Efficient Extraction of Insights from large scale Social Media Data through Distributed Deep Learning

[1]*Dr. K. Syed KousarNiasi
Assistant Professor, Department of Computer Science, Jamal Mohamed College (Affiliated To Bharathidasan University), Tiruchirappalli-620020.Tamilnadu, India.
Mail id: skn@jmc.edu
[2]Dr.J.Jaganpradeep
Professor, Department of Electronics & Communication Engineering, SSM College of Engineering, Komarapalayam, Tamilnadu-638183
Mail id: jgnprdp@gmail.com
[3]Prakash K
Assistant Professor, Department of Mathematics,Bannari Amman Institute of Technology,Sathyamangalam - 638 401, Erode, Tamil Nadu, India.
Mail id: prakashk@bitsathy.ac.in
[4]Dr. M. Balamurugan
Head of the Department & Assistant Professor, Department of Computer Science and Engineering, The Kavery Engineering College, Mecheri, Salem
Mail id: hodcse@kavery.org.in

Abstract:

The requirement for effective pattern and insight extraction has grown critical due to the exponential growth of social media data. The authors of this study offer a fresh strategy to overcome this difficulty by using deep learning methods on massive amounts of social media data stored in different places. To get valuable insights into user behaviour, preferences, and social interactions, the primary goal is to facilitate the fast and accurate detection of relevant patterns, feelings, and trends. The proposed system integrates distributed computing with deep learning approaches to handle and analyse massive volumes of social media data concurrently. Leveraging the power of distributed systems significantly improves both scalability and processing speed, enabling real-time or near-real-time analysis of dynamic social media material. Unstructured text, video, and user interactions are only some of how social media data differs from traditional datasets, and the deep learning models are altered and improved to manage them. This research presents an approach to evaluating a CNN model's interpretability. The suggested Dolphin Echolocation Algorithm (DEA) was added during the feature selection phase and used to fine-tune the CNN's filter weights. Through a process of backtracking analysis on model prediction results, the approach we present can conduct multi-angle analysis on the discriminant outcomes of multi-classified text and multi-label classification tasks. Data diversity, volume, and velocity challenges are typical in large-scale social media datasets, and this study helps address those issues. The suggested method also strives to use as few resources as possible, which reduces costs and helps the environment both of which are crucial when dealing with distributed databases of any size. Extensive tests are performed on various datasets obtained from major social media sites to validate the efficacy and efficiency of the proposed framework. The outcomes show that our strategy is more precise, faster, and scalable than the status quo.

Keywords: Deep learning, Distributed systems, Insights, Massive data, Patterns, and social media data.

1) Introduction

The widespread use of social media platforms and the constant accessibility of internet connectivity have resulted in a significant number of individuals, exceeding 2 billion users, accounting for approximately 35% of the global population, engaging with social media. Social data analysis presents a wide range of difficulties and possibilities for investigators in the field of natural language processing (NLP). This includes a vast amount of data consisting of tweets, websites, and analyses from various domains. Exploring this data can lead to the discovery of valuable information, but it also requires researchers to address several obstacles. Furthermore, the availability of such data allows for the analysis of individuals' perspectives on a particular subject. It presents valuable insights that can be utilised for predictive purposes in various domains such as product sales, stock market trends, elections for politicians, and even more. Sentiment analysis is vital in business studies, as it enables timely decision-making based on people's reviews. This research domain is highly sought-after due to its ability to provide valuable insights for making informed decisions. Sentiment analysis, also known as opinion mining, is a research field that focuses on analysing individuals' sentiments towards various entities such as goods and services, requirements, businesses, and more [1]. Previous research studies have extensively explored the application of sentiment analysis across different levels of granularity, including document, phrase, dimension, subject, and more. These studies have investigated various approaches and techniques to analyse sentiments in textual data. The proposed methods in the research primarily focus on utilising a single modality, specifically textual data, to infer sentiment.

Social network (SN) sites have emerged as a versatile and activestage increasingly being leveraged for various purposes [2]. However, it is essential to acknowledge that these platforms are not immune to misuse, as they have also become a breeding ground for illegal activities. In general, individuals utilise SN platforms to engage in social interactions with individuals who share similar interests and professional connections. Moreover, this tool is commonly employed to communicate with customers, and its data can be precious for identifying emerging trends in business analytics.The potential of social media platforms massive amounts of data as a tool for understanding user behaviour was investigated in a prior study [3].

[1]B. Liu, Sentiment analysis and opinion mining, Synth. Lect. Hum. Lang. Technol. 2012.
[2]Naseem et al., Transformer-based deep intelligent contextual embedding for Twitter sentiment analysis.
[3]Tufekci et al., Big questions for social media big data: representativeness, validity and other methodological pitfalls. In: Eighth international AAAI conference on weblogs and social media. 2014.

Since this issue has received so much focus from academics, many have analysed and dissected it at length. The authors of this study [4] investigated the viability of using big data analytics and machine learning algorithms to track online conversations about high-end hotels and deduce guests' opinions of those establishments from those discussions. They have zeroed in on improving brand management tactics for high-end hotels by using cutting-edge visual data analysis methods.

A dataset of 8434 startups was obtained from Twitter and used in a recent study [5]. The authors built a machine-learning model using attributes derived from social media data. The primary purpose of this approach was to foretell each startup's degree of involvement on social media. This study's results show that deep learning models can anticipate users' interests with more precision than other methods. According to this research, a company's social media marketing success may be gauged by the number of tweets, retweets, and likes it receives. The purpose of this study is to use big data analytics to look at how customers feel about and interact with social media.

This research addresses the pressing need for effective pattern and insight extraction from large-scale social media data by proposing a fresh and innovative strategy that leverages the power of deep learning methods in conjunction with distributed computing. The primary objective is to facilitate the rapid and accurate detection of relevant patterns, sentiments, and trends within the vast expanse of social media content. Harnessingdistributed systems' capabilities enhances scalability and significantly improves processing speed, enabling real-time or near-real-time analysis of dynamic social media material. The diversity of social media data, including unstructured text, images, and user interactions, necessitates adapting and optimising deep learning models to extract patterns and insights effectively.

A key focus of this research is the evaluation of interpretability in Convolutional Neural Network (CNN) models. During the feature selection phase, we propose the Dolphin Echolocation Algorithm (DEA) and demonstrate its effectiveness in fine-tuning the CNN's filter weights. Our approach enables multi-angle analysis of discriminant outcomes for multi-classified text and multi-label classification tasks through backtracking analysis on model prediction results.

1) Related Works

The authors of this study [6] investigate what makes online customer evaluations exciting and valuable to their audience. Within big data analytics, they use a technique called "sentiment mining." According to this research, longer and more established online customer evaluations tend to attract more readers and be rated as more valuable by those readers.

_____-

[4]Giglio S, Pantano E, Bilotta E, Melewar TC. Branding luxury hotels: evidence from analysing consumers' 2020.
[5]Jung SH, Jeong YJ. Twitter data analytical methodology development for prediction of start-up frms' social media marketing level. 2020.

Unstructured data, which accounts for at least 95% of extensive data, is the main topic of prominent research [7], emphasising the importance of big data analytics in this context. The researchers have examined the several analytics methods used for handling data in sound, video, text, and social media. Also, they've helped a lot by developing brand-new predictive analytics tools and techniques that work well with structured data. It's important to remember that noise, illegibility, and dependencies are commonplace with massive data.

In this research, we look at how consumers' preferences affect our ability to foresee their actions and choices in the store. Businesses may improve their marketing efforts by learning how their target audience perceives a brand and using that information to create more relevant ads. Predicting a user's character from their social media activity was proposed in a prior study [9]. A recommender framework that considers personality was created to better understand how consumers' traits affect their reactions to product suggestions. This paper's structure is built on the five-factor theory of personality, which is discussed in references [10, 11]. The authors [12]of this study dive into the aspects that affect how people feel about a brand and examine how consumers' feelings about the brand are formed. The study aims to learn more about the causes of people's good and negative emotions and how to better respond to them.

A database of 2204 coded tweets is used to investigate brand authenticity and sentiment polarity. The author employs a qualitative analysis of tweets to understand better how people feel about a brand's sincerity. This qualitative research builds the groundwork for a quantitative framework that predicts brand authenticity and reported sentiment's polarity. Quality, dedication, legacy, distinctiveness, and symbolism were criteria for organising the tweets.

In [14], a literature review is conducted on sentiment analysis for predicting epidemics and outbreaks and other application areas by employing several ML approaches over social media data. It examines the pros and cons of using ML, linguistics, and hybrid methods for sentiment analysis.

---

[6]Salehan M, Kim DJ. Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics.
[7]Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J InfManag. 2015.
[9]Buettner R. Predicting user behaviour in electronic markets based on personality-mining in large online social networks. Electron Mark. 2017.
[10]Chu SC, Chen HT, Gan C. Consumers' engagement with corporate social responsibility (CSR) communication in social media: evidence from China and the United States. J Bus Res. 2020.
[11]Costa PT. McCrae RR: Revised NEO Personality Inventory (NEO PIR) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual. Odessa: Psychological Assessment Resources.1992.
[12]Shirdastian et al., M-oO. Using big data analytics to study brand authenticity sentiments: The case of Starbucks on Twitter. Int J Inform Manag. 2019; 48:291–307.

It has been shown that LSA's use improves the accuracy with which one can forecast the authenticity dimension of a brand and identify the polarity of sentiment surrounding that brand. This study [13] analyses the current state of knowledge concerning social media by reviewing the current literature. The study team looked closely at the most up-to-date tools and metrics for measuring social media's impact.

Except for very brief sentences, ML techniques are preferred over linguistic ones.In[15], researchers analysed 85.04M tweets from 182 countries about COVID-19 from March to June 2020. They observed that the distribution of feelings changed over time and between countries, shedding light on how people felt about new rules like social distancing and telecommuting. The authors suggest that analysing social media in non-English languages and platforms is essential for spotting false data and misleading online debates.In[16], the authors assess the outreach impacts of Facebook posts by Public Health Authorities (PHAs). We found that across all PHAs, toxic comments were uncommon and that the average number of daily posts ranged from 1.4 to 5. We also found that the average number of comments per post ranged from 12.5 to 255.3 and that the mean sentiment polarity was positive.In[17], the authors try to figure out whether or not tweets are helpful by analysing the tone and content surrounding major issues like pandemics. Up to 81% accuracy is achieved by the deep learning classifiers used in the model suggested for sentiment analysis. A proposed second model uses fuzzy logic and is executed using SVM to achieve an accuracy of 79%.

In [18], a convolutional neural network (CNN) sentence-based classifier is constructed to place a piece of text into one of six pre-defined emotion classifications (happiness, disgust, rage, shame, and sorrow) based on Ekman's concept. The experimental assessment demonstrates that the deep learning model (CNN) surpasses the three conventional classification algorithms in terms of overall accuracy in the classification of emotions. The precision of text categorisation is enhanced by stemming the text first. The 6,000 Facebook postings written

---

[13]Ghani, NA, et al. Social media big data analytics: A survey. Comput Hum Behav. 2019; 101:417–28.
[14]Singh, R.; Singh, R.; Bhatia, A. Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. Int. J. Adv. Sci. Res. 2018, 3, 19–24.
[15]Sharma, K.; et al., COVID-19 on social media: Analyzing Misinformation in Twitter Conversations. arXiv 2020, arXiv:2003.12309.
[16]SesagiriRaamkumar, et al., Measuring the Outreach Efforts of Public Health Authorities and the Public Response on Facebook During the COVID-19 Pandemic in Early 2020:
[17]Chakraborty, K.; Bhatia, S.; Bhattacharyya, S.; Platos, J.; Bag, R.; Hassanien, A.E. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers.
[18]Skenduli et al., D. User-Emotion Detection Through Sentence-Based Classification Using Deep Learning: A Case-Study with Microblogs in Albanian. In Proceedings of the Foundations of Intelligent Systems, 2018.

by Albanian politicians that were used to train the algorithm are also a significant contribution. In addition, in [19], the authors use clustering to identify sets of generic phrases typical of a particular emotion category, further developing their theory. For an emotion detection model, the authors suggest looking at deep neural network designs [20-23] like LSTM that account for the sequential structure of text input. A domestic abuse detection system based on deep learning was proposed by Subramani et al. [27]. For the suggested system, data was gathered from Facebook. Tocategorise Twitter user-generated material (tweets) into two groups, Ahmad et al. [28].

Combined a deep learning model with a sentiment analysis approach. Budiharto and Meiliana[29] attempted to foretell the outcome of the Indonesian presidential election by employing sentiment analysis methods. Tweets were analysed using sentiment analysis methods by Al Shehhi et al. [30].

A sentiment analysis method for evaluating comments on classroom performance was proposed in [31].Self-attention models (which have obtained state-of-the-art results in several machine translation tasks) were tested for their ability to identify instances of cyberbullying by Pradhan et al. [24].This design swaps the encoding and decoding recurrent layers for a multi-headed self-attention layer.

---

[19]Skenduli et al., Classification and Clustering of Emotive Microblogs in Albanian: Two User-Oriented Tasks, 2020.
[20]Jothimani et al.,THFN: Emotional health recognition of elderly people using a Two-Step Hybrid feature fusion network along with Monte-Carlo dropout. Biomedical Signal Processing and Control.
[23]Jothimani et al., Advanced Deep Learning Techniques with Attention Mechanisms for Acoustic Emotion Classification." 2022.
[27]Subramani et al.,Domestic violence crisis identification from facebook posts based on deep learning. IEEE Access 2018.
[28]Ahmad, S.; Asghar, M.Z.; Alotaibi, F.M.; Awan, I. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques.2019.
[29]Budiharto et al.,Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis.
[30]Al Shehhi et al.,A cross-linguistic Twitter analysis of happiness patterns in the United Arab Emirates.
[31]Pong-inwonget al., Sentiment analysis in teaching evaluations using sentiment phrase pattern matching (SPPM) based on association mining.2018.
[24]Pradhan et al., In Proceedings of the 2020 International Conference Cyber Situational Awareness, Data Analytics and Assessment (CyberSA),2020.

The outcomes obtained using the suggested approach were positive. Existing methods have limitations, such as limiting detection to a single social media platform, narrowing detection to a single type of hate speech, and relying on handcrafted features that traditional algorithms for machine learning provide[35-37]. To get over these restrictions, Agrawal et al. [25]provided a framework and empirically demonstrated its efficacy. The authors explored four distinct types of deep learning systems to circumvent these limitations. The authors further categorised hate speech on social networking sites as either harassment, xenophobia, sexism, or assault. They turned to transfer learning to apply deep learning's insights outside the original dataset. Extensive tests were conducted on Twitter, Wikipedia, and Spring to evaluate the examined architectures. Approximately 16,000 tweets were utilised in the analyses presented in [24, 25].Plaza et al. [26] devised a method to identify cyberbullying in Spanish-language social media. The authors looked at using deep learning to detect hate speech in Spanish. More specifically, the authors trained the deep learning models to increase efficiency, using a transfer learning strategy to deal with minor sample issues.

[35]Kaur et al., A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis.
[37]Jain, Rachna et al.,"Explaining sentiment analysis results on social media texts through visualization." Multimedia Tools and Applications.
[24]Pradhan et al.,Self-attention for cyberbullying detection. In Proceedings of the 2020 International Conference Cyber Situational Awareness, Data Analytics and Assessment (CyberSA).
[25]Agrawal, S.; Awekar, A. Deep learning for detecting cyberbullying across multiple social media platforms. In European Conference on Information Retrieval, 2018.
[26]Plaza-del-Arco, F.M.; Molina-Gonzalez, M.D.; Urena-Lopez, L.A.; Martin-Valdivia, M.T. Comparing pre-trained language models for Spanish hate speech detection. Expert Syst.

2) Methods and Materials

Massive volumes of social media data are collected from various stands, including text, video, and user interactions. Data preprocessing techniques are applied to clean, normalise, and convert the raw data into a suitable format for deep learning analysis. This step helps to handle the challenges associated with data diversity, volume, and velocity in large-scale social media datasets.Deep learning models extract insights from the pre-processed social media data. Traditional deep learning models are altered and improved to accommodate the unique characteristics of social media data, such as unstructured text, video, and multi-label classification tasks. The Fig.1.shows theoverall structure of Extraction of Insights from large-scale Social Media Data.

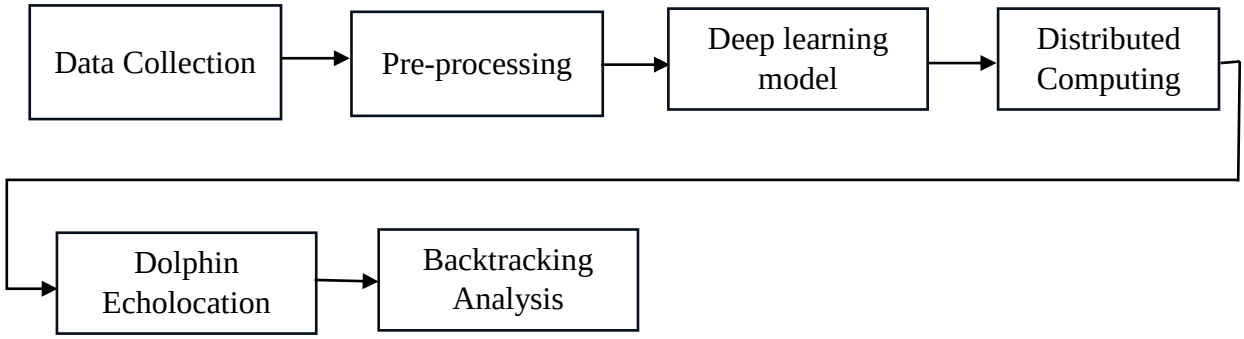


Fig.1. Overall Structure of Extraction of Insights from large scale Social Media Data

Distributed computing techniques are integrated with deep learning approaches to handle and analyse massive social media data concurrently. Leveraging the power of distributed systems enhances scalability and processing speed, enabling real-time or near-real-time analysis of dynamic social media material.The suggested Dolphin Echolocation Algorithm (DEA) is introduced during the feature selection phase. DEA is employed to fine-tune the filter weights of CNN models, enhancing their interpretability and making them more effective in extracting meaningful patterns and trends.The approach involves backtracking analysis on the model prediction results, enabling multi-angle analysis on the discriminant outcomes of multi-classified text and multi-label classification tasks. This step aids in gaining deeper insights into user behaviour and preferences.

3.1) Data Collection

The experiments were conducted using three datasets consisting of tweets: Sentiment140, Tweets Airline, and Tweets SemEval. Sentiment140 [32] stands out as the most extensive, comprising 1.6 million tweets in the realm of tweet datasets. Each tweet in this dataset is meticulously labelled with either a positive or negative sentiment. The Tweets Airline [33] dataset consists of 14,640 tweets, and the Tweets SemEval [34] dataset contains 17,750 tweets. Both datasets encompass tweetslabelled as positive, negative, or neutral in terms of sentiment.

The Sentiment140 dataset comprises anenormousgathering of 1.6 million tweets, serving as a valuable resource for sentiment analysis tasks in the field of NLP. The dataset's structure follows a CSV format, where each tweet is presented as a single line with three distinct columns. Lastly, the third column contains the actual tweet text, enabling researchers and practitioners to analyse and extract insights from the textual content to understand the prevailing sentiments expressed by users on social media platforms. The Sample data distribution from Tweets Airline is shown in Figure 2.

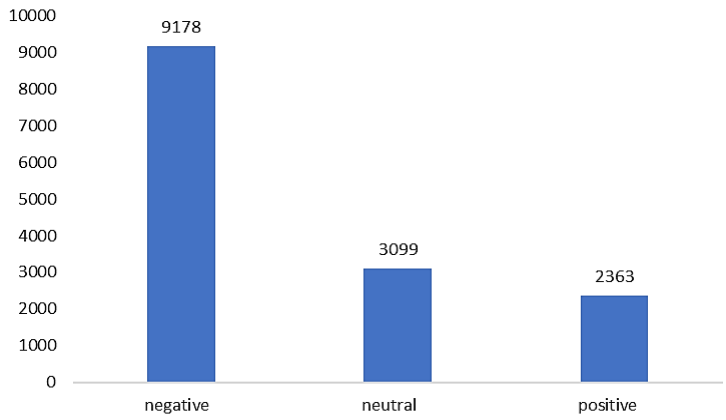


Fig.2. Sample data distribution from Tweets Airline

[31]Sentiment140 - a Twitter sentiment analysis tool,"http://help. Sentiment140.com/site-functionality.
[32]"Twitter US Airline Sentiment," Available from: (accessed on 10 December 2020), https://www.kaggle.com/crowdflower/ twitter-airline-sentiment.
[33]"International Workshop on Semantic Evaluation 2017, Available from: (accessed on 10 December 2020).

3.2) Pre-processing

The data cleansing and preliminary processing phase typically consists of three sub-phases. The raw tweet dataset undergoes a series of procedures to generate the finalised data, as outlined in the preceding dataset. Various noise removal techniques are applied to the text during the initial sub-phase. These techniques include eliminating URLs, removing hashtags and mentions, eliminating punctuation and symbols, and transforming emoticons. During the second sub-phase, Out of Vocabulary Cleansing techniques are applied, including spell checking, acronym development, slang modification, and elongated (repeated characters) removal. Various tweet transformations are performed during the concluding sub-phase, including converting text to lower-case, stemming, word segmentation (tokenisation), and applying stop word filtering. The sub-phases are conducted tooptimise the quality of tweets and enhance the accuracy of feature extraction and classification.The unsupervised learning approach, GloVe (Global Vectors for Word Representation),produces word embeddings and dense vector illustrations of words in a constant vector field. The mathematical equation for GloVe can be described as follows:

V: The total vocabulary size, i.e., the number of unique words in the corpus.W: The co-occurrence matrix of size $V \times V$, where $W[i,j]$ represents the number of times the word i and word j co-occur in the context within a specific window size.X: The word embedding matrix of size $V \times d$, where d is the dimension of the word embeddings.b: The bias term vector of size V.The objective of the GloVe algorithm is to learn word embeddings such that the dot product of two-word embeddings approximates the logarithm of their co-occurrence probability.TheGloVe loss function is defined as follows:

$$Loss = \sum(i,j)\,[f(W[i,j])\,(X[i]^T * X[j] + b[i] + b[j] - log(W[i,j]))^2] \quad (1)$$

(i,j)iterates over all word pairs in the co-occurrence matrix W.$f(W[i,j])$is a weighting function that accounts for the common words that occur frequently and are less informative. It can be defined as

$$f(x) = min\left(1.0, \left(\frac{x}{max\_cooccurrence\_value}\right)^{\alpha}\right) \quad (2)$$

where$max_cooccurrence_value$ is the maximum value in the co-occurrence matrix, and α is a hyperparameter (typically set to 0.75).$X[i]^T$represents the transposition of the ith word's embedding vector in X.$b[i]$is the bias term for the ith word.The optimisation goal is to minimise this loss function by updating the word embedding matrix X and the bias term vector b using stochastic gradient descent (SGD) or other optimisationmethods.GloVe combines the global co-occurrence statistics of words with the local context-based information to learn word embeddings that capture semantic and syntactic relationships

between words in the corpus. The resulting word embeddings can be used for various NLP tasks such as word similarity, word analogy, and text classification.

3.3)Dolphin Echolocation Algorithm (DEA)

The Dolphin Echolocation Algorithm (DEA) is a novel optimisation technique that draws inspiration from the remarkable echolocation ability exhibited by dolphins. This algorithm has been developed with the primary objective of addressing various optimisation problems encountered in diverse fields.The concept of DEA is inspired by the unique click sounds produced by dolphins during their hunting activities. These click sounds possess distinct characteristics that have been studied and analysed for their potential applications in various fields. In the process of foraging for food, dolphins employ a unique method known as echolocation. This biological mechanism involves the generation of clicks by dolphins, which are emitted into their surrounding environment. These clicks serve a crucial purpose as they interact with potential prey items. Upon contact with the prey, the clicks are reflected to the dolphin as echoes. This enables the dolphin to gather valuable information about the prey's location, size, and other relevant characteristics, facilitating successful foraging endeavours.

The process of dolphins estimating the location and distance from prey is known as dolphin echolocation. This ability involves analysing echoes and is utilised by dolphins to optimise their hunting strategies. In the initial phase, the population of dolphins is established. Subsequently, each feature is examined, and the search space alternatives are arranged in ascending or descending order. In this sorting method, feature vectors are generated by representing the columns of the alternative’s matrix. The NL locations for the dolphins are randomly selected, simulating a research scenario. The change of convergence factor is determined by calculating the PP (Probability of Progress) of the current loop, as commonly done in research studies.

Algorithm 1: DEA based feature selection

1. **Start** a dolphin's NL locations randomly, then sort the alternatives in the search space into a matrix called "Alternatives."
2. **Use** Eq. (1) to get the loop's PP.
3. **make** a best guess as to how suitable each area is.
4. Use Eq. (2) to **calculate** the cumulative fitness.
5. **Vary** the array's value to disperse the search space.
6. Figure out where the current loop should be placed and set their AF to 0.
7. **For** each variable, determine its probability, $\bar{P}_{ij}$.
8. Using Eq. (5), **assign** the probability $\bar{P}_{ij}$To the best option for each variable in the best location set, and the balance of the probability to the other options.
9. **Determine** where to go next based on the probability you've given each possible course of action.
10. **If** the termination condition is not reached, **return** to Step 2.

3.4) Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a deep learning model commonly used for text classification tasks. In text classification, the input is a sequence of words (text), and the output is a class label representing the category or sentiment of the text.The application of filters in this layer results in the creation of feature maps and the extraction of detected

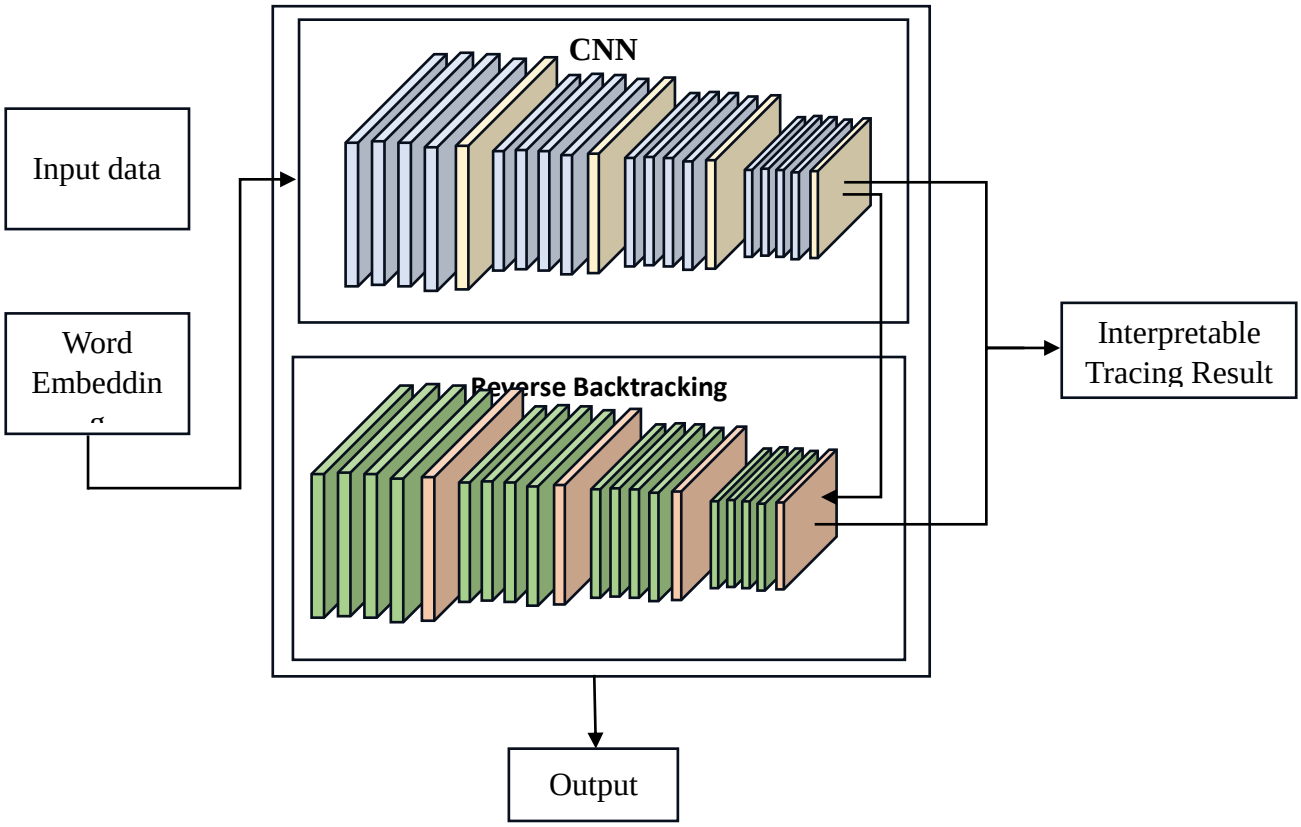


Fig.3. Convolutional Neural Network Architecture

features from the input. The process involves the utilisation of a small matrix of numerical values, commonly referred to as a kernel or filter. The CNN architecture is shown in the figure 3

This kernel is applied to the matrix representing the paragraph, resulting in a transformation of the paragraph matrix based on the values within the filter. In this research, we consider the input paragraph matrix, denoted as $E_w(m, n)$, where m and n represent the dimensions of the matrix. Additionally, we introduce a two-dimensional matrix H, which has a kernel size of $(2g + 1, 2d + 1)$. Here, g and d are constants that are predetermined. The representation of the outcome of the convolutional layer is typically denoted by,

$$y_{i,j} = \sum_{m=-h}^{h} \left(\sum_{n=-i}^{i} K[m,n] H[i-m, j-n] \right) \quad (3)$$

The ReLu activation layer is commonly used after each convolution layer to introduce non-linearity and enhance the expressive power of the neural network.It helps to normalise the output by selectively activating only the positive values and setting the negative values to zero. This non-linear activation function aids in capturing complex patterns and features within the input data, thereby improving the network's ability to learn and make accurate predictions. The inclusion of this layer in the model facilitates the acquisition of complex and intricate knowledge while minimising the risk of encountering the vanishing gradient problem and maintaining low computational expenses. In this context, $y_{i,j}$ represents the input to the Rectified Linear Unit (ReLU) function.

$$A(y_{i,j}) = max(0, y_{i,j}) \quad (4)$$

The suggested method combines and feeds into the fully connected layer feature maps acquired using different kernel sizes. Using this strategy, data from various kernel sizes may be combined to improve the model's representation and functionality. Connecting all the activations from the preceding layers is the job of the completely connected layer, which may also be referred to as a dense layer or a multilayer perceptron. Matrix multiplication is the mathematical procedure by which the weights associated with each neuron are multiplied by a specific matrix and the resulting product is then offset by a given number to decide which neurons are activated. Overfitting may be prevented in part by using a dropout layer, a technique that is becoming popular. To function, this layer intermittently toggles on and off the outbound connections of concealed units throughout the training phase's iterative updates. The dropout layer helps mitigate overfitting because of this unpredictability. At last, the classification layer uses the attributes acquired in earlier layers to complete classification tasks.Figure 4 shows the layered structure of the CNN model.In the context of multi-class classification, CNNs are commonly employed and trained using a technique known as categorical cross-entropy. This approach minimises the categorical cross-entropy loss function, calculated over a softmax activation function. The loss above functions can be mathematically formulated as,

$$L = -\frac{1}{K} \sum_{a}^{k} \sum_{b}^{l} y_{i,j} log \left(\frac{e^{\hat{y}_{i,j}}}{\sum_{s=1}^{l} e^{\hat{y}_{i,j}}} \right) \quad (5)$$

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 1000)	0	
embedding_1 (Embedding)	(None, 1000, 100)	17407500	input_1[0][0]
reshape_1 (Reshape)	(None, 1000, 100, 1)	0	embedding_1[0][0]
conv2d_1 (Conv2D)	(None, 998, 1, 512)	154112	reshape_1[0][0]
conv2d_2 (Conv2D)	(None, 997, 1, 512)	205312	reshape_1[0][0]
conv2d_3 (Conv2D)	(None, 996, 1, 512)	256512	reshape_1[0][0]
max_pooling2d_1 (MaxPooling2D)	(None, 1, 1, 512)	0	conv2d_1[0][0]
max_pooling2d_2 (MaxPooling2D)	(None, 1, 1, 512)	0	conv2d_2[0][0]
max_pooling2d_3 (MaxPooling2D)	(None, 1, 1, 512)	0	conv2d_3[0][0]
concatenate_1 (Concatenate)	(None, 3, 1, 512)	0	max_pooling2d_1[0][0] max_pooling2d_2[0][0] max_pooling2d_3[0][0]
flatten_1 (Flatten)	(None, 1536)	0	concatenate_1[0][0]
dropout_1 (Dropout)	(None, 1536)	0	flatten_1[0][0]
dense_1 (Dense)	(None, 20)	30740	dropout_1[0][0]

Total params: 18,054,176
Trainable params: 646,676
Non-trainable params: 17,407,500

Fig.4. The Layered structure of CNN architecture

The CNN text categorisation outcomes form the basis of the Backtracking Analytical Model. A backtracking study of model-predicted labels reveals the most influential variables in the prediction process. Convolutional neural networks indicate learned visual concepts, making them ideal candidates for visualising. A weighted space diagram that maps input features such as "class of each channel to the significance" and "the strength of the activation of the various channels in the input text" can reveal the convolution and pooling layer in the network's output during training. The CNN text classification model was used to categorise a text, and then the process was reversed by labelling the categories. The anticipated results were back-calculated through a series of tightly connected layers, pooling layers, and convolutional layers to determine the relative importance of each input text vector value component. This significance was quantified at each position along the input vector and served as the primary information for the model's following accessibility study. The deconvolution network might help with model creation, troubleshooting, and viewing and explaining the CNN model's underlying structure. Text restoration was used to identify the influencing keywords, and an improved classification model was obtained from the inside.

3) Result and discussion

Performance evaluation in the context of text classification involves assessing how well a trained model predicts the class labels of text samples. Common evaluation metrics include accuracy, precision, recall, F1 score, and confusion matrix.In machine learning and data analysis, "True Positives" (TrP) refers to the number of positive samples a given model or algorithm has accurately predicted. It is a crucial metric used to evaluate the performance and effectiveness of classification models. True Positives represent the instances where the model correctly identifies positive samples from a dataset. These samples can be anything from identifying spam emails to detecting them. In predictive modelling and classification, "True Negatives" (TrN) refers to the number of samples correctly predicted as unfavourable. It is an important metric used to evaluate the performance of a classification model. When a classification model is trained and tested on a dataset, it makes predictions about the class labels of the samples. In binary classification problems, there are two possible classes: positive and

The concept of false positives (FaP) refers to the number of positive samples that are incorrectly predicted or identified as positive by a specific model or algorithm. In various research studies and experiments, false positives are often used as a metric to evaluate the performance and accuracy of predictive models or classification algorithms. The calculation of false positives involves comparing the predicted positive samples with the ground truth or actual positive samples, by quantifying the number. False negatives (FaN) refer to the number of negative samples incorrectly predicted by a model or algorithm. In other words, false negatives represent instances where the model fails to identify a sample as negative when it should have. This metric is commonly used in various fields of research and analysis, such as machine learning, medical diagnostics, and quality control, by quantifying the number of false negatives, researchers.

Accuracy is a commonly used metric in machine learning to evaluate a model's performance. It quantifies the degree of correctness in the model's predictions.

$$Accuracy = \frac{(TrP + TrN)}{(TrP + TrN + FaP + FaN)} \quad (6)$$

Precision is used to evaluate the level of accuracy in the positive predictions generated by a given model.

$$Precision = \frac{TrP}{TrP + FaP} \quad (7)$$

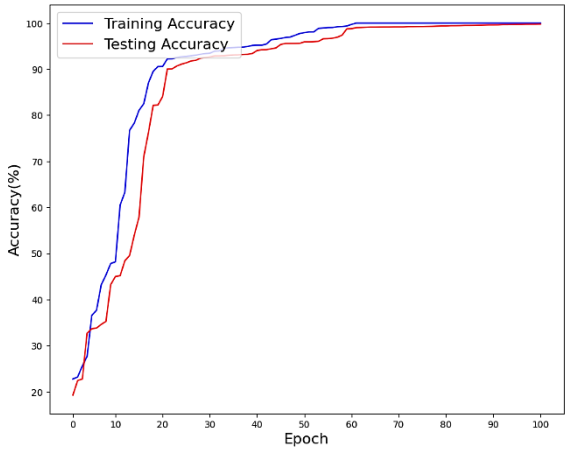
The recall metric quantifies the ratio of correctly foretold positive samples to the total number of actual positive samples in the model's predictions.

$$Recall = \frac{TrP}{TrP + FaN} \quad (8)$$

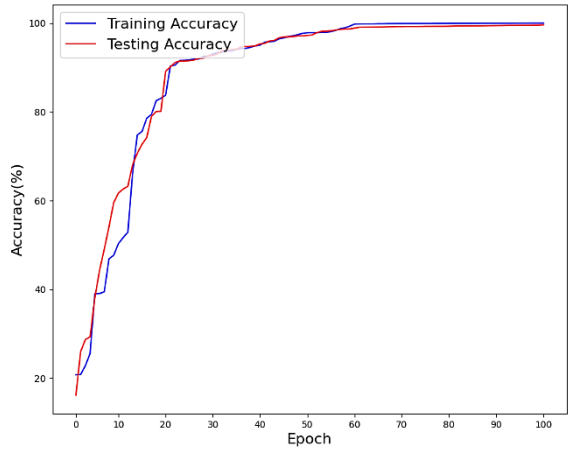
The F1 score is a widely used evaluation metric in various fields, including machine learning and information retrieval. It is calculated as the harmonic mean of precision and recall, which allows for a balanced assessment of a model's performance. By considering both precision and recall, the F1 score provides a comprehensive measure of how well a model can correctly identify relevant instances while minimising false positives and negatives. This balanced evaluation metric is handy when dealing with imbalanced datasets or when precision and recall are equally important in a given application.

$$F1 - Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (9)$$

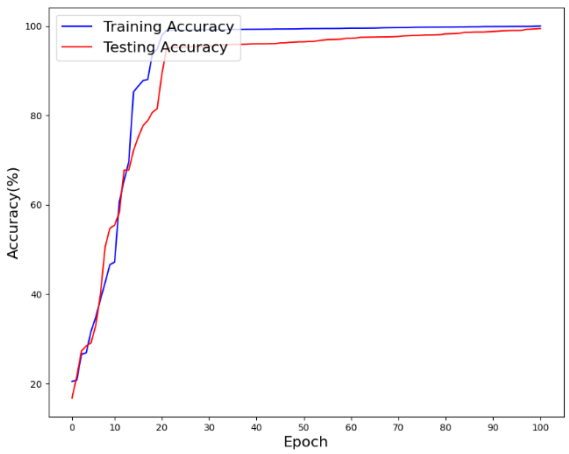
The confusion matrix is widely used in machine learning and data analysis. It is a table that provides a comprehensive representation of the performance of a classification model. The matrix is constructed by comparing the model's predicted labels with the data's actual labels. It consists of four key elements: true positives (TrP), true negatives (TrN), false positives (FaP), and false negatives (FaN). These elements allow a detailed analysis of the model's accuracy, precision, recall, and other performance metrics. The confusion matrix is a valuable resource for evaluating and fine-tuning classification models, enabling the presented data to provide an overview of the number of true positive, true negative, false positive, and false negative predictions.



(a)



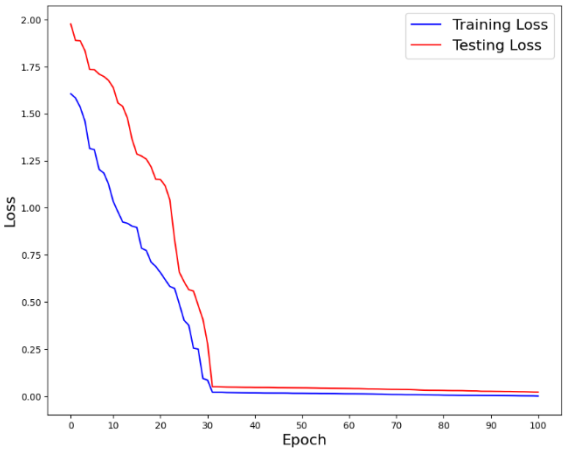
(b)



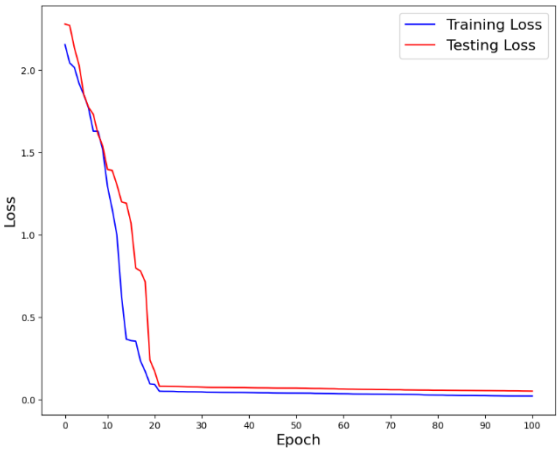
(c)

Fig.5. Accuracy of Insights from Large Scale Social Media Data through Distributed Deep Learning on (a) Sentiment 140, (b)Tweets Airline, and (c)Tweets SemEval.

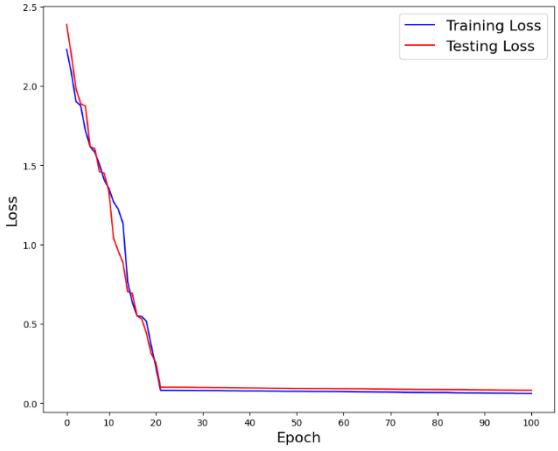
Fig.5 shows the accuracy results of a distributed deep learning approach applied to different large-scale social media datasets. Fig.5 (a) indicates that the distributed deep learning approach achieved an accuracy of 99.8% on the Sentiment140 dataset. The Sentiment140 dataset is widely used for sentiment analysis, containing tweets labelled with positive or negative sentiment. The 99.8% accuracy suggests that the model correctly predicted the sentiment of most tweets in the dataset. Fig.5 (b) indicates that the distributed deep learning approach achieved an accuracy of 99.7% on the Tweets Airline dataset. This dataset likely contains tweets related to airline experiences, and the high accuracy suggests that the model performed very well in predicting sentiment or other classes associated with the airline-related tweets. Fig.5 (c) indicates that the distributed deep learning approach achieved an accuracy of 99.6% on the Tweets SemEval dataset. The SemEval dataset is another popular dataset used for various NLP tasks, and the high accuracy indicates that the model successfully predicted the class labels or sentiments for the tweets in this dataset.



(a)



(b)

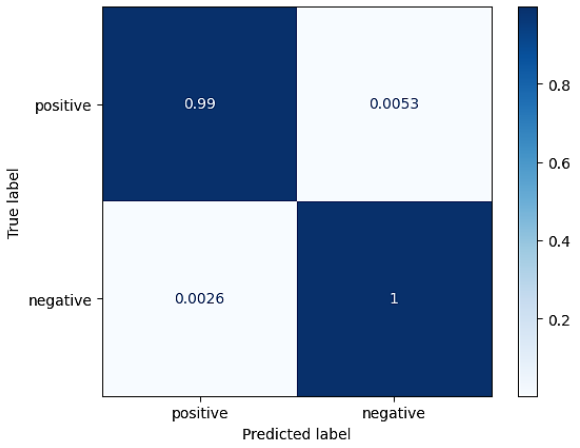


(c)

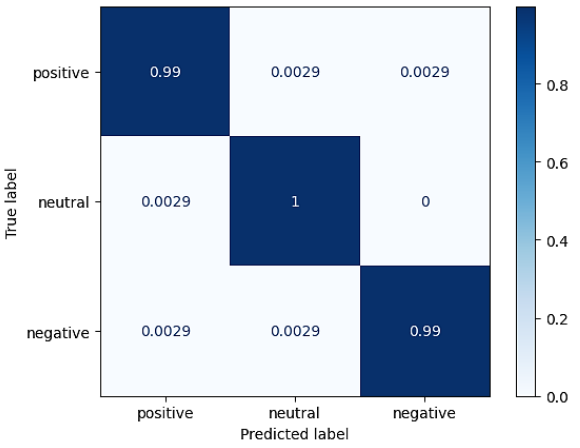
Fig.6. Loss of Insights from Large Scale Social Media Data through Distributed Deep Learning on (a) Sentiment140, (b)Tweets Airline, and (c)Tweets SemEval.

Figure 6 depicts the loss values acquired while implementing a distributed deep-learning approach on extensive social media datasets. Within deep learning, the loss function plays a crucial role in quantifying the dissimilarity between the predicted output of a model and the true labels associated with the given data. The primary objective during the training phase is to minimise the loss function, which measures how effectively the model can capture and represent the underlying patterns in the data.According to the findings presented in Figure 6.a, the distributed deep learning approach demonstrated a notable achievement by attaining a loss value of 0.02 on the Sentiment140 dataset. The low loss value observed in the model indicates a high level of accuracy in predicting sentiment labels that closely align with the actual sentiment labels present in the dataset. The model exhibits strong performance by generating precise predictions with negligible errors.

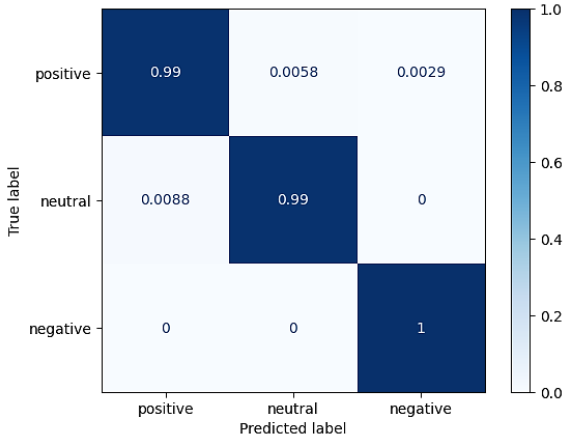
The study observed that the distributed deep learning approach yielded a loss value of 0.05 when applied to the Tweets Airline dataset, as depicted in Fig.6.b.Similarly, a dataset with a low loss value indicates that the model is proficiently making predictions regarding the sentiment or other class labels linked to tweets related to airlines. According to the findings presented in Figure 6.c, the distributed deep learning approach demonstrated a loss value of 0.05 when applied to the Tweets SemEval dataset. Similar to previous datasets, a low loss value indicates that the model effectively predicts the class labels or sentiments for the tweets included in this dataset. The statement emphasises the efficacy of employing a distributed deep learning approach to reduce loss values across various social media datasets. Observing low loss values suggests that the model effectively acquires knowledge from the provided data and generates precise predictions. The findings of this study indicate that the proposed methodology is highly suitable for extracting insights and sentiment analysis tasks when applied to social media data.



(a)



(b)



(c)

Fig.7. Confusion matrix of Insights from large-scale Social Media Data through Distributed Deep Learning on (a) Sentiment140, (b)Tweets Airline, and (c)Tweets SemEval.

The confusion matrices in Figure 7 illustrate the outcomes of employing a distributed deep learning technique on various extensive social media datasets for sentiment analysis. Figure 7a illustrates the confusion matrix about the Sentiment140 dataset. The results demonstrate that the model attained a remarkable accuracy rate of 99% in accurately predicting positive samples, also known as true positives. Additionally, all negative samples were correctly classified as true negatives. The results suggest that the model accurately predicted sentiments, regardless of whether they were positive or negative. In the research conducted on the Tweets Airline dataset, Fig.7.b illustrates the confusion matrix. The results indicate that the model demonstrated a notable level of accuracy in its predictions for positive, neutral, and negative samples. The accuracy rates were 99% for both positive and negative samples, while neutral samples achieved a perfect % accuracy rate of 100%. In this research, Fig7.c presents a detailed depiction of the confusion matrix constructed for the Tweets SemEval dataset.

The confusion matrix is a valuable tool in evaluating the performance of a classification model by illustrating the distribution of predicted labels against the actual labels. By analysing the confusion matrix, researchers can gain insights into the accuracy and effectiveness of the classification model in correctly classifying tweets within the SemEval dataset. The results suggest that the model demonstrated a notable level of accuracy when predicting samples classified as positive, neutral, and negative. The model demonstrated exceptional performance, achieving a remarkable accuracy rate of 99% for both positive and neutral samples. Furthermore, it exhibited flawless accuracy, achieving a perfect score of 100% for negative samples.

4) Conclusion

In conclusion, the exponential growth of social media data has posed a significant challenge in extracting valuable patterns and insights. However, the authors of this study have presented a novel strategy that leverages deep learning methods to overcome this difficulty. The proposed system efficiently handles massive volumes of social media data by harnessing the

power of distributed computing and integrating it with deep learning approaches, enabling fast and accurate detection of relevant patterns, feelings, and trends.The primary goal of this research is to gain useful insights into user behaviour, preferences, and social interactions. To achieve this, the study adopts deep learning models, adapted and improved to handle social media data's diverse and unstructured nature, including text, video, and user interactions. The approach also introduces the Dolphin Echolocation Algorithm (DEA) during the feature selection phase, fine-tuning the convolutional neural networks (CNN) filter weights and enabling multi-angle analysis of multi-classified text and multi-label classification tasks.Addressing data diversity, volume, and velocity challenges in large-scale social media datasets is a crucial aspect of the proposed method. By using as few resources as possible, the study reduces costs and promotes environmental sustainability when dealing with distributed databases of any size.To validate the efficacy and efficiency of the proposed framework, extensive tests were performed on various datasets collected from major social media sites. The outcomes demonstrate that the strategy outperforms the existing methods regarding precision, processing speed, and scalability. The model achieves an accuracy of 99.6% to 99.8 %. The proposed approach shows great promise in unlocking valuable insights from the vast ocean of social media data, helping researchers and practitioners better understand user behaviour, sentiment, and interactions in real-time or near-real-time scenarios.

Future work in social media data analysis and deep learning could focus on several aspects to further enhance the understanding of user behavior, sentiment, and interactions. One approach could be to explore the integration of multiple data modalities, such as text, images, and videos, to gain a deeper understanding of user engagement and sentiment on social media. This could involve developing advanced deep learning models to process and interpret multi-modal data effectively.