng, Yuefen Wang, Hongbo Xu, "Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis", *SPRINGER* 2009