Received: September 13, 2023. Revised: November 3, 2023.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Social Media Data-Based Business Intelligence Analysis Using Deep Learning

Ruwaida Mohammed[1]* Israa Shihab[1] Mustafa Musa[1]

[1]*Informatics Institute for Post-Graduation Studies,*
IIPS Iraqi Commission for Computer & Informatics (ICCI), Iraq
* Corresponding author's Email: roueida.m.yas@iips.icci.edu.iq

Abstract: The digitization of goods and services is predicated on the production of value by contemporary organizations. What customers say about a firm on social media has become a significant part of the big data revolution. Furthermore, social media data analytics is a challenging topic to learn because of the partiality of text assessment and the extra variables. Social media data analytics using the deep learning (SMDA-DL) framework is proposed in this paper. A two-phase framework is suggested in the SMDA-DL framework: Pre-processing stage and selecting an efficient deep learning technique for this information is the emphasis of phase one; the second phase relies on existing components of large data structures and the model developed in phase one. First, small and large databases in non-big data sets will be analyzed in the first step of the suggested approach. While the first step of the deep learning model may be used to analyze massive data databases, the second stage is recommended. For the first layer of the architecture, there is a case study for assessing social media assessments. The proposed framework yields good simulation results in an accuracy ratio of 98.7%, precision ratio of 99.7%, recall ratio of 98.4% and F score of 98%.

Keywords: Deep learning, Artificial intelligence, Business intelligence, Social media, Data analytics, Application programming interfaces (APIs).

1. Introduction to social media

Social media is a good process that everyone owns, from kids to kids to adults. Facebook, Twitter, Line, and Tumblr have joined social media to make social media more vibrant [1]. Effective digital marketing advertising is around using business postings on social media to target a particular target. You'll also be required to construct market profiles if users don't already have a solid understanding of who the main demographic is. Their significance in making it simple for users to engage, share, and produce content, such as blogging, social networks, online communities, newsgroups, and virtual reality, increases social media for all consumers [2]. It is backed up by data from smartinsights.com, which shows that the number of available social media has surpassed 1.69 billion, Twitter has exceeded 340 million, Snapchat has exceeded 450 million, and so on [3]. According to these figures, Facebook is the most popular social networking platform with the most registered users.

The effectiveness of social media is significantly influenced by the number of active users indicated [4]. The effect is that social media, formerly used to engage in, exchange, and form communities, can now advertise and engage with customers via social media. It recognizes that digital marketing needs constant monitoring to avoid misleading advertising, campaigning, and customer interaction [5]. Let's examine its effects if we wish to use internet advertising to expand the company: Utilize SEO to generate business who users never would have met else. Recognize the product offerings that are important to the business. effectively link them with both present and potential consumers by communicating their brand. This is so that interactive advertising can help clients navigate the selling process by delivering more specialized and tailored material as necessary. Source creation: Generating quality traffic is the main goal of any promotional campaigns. Sentiment analysis is a type of text

mining process to determine subjective attitude, such as sentiments from written texts, and hence has become a main research interest in domains of natural language processing and data mining [6]. This monitoring is performed directly on social networking sites such as Facebook, which is the most common for marketing efforts, followed by Instagram, Tumblr, Twitter, and Pinterest [6].

This study can develop and document the data storage concept and technology for the business intelligence technology. A strategic organizational toolbox for gathering and analysing internal information is referred to as corporate information. A variety of IT products, from equipment as well as real-world sensing to application systems or data storage, could be a part of such technology. Its features are self-serve reporting, quick deployment, in-memory analytics, data warehousing, visual analytics, OLAP, reporting timing, and enhanced security are just a few of the features available. Data sources can be gathered from LinkedIn and Instagram account information. Both can be divided into entertainment, economics, health, culture, automobile, political, sports, technologies, and transportation. Three algorithms, including Nave Bayes, Random Forest, and Support Vector Machine, can be used to test the text categorization approach for social media [5].

The best possible results from evaluating the three methods can be chosen to build a classification model, which can be extremely useful for data storage deployment [1]. The Kimball technique is the design strategy to employ for database systems. A data warehouse is designed as well as developed using the Kimball set of established methodologies, procedures, and procedures. It is additionally alluded to by other terms, including Kimball's multidimensional modelling, the bottom-up method, or the Kimball relational database life cycle concept. Select the business operations, define the granularity, specify the measurements, and define the facts are the four processes that must be completed to construct a database system using this approach [7]. The ETL procedure (extraction, transformation, and load) is a step that must be completed before a relational database can be built. The next step is to build business intelligence utilizing Carlo Vercelli's process, which has four key stages: evaluation, development, implementation, and management [8].

Several quantitative financial models relying on the concept of equilibrium price and the no-arbitrage premise have been presented to capture interest variations in recent years. Moreover, probabilistic process concepts are commonly distinguished [4, 10, 6] These systems are driven by economic theory. They cannot exchange rate patterns during a crisis. As a result, the requirement for a more realistic and adaptable model to represent interest rate changes has developed [9]. These regression systems can detect enabling legislation and cannot produce prediction measures.

These systems are predicated on the presumption that disruption terms are normal and variables have a linear connection. [3] adopting artificial intelligence particularly machine learning systems to provide precise and reliable findings is unavoidable [10] These models are more favourable and effective since they capture non-linear trends in time series analysis. There has been very little research into the use of social networks in interest rate forecasting [11]. No research examines the influence of major local and worldwide occurrences on interest rate forecasting. User-generated content, commonly referred to as UGC or consumer-generated material, is unique material made by clients specifically for a business and shared on social media or through other means. UGC may take several different formats, such as pictures, movies, comments, a recommendation, or a webcast. UGC may be utilized in promotional campaigns outside of social media, rendering the plan an integrated one. For instance, one might include UGC photographs in a message sent to a customer who has abandoned their shopping cart to encourage them to acquire something, or one could incorporate user-generated content into important webpages to boost conversions. Three things define user-generated content (UGC): (1) it is created by consumers, not by the company that distributes it; (2) it is innovative in design as well as the consumer contributes anything unique; and (3) it is uploaded to the internet and is widely available. A common illustration of a feedforward artificial neural system is an MLP. The ith activating component in the lth level is represented by the letter ai in this diagram (l). The variable settings of a neuronal system, which include the quantity of layers as well as transistors, require tweaking.

This research proposes a computational methodology for managing big data (BD) that focuses on sets of knowledge comprising user-generated content (UGC), not entirely. This paper makes two commitments:

1) This paper presents a BD and deep learning (DL) architecture for analyzing both subjective (text appraisals) and numerical (user evaluations) data for predicting content analysis

2) The proposed uses a combination of DL and natural language processing (NLP) methods to classify customer reviews negatively or positively in a sample of the Yelp data.

3) The findings show that by employing a multi-layer perceptron, a high level of precision was attained for the classification algorithm.

The upcoming section is as follows. Section 2 deals with the background and the literature survey of the social media data analytics methods. The proposed social media data analytics using the deep learning (SMDA-DL) framework is designed and implemented in section 3. The software analysis and performance evaluation are illustrated in section 4. The conclusion and future scope are illustrated in section 5.

2. Background to social media data analytics method

In [12] described the research using data analysis for social networking business information. The foundation of their study is that business intelligence improves the analysis of data on social networking simpler and can even extract hidden insights from data. This might happen when data mining techniques are coupled with business intelligence. With the development of data mining was now possible to extract new knowledge from enormous amounts of information such as classification and patterns to forecast anything based on previous information. There are numerous strategies in data analysis that are split into two categories: descriptive and predictive tasks.

The detection of satisfaction level of customer joy for online businesses on social media especially from Facebook and twitter was analyzed using machine learning. Stemming, normalisation, and preventing the removal of words are three of the most well-known NLP technologies used in large-scale data analysis by businesses on social media. The performance was analyzed using classifiers of type random forest (RF) and support vector machine (SVM) achieved 90.9 and 93.4% of F1-measures with the addition of deep learning algorithms for word embedding and bag of words calculation [15].

The most often used methodology for forecasting customer happiness was logistic regression. This data-gathering methodology can tackle problems like identifying user emotions in social media posts, recognizing behaviours, and conducting demography on Facebook users [13] conducting information and learning more about consumers. Logistic regression (LR) was used as a prediction model that reads each tweet and categorized using a dictionary and sentiment analysis to show the risk type with an accuracy of 93.2%. The sample was categorized only for 3 types it was a major limitation [14].

In addition, develop and analyze the neighbourhood's personal and group problem-solving solutions. Data mining has numerous parts for managing electronic commerce information and resolving some of the mentioned challenges [18, 19]. These parts start with data gathering and then move on to statistical analysis, which includes preprocessing and modelling data using a data mining technique to recover knowledge from the database extraction results that may be used to make decisions, anticipate user behaviour, and define corporate strategies.

In [15] investigated the significance of data mining incorporating process automation. Data mining would be a fascinating technology that may help companies concentrate on the most relevant data in current data warehouses by extracting hidden give valuable information from big databases. Organizations may use data mining technologies to anticipate future trends and practices and make data-driven choices.

Data mining's automated, speculative examinations go beyond the breakdowns of past events provided by theoretical type of decision assist network members. Data mining tools can help answer previously too time-consuming questions [20]. They create datasets to find hidden cases and foresight data that experts may overlook was beyond their scope of interest. With data gathering on business intelligence, solutions such as identity verification, corporate finance, customer behaviour assessment, competitive analysis, and purchase orders can be developed. [16] created business intelligence solutions that support consumer happiness. To conduct customer satisfaction research, it is essential to know the demographics of Facebook and Instagram users whose information may be retrieved. Multinomial Naïve Bayes (MNB) was used for classifying the text and controls the decision boundary of the prediction model of Arabic tweets. The model was built with the grid search validation and stopping criteria for optimal parameter for the count vectorizer and achieved average accuracy as 56.2%. Since, the Arabic tweets attributes were found to be noisy it impacts the classification accuracy of the model [22]. Decision tree (DT) was applied as a credit scoring based on LinkedIn social media information to evaluate creditworthiness. The work evaluates creditworthiness by analysing posts made by borrowers, which include discussions and comments on social media. The results achieved with an accuracy of 85.9%, precision of 91.50%, recall of 73.33, and f1-score was 81.41%. An expert judgment for threshold analysis may create bias it was a major

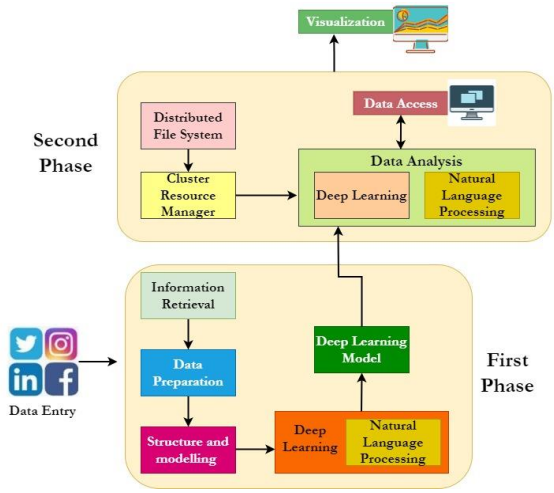


Figure. 1 The architecture of SMDA-DL framework

drawback [23]. Text mining, grouping, which splits information into relevant or useful groupings, and visualization as a platform for visualizing information to end-users are detailed approaches used to construct business intelligence systems. Because some types of visualization representations are relatively expensive to produce, one problem for the visualization industry is to keep up with the fast-growing amount of data found in many potential applications. The enormous range of data types used in various scientific fields, which often involves data interpretation before it can be visualized, is yet another clear hurdle to the widespread use of visualization tools in research. Social media data analytics using deep learning (SMDA-DL) framework is proposed to overcome these drawbacks.

3. Proposed social media data analytical using deep learning (SMDA-DL) framework

3.1 Framework for social media data integration and big data analytics

The suggested framework's theoretical aspects are presented in this section. The architecture is divided into data management and business process automation.

Fig 1 shows the social media data analytics architecture using the deep learning (SMDA-DL) framework. It consists of two phases. Companies can use accessible data to take advantage of the latest and obtain crucial knowledge about the ideas concealed in values as BD becomes more significant. Handling enormous amounts of data and analysis procedures is a major challenge that must be considered. The rewards are represented in gaining an advantage over rivals or improvements in internal procedures. The structure's two phases can be summarized as follows: it takes huge amounts of unorganized text data as inputs, uses a DLM to extract significant insights, and then visualizes the findings using high-level approaches.

- First phase

The framework's initial phase presents a sequence of phases that describe data sequence, from entering data to DLM creation. A percentage of big data (BD) should indeed be acquired from big small data (BSD) sources for better results in this stage, which analyzes data that may be big data (BD) or big small data (BSD). Small data is, unsurprisingly, shorter whereas big data comprises bigger amounts of data. One method of viewing at it is that huge quantities of large amounts of data are frequently referred to as big data. On the other side, data points focus on finer, more manageable parameters. This approach is advantageous since it provides a later sense of data structure and format. Furthermore, this methodology suggests text data collection; other data resources could be effectively examined using different components.

Step 1: data entry

Text data information: unorganized data are used to enter data into the system. Information from social media outlets, blogging, chat rooms, and microblogging sites can all be used. The data to be studied in this research is Twitter data, particularly user evaluations of hotels and restaurants. However, the approach could be applied to any data source that contains social textual information.

Step 2: phase of data collection

The platform's first phase is energy processing, which entails taking the appropriate steps to gather necessary data, even if many businesses can have their text data sources. A group of businesses in the field of energy supply as well as manufacturing make up the electricity industry. Energy firms may control rolling blackouts, determine seasonal demand, and establish energy cost by using predictive analytics to discover power usage and fuel savings. Machine learning could be utilized by oil and gas companies to assist manage transportation and refining operations and instantly respond to customer needs. Corporations collect data for extensive research projects, such as consumer feedback or sensor readings. The collection of unnecessary data must be prevented during the data capture procedure. Information gathering solutions that cut expenses by

lessening data centres' computing and storage demand are beneficial.

Data collecting techniques assist in gathering a variety of information for BD storage. Web shears are either developed in-house or purchased from a private entity. Application programming interfaces (APIs) supplied by social networking platforms such as Instagram, Flickr, and Twitter can be used to gather information. Web APIs usually defines the format of messages contain as well as utilize HTTP for incoming and outgoing traffic. XML or JSON files are frequently used for this system receives. Since they offer information in a style that makes it simple for other applications to handle, XML as well as JSON are both favoured standards. It provides a useful way for collecting data straight from web pages using user-generated content (UGC).

Data collecting must guarantee that all necessary information for the study is acquired; a noisy database or a lack of reporting may happen. One viable solution is to gather data using evolutionary algorithms to avoid a data deficit. The methods' adaptability and their capacity to self-adapt the research for the finest alternatives on the spot are two particular benefits. The use of simulated annealing is becoming commonplace as personal computers get faster. Returning to this step from later steps is always possible to would affect the development's timeframe, delaying its conclusion. Because it is accountable for delivering the information that can be needed in the future stages, the output of this step immediately enters the data preprocessing step.

Step 3: phase of data preparation

During the data discovery phase, the tasks performed were connected to acquiring BSD from varied processes: these data sets must now be cleaned and readied for more examination. The second step, data preparations are the most significant; other related phases include data pretreatment and integration procedures. Information gathering takes up a significant amount of time for data analytics teams. Seventy-five percent of the overall project duration is projected to be spent on preprocessing and 25 percent on evaluation, known as data "cleansing."

The data preparation step relates to identifying exceptions or inconsistent numbers compared to the rest of the information. When dealing with quantitative data, they should generally be within a scale that meets the survey's requirements. However, readings are usually considerably above or much below the predicted signs. While working with communication costs or sequences, many are null or relate to other languages, causing the messages to alter.

In any of these circumstances, data processing must involve applying various techniques (mainly analytical) to give those records specific care. A considerable variation in the study outcomes can occur if improper data is not detected. The preprocessing phase's result is material with a homogenous structure and statistics that meet the research's goals and no anomalies in the existing information. This step can allow BSD approaches to receive data, making the process easier by removing the noise. The preprocessing step can be categorized into two: the raw data is "cleared away" and used in the following step, which is accountable for evaluating the architecture of the material and creating schematic modelling that covers the total collection of data.

Step 4: structure and modelling

The architecture and optimization phase aims to determine these big data by creating a data model that explains their internal relationships and content. By reviewing all of the components, accessing the information, and understanding their relationships, a data architecture can be easily created if the information is single-sourced. Alternatively, random searches or direct data selections on the object classes can be needed to assess: the pieces that make up the details, relationships, and the various sorts of information. The architecture and modelling step is comparable to the technique of developing database systems, in which the outcome is a framework of interconnected rows and their columns. Organizing and modelling BSD gets more complicated when information originates from two or more data resources. It is one of the data analytics team's most difficult responsibilities, where the most important discoveries are concealed. The method of this step can be repeated for each data source to solve this problem. After creating all the database schemas, it needs to create the outside relationships related to each resource and its objectives. Since this step entails producing an organized model from unorganized data, various measures must be made when building the data model. Because information from social networks like LinkedIn and Instagram is orientated to a directed graph, a linear database management systems model into a collection of tables, including rows and columns, could be a better match for some unorganized data streams. Fundamentally, a database management system (or DBMS) is just a computerized data-keeping mechanism. Customers of the platform are provided

with the ability to carry out a variety of actions on this type of platform for either managing the data structures on its own or manipulating the information contained in the database. A social network's relationships between individuals, communities, or organizations are depicted in a proposed network. The phrase is additionally utilized to refer to a person's social media platform. A sociological structure emerges as a collection of network elements linked together by lines whenever it is represented as a mapping. As a result, an effective model for representing unorganized data should include various products, including linear and network designs. This step aims to learn about data: its nature, internal structure, the value set for each area, and the complex relationship between data types. This information structure can apply analytical methods that can better comprehend information.

Step 5: deep learning and NLP

This phase claims that DL, in conjunction with NLP approaches, can improve the evaluation of unorganized text data produced during the data preparation phase. With the help of NLP, we could precisely retrieve the knowledge and insights from the material as well as classify them into the appropriate groups. For instance, once a customer browses for anything on the Google search engine, and Google's algorithm uses NLP methods to display all the important files, websites, as well as publications. Deep learning transforms several research fields from text identification to driverless vehicles. RNNs have been utilized to research a variety of NLP problems, including computational linguistics, picture tagging, as well as language processing, between everyone else. An RNN system can do some natural language problems as well as or perhaps higher than a Classification algorithm, although it is not always preferable. NLP is a branch of computer technology that studies natural language to let people interact with machines like they do with others. When observing the results, data analysis can benefit from deep learning and natural language processing technologies. Although digital information can be studied using other methods, it is suggested that DL and NLP methods be used for the current data for the purposes mentioned above: first, because they are the most widely acknowledged and proven techniques among computer scientists, and secondly, since there is a professional and technical society that promotes and encourages these technologies. Many data analysis studies are done using various DL algorithms; NLP is considered in

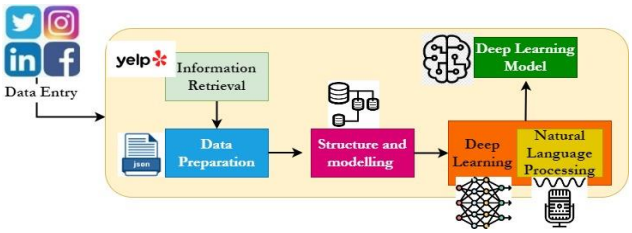


Figure. 2 First phase architecture of SMDA-DL framework

this step. These rely on some computing tools designed expressly for natural language problems. As a result, when NLP is combined with a deep learning algorithm, the highest predictive outcomes are possible. Data analysis using DL and NLP has many applications. Text mining, recommendation engines, and review score assessment are three of the most important. This step produces a mechanism for performing data analytics that combines DL algorithms with NLP software. The characteristics of NLP technique are Cryptocurrency, named entity identification, part-of-speech tagging, dependent decoding, tokenizing & originating, lemmatization elimination, word sense disambiguation, and domain querying.

Step 6: a model of deep learning

The final part of this step produces a method that integrates deep learning and natural language processing methods to analyze textual information; this approach can serve as the foundation for a deep learning model (DLM). The following factors are considered in this framework: (i) the system can be used to analyse pre-existing data to extract information, such as user preferences, extraction of features, or sentiment classification. The model's performance in forecasting results may vary depending on the DL and NLP approaches employed in the model; text mining is an ever-growing discipline that includes optimizing DL algorithms and incorporating many other approaches such as arithmetic, economics, and simulation methods. ii) The system is designed to run on a non-BD structure, which means that big data can be examined without requiring high-performance capabilities. Removing the barrier of needing a highly advanced computing infrastructure allows for more dependable results to be produced with higher performance. iii) Because the second step requires BD analyses, this framework can serve as the foundation for the DL step. As a result, the DLM might be used to analyze BD in its entirety or partly.

Fig. 2 shows the first phase architecture of the social media data analytics using the deep learning (SMDA-DL) framework. It consists of six steps:

information retrieval, data preparation, structure and modelling, DL, NLP, and the deep learning model explained earlier.

- Second phase

As mentioned earlier, BD is a hot topic in technology, industry, and development. Using data gathering, it is possible to get valuable insights from vast volumes of data that would otherwise be impossible to obtain. Major IT businesses such as Facebook, eBay, and Instagram are examples of how strategic outcomes may be produced with the collaboration of academics devoted to BD evaluation, which can be a deciding factor in gaining a competitive edge. A set of layers targeted at constructing a BD infrastructure capable of performing BD analyses of large unorganized datasets is suggested for the architecture's second phase.

Layer 1: distributed filing systems

BD architecture is described by the distributed filing system (DFS) layer. Customers of spatially parallel computing can communicate their information and resources through utilizing a common file system owing to the distributed file system (DFS), which was designed for this function. Network participants can utilize many resources as if they were locally memory because distributed file networks can transfer information from one computing device across multiple computers. It gives businesses easy, flexible, safe, and practical data accessibility. The key resources of distributed systems are planning phase, extensibility, disclosure, utilization of shared resources, interoperability, and patience of errors. It suggests some different BD goods and services. It was concluded after reviewing the application that one of these solutions could be incorporated into this layer. The spark strategy is developed as the layer's centre because it can offer the next stage of development with the essential assistance to perform statistical analysis at higher levels. Data analysts all across the major difference Spark, an appraisal system, for big data analysis. It can handle both batches or streamed data processing and is developed on top of Hadoop. Spark is a BD managing architecture that has been utilized in previous research for various academic and industrial purposes, including text data collection. It's worth noting that cloud service solutions might be included in this tier. Although google cloud platform and online storage data processing support many spark applications, their pay-as-you-go models are prohibitively expensive for small businesses. They give a meta-analysis of 295 papers covering these subjects and how they connect to the marketing and corporate domains, revealing how information technology has been used in these disciplines and the issues its experts are experiencing.

Layer 2: cluster resources management

A cluster resources management (CRM) is necessary if the DFS is deployed locally to supervise the task dispersed across the nodes. The capabilities allocated to a community reserve must be mapped to specific cluster members by a clustered managing resource. The mappings are significant positive according to the current load circumstances and therefore is independently of the cluster's demand distribution model. A cluster is a group of machines linked by a network and simultaneously completes multiple tasks. The clustering connects several machines, allowing them to use their processing capabilities to solve problems while offering error detection because others can readily substitute failed nodes. Because BD approaches are designed to do parallel and cloud services, it is proposed that a group be employed to boost processing power. Resources management is a services management that runs jobs, analyzes node efficiency and functions, and organizes machines. Apache Mesos is suggested as the CRM if the group is installed in the local computing environment. An open-source network agent called Apache Mesos manages tasks in a multiuser environment by dynamically distributing capital and isolating them. Programs may be deployed and managed with Mesos in large-scale heterogeneous environment. Mesos is made up of Mesos modules, which execute activities on these clients, as well as a supervisor computer that controls agent daemons operating on each cluster node. By submitting resources requests, the supervisor allows fine-grained data exchange among modules (CPU, RAM, etc.). The study evaluated many CRM and found Mesos to be the most effective overall. While building up a local group can take a lot of time and work, cloud technology has many benefits because services can quickly and efficiently set up groups.

Layer 3: data access

Now that the dispersed and coordinated levels have been defined database may determine ways to manage access to data that is loaded to terminals later. The data model is under the jurisdiction of the BD platform. It supports various languages for access to information, including Java, Php, R, and NoSQL, among many others. The information in the internal

structure results from this level can be analyzed using BD approaches in the next interface.

Layer 4: data analysis

The preceding three layers were responsible for creating a BD structure that could retrieve current information and not conduct research. As a result, the business intelligence layer is included in the design. Because this layer is where most BD analysis takes place, the study recommended using deep learning techniques seen in many BD platforms to do such analysis. When building a DLM, the framework recommends looking at massive amounts of textual data in a non-BD setting. This method is now anticipated to work at this level after being included in a BD analysis job that employs a distinct DL program. This strategy has a drawback: special caution must be exercised when employing deep learning. Deep learning in the chosen BD system may not be compatible with the DLM methods in the first stage. Because BD technology is still evolving, several deep learning methods for BD-like operations are currently being developed. Each component can utilize a single-node training library to resolve this issue. In addition, an NLP module is included in this level to conduct various text data processing as needed.

Layer 5: visualization

The preliminary stage outputs a set of values that reflect the behavior being evaluated, typically determined at the start of the study. Visualization tools can be required, allowing more understandable data analysis on a broad scale to depict these values. A set of approaches for deeper comprehension and getting insights from information is done as data visualization. It can be used as a service software, explaining and presenting the information. Because this layer is concerned with constructing state-of-the-art graphs that enable quick analysis of a vast quantity of information easily and flexibly by visualizing and engaging with information, graphical analytical methods provide participants with decision-making examples. This layer produces tools for dynamically visualizing vast amounts of text data.

3.2 Sentiment analysis

The social media text is analyzed in this section. The special symbol recognition and analysis based on the deep learning method is proposed. Sentiment analysis, often referred as information extraction, is a natural language processing (NLP) method for identifying the positivity, negativity, or neutrality of information. Companies typically do text analytics on text documents to track the perception of their merchandise and brands in user reviews and to help explain their target market. Annotation assessment results can be compared to determine the correctness. Nevertheless, employing TP (true positive) and FP (False Positive) in addition to high level of performance F-measure will also be helpful.

Embedding with two senses

Recent research has demonstrated that word embedding activities like word2vec and fast text are very successful. It uses fast text to create emoji descriptors by treating each emoji as a unique word with its own set of deep learning. The capacity to acquire and interpret information from social media platforms to encourage business choices as well as track the effectiveness of steps taken in response to such choices by social media is referred to as social media influencers. The catch is that, unlike traditional approaches in which each emoji reacts to a single embedding matrix, it integrates each symbol into two separate matrices (bi-sense symbol encoding). First, it allocates each emojis to two separate symbols, one for usage in favourable emotional settings and the other for usage in unfavourable sentiment situations. Each token is embedded into a separate vector using the identical fast text training method, yielding favourable and unfavourable sensation descriptors for each symbol. The skip-gram method founded on the word2vec aims to maximize the likelihood function by adding the possibilities of presenting adjacent words given a collection of meanings. One unsupervised learned method utilized to discover the phrases that are most connected to a lexical item is the skip-gram. To determine the background phrase for a single target phrase, skip-gram is employed. It is the CBOW method inverted. Here, the targeted phrase is entered and the surrounding syllables are produced. When provided a recent term, the continuous skip-gram algorithm achieves by anticipating the syllables that will be around it. To put it another way, the continuous skip-gram theory foretells keywords that will appear preceding and following the present syllable in the identical phrase inside a required area. The continuous bag of words makes word predictions given nearby background. The fastest model differs because it presents the issue as a binary classifier. The goal is to anticipate the appearance of each contextual word with unfavourable examples drawn at arbitrary from the language model that is missing. The objective of the deep learning algorithm is expressed in Eq. (1).

$$L(w_t, w_c) = \sum_{t=1}^{T}\left(\sum_{w_c=1}^{W_{c_t}} L\big(s(w_t, w_c)\big) - \sum_{w_n=1}^{W_{n_t}} L(s(w_t, w_n))\right) \quad (1)$$

As shown in Eq. (1) objective of the deep learning algorithm has been expressed. The target model is designed using logistic regression loss depending on an input word sequencing $\{w_1, w_2, \cdots, w_T\}$.The initial term $L(w_t, w_c)$ represents the sum of losses L for word pairs (w_t, w_c).The contextual term set W_{c_t} and the collection of unfavourable word sampling W_{n_t} of the previous frame w_t . Where $L(s(,))$ is the logistical losses of the scoring variable $s(,)$ which is calculated by adding the scalar components of the present word's n-gram descriptors and the interpretation sentiment analysis, unlike word2vec, where the scoring is the vector product of the previous frame and the contextual sentiment analysis. Fast text is chosen over word2vec because of its computational time. On average, the two models produce comparable results, and an evaluation of word embeddings is outside the scope of this paper. In conclusion, it solely displays the effectiveness of fast text activation.

Word dictionary based on LSTM

Values represent been encoded using long short-term memory (LSTM) units frequently. A text learning algorithm, LSTMs texture, and fully-connected levels for additional functions such as text classes depending on the recorded information make up the basic encoding model. Eq. (2), Eq. (3), Eq. (4), Eq. (5), Eq. (6), Eq. (7) expresses the processes in an LSTM unit for time step t:

$$i_t = \sigma(W_i X_t - U_i - b_i) \quad (2)$$

$$f_t = \sigma(W_f X_t - U_f h_t - b_f) \quad (3)$$

$$o_t = \sigma(W_o X_t - U_o h_t - b_o) \quad (4)$$

$$g_t = tanh(W_c X_t - U_c h_t - b_c) \quad (5)$$

$$c_t = f_t \cdot c_t \cdot i_r \cdot g_t \quad (6)$$

$$h_t = o_t \cdot tanh(c_t) \quad (7)$$

As found in Eq. (2), Eq. (3), Eq. (4), Eq. (5), Eq. (6), Eq. (7) LSTM process has been identified. Where h_t represents the intermediate feature variables, X_t are the existing LSTM data and the encoding W_t of the existing phrase W_t is used here, and W and U indicate the weighting vectors. It changes the input level into the LSTM units to use the bi-sense symbol encoding. The LSTM variables are denoted as $i_t, f_t, o_t, g_t, c_t and\ h_t$. Emojis have numerals allocated to them according to the Unicode Specification. This is the way it goes. Every emoticon is referenced by a valued (a hexadecimal integer) in the Unicode Standards, such as U+1F063 for illustration. The senti-emoji encoding is initially calculated as a balanced mean of the bi-sense symbol encoding based on self-attentiveness.

The encoded symbol, weight of the symbol, and the senti matrix are expressed in Eq. (8), Eq. (9), Eq. (10):

$$u_{t,i} = f_{att}(e_{t,i}, W_t) \quad (8)$$

$$\alpha_i = \frac{exp(u_{t,i})}{\sum_{i=1}^{m} exp(u_{t,i})} \quad (9)$$

$$v_t = \sum_{i=1}^{m} \alpha_{t,i} \cdot e_{t,i} \quad (10)$$

As calculated in Eq. (8), Eq. (9), Eq. (10) encoded symbol, the weight of the symbol and sentimental matrix has been computed. Let $e_{t,i} \in (1, m)$ symbolize the i-th sense encoding of the symbol e_t (m = 2 in the bi-sense encoding), and $f_{att}(., W_t)$ signify the attentive variable based on the current deep learning, the attentiveness weighting α_i and the senti-emoji inserting matrix v_t . It uses a fully connected level with ReLU activity as the concentration mechanism, and it combines the interest matrix v_t with the word encoding as the LSTM's given configuration. As a result, in Eq. (2), the input data x_t yields $[W_t, v_t]$. The outcome of the last LSTM module is then inputted into a complete layer with sigmoid activated to produce the Twitter emotion, with bipolar bridge loss serving as the objections response, with *N* being the entire number of iterations.

The idea for this method is that each contextual word directs the attentiveness weighting causing the machine to self-select the embedded sensation this can focus. As a result, this system is referred to as the Word-guide Interest LSTM with Bi-sense symbol encoding. The interest of the symbol is denoted in Eq. (11)

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} y_i\, log\,(p_i) + (1 - y_i) log\,(p_i) \quad (11)$$

As explored in equation (11) interest in the symbol has been derived. The senti-emoji is denoted as y_i and the probability of the symbol is denoted as p_i. The total number of iterations is denoted as N.

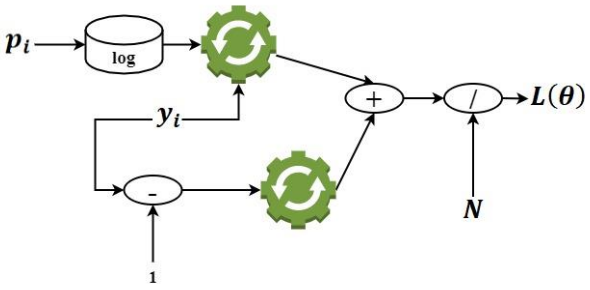


Figure. 3 Pictorial representation of L(θ)

Fig. 3 shows the pictorial representation of $L(\theta)$. The senti-emoji yi and the probability of the symbol p_i and the total number of iterations used for symbols N is used to calculate the variable $L(\theta)$ is denoted as the interest of the senti-emoji.

LSTM with multi-level consideration

Another way to express the feature representations is to use attention weighting to describe how picture information is dispersed through word vectors, as recommended. By substituting W_t with the ultimate state matrix h transferred from the last LSTM module, the revised senti-emoji encoding matrix v is at the tweet-level rather than the phrase in Eq. (12), Eq. (13):

$$\alpha_i = \frac{exp(f_{att}(e_i,h))}{\sum_{i=1}^{m} exp(f_{att}(e_i,h))} \quad (12)$$

$$v' = \sum_{i=1}^{m} \alpha_t \cdot e_i \quad (13)$$

As initialized in Eq. (12), Eq. (13) sentiment encoding matrix has been described. The attentiveness weighting is denoted as $\alpha_i$and the senti-emoji inserting matrix is expressed as v'. The senti-emoji is expressed as e_i and the state of the matrix is denoted as h. The attenuation function is denoted as f_{att}.

The added layer of attentiveness following is calculated using the derived senti-emoji encoding v'. The attentiveness weighting $\alpha'_t, t \in (1,T)$ condition on the senti-emoji encoding is defined as the following assuming the input Twitter sequences $\{w_1, w_2, \cdots, w_T\}$ is expressed in Eq. (14):

$$a'_t = \frac{exp(f_{att}(W_t,v'))}{\sum_{i=1}^{m} exp(f_{att}(W_t,v'))} \quad (14)$$

As deliberated in Eq. (14), input twitter sequences have been discussed. The attenuation variable is expressed as f_{att}. The word sequence is expressed as W_t and the senti-emoji inserting matrix is expressed as v'.

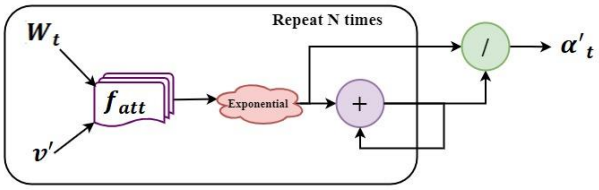


Figure. 4 Pictorial representation of α'_t

Fig. 4 shows the pictorial representation of α'_t. The attenuation variable f_{att}, The word sequence W_t and the senti-emoji inserting matrix is v' are utilized to evaluate the variable α'_t . To redistribute the senti-emoji data to each phase, it creates a new entry u_t for each LSTM block by combining the original word encoding and the attentiveness matrix. Multi-level Media exposure LSTM with Bisense Emoji Encoding is the name of this system. It uses the same linear cross-entropy with similar network architecture as the gradient descent is expressed in Eq. (15).

$$u = [W_t, \alpha'_t, v'] \quad (15)$$

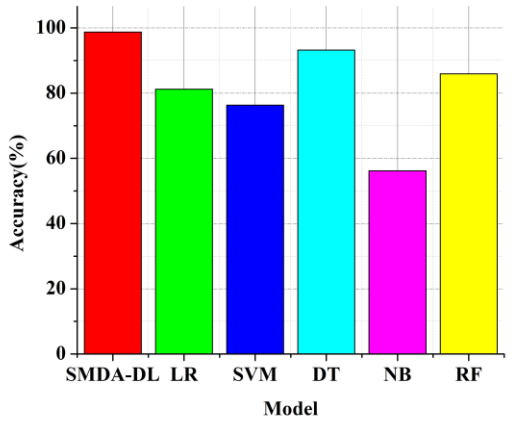
As demonstrated in Eq. (15) gradient descent has been evaluated. The word sequence is denoted as W_t.The encoded senti-emoji is denoted as α'_t and the inserted social media data matrix is denoted as v'. In this way, the emojis and texts used in the social media data are analyzed in the proposed SMDA-DL framework.

The proposed SMDA-DL achieves high accuracy, enhances recall, and increases f-score and precision based on deep learning techniques.

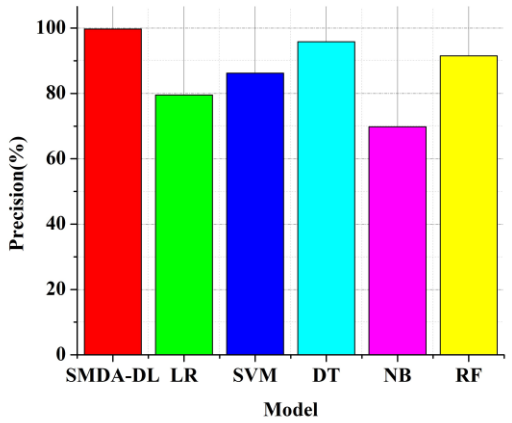
4. Software analysis and performance evaluation

Data and methods

Due to privacy concerns, most businesses refuse external researchers access to their information. Other state-of-the-art technologies use Web scraping methods to extract publicly available data. The practice of deploying computers to gather information and material from a webpage is referred to as scraping. Web scraping collects the fundamental HTML code as well as, with it, information kept in a computer, in contrast to data extraction, that just scrapes images seen immediately. After that, the scrape can duplicate a whole web page somewhere. Furthermore, the permissions to handle data received from social media sites using these methods are unclear. The Yelp Corporation makes a dataset available for educational use. Yelp is a worldwide corporation that operates Yelp.com and the Yelp



(a)



(b)

Figure. 5: SMDA-DL framework: (a) Accuracy analysis and (b) Precision analysis

application, which broadcasts ratings of local establishments and offers Yelp booking systems. The Yelp database has been utilized in several research papers. The Yelp database is a portion of Yelp's listings, comments, as well as user information that has been rendered accessible to the general public to be utilized in private, scholarly, and instructional contexts. Utilize it to educate children about networks, practise NLP, or utilize example performance data while they study how to create phone devices. It is accessible as JSON documents. Yelp is significant since it is quickly taking over as the premier website for customer feedback of several different kinds of companies and that, according to my studies, it consistently ranks well in the Google Search Engine outcomes. It is broadly acknowledged in computing. The items included in the Yelp database did not agree with the JSON specification. A text-based convention for encoding organized statistics based on JavaScript object grammar is entitled JavaScript object notation (JSON). It is frequently employed for data

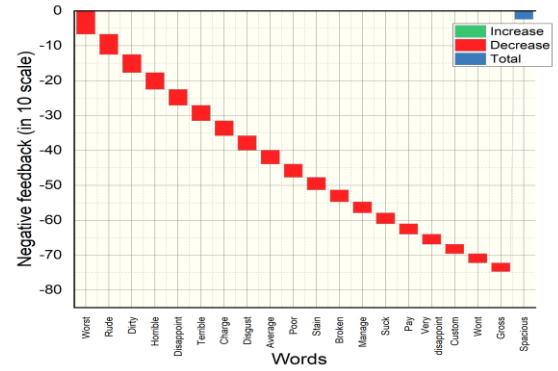
Table 1. Performance analysis of the social media data analytics using deep learning (SMDA-DL) framework

Method	F1 score (%)	Recall (%)
SMDA-DL	98	98.4
RF	90.9	91
SVM	93.4	88
LR	85	80
MNB	78	79
DT	81	73.3

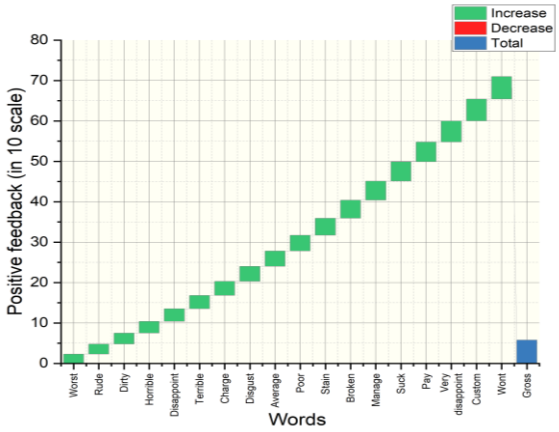
transmission in online apps (For instance, transmitting the client's information to the host so that it may be published on a web page, or the opposite). It was required to write a copy of the program and fix them. After reading the documents' contents, it was decided that they should be exported to a database system for improved data collection. Python and R are two languages used to perform deep learning. Python was chosen as the computer language due to familiarity, and R as a secondary environment for DL development is encouraged. The greatest often utilized open-source Python library for data scientists, data processing, and network learning opportunities is called Pandas. It is constructed on top of Numpy, a different program that supports multi-dimensional matrices. Libraries like NLTK (natural language toolkit) and Pandas (statistical analysis and assessment) were employed for the NLP work and data preparation. Lastly, precision and recall testing can be used to examine the data produced. https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset

Fig. 5 (a) and (b) show the accuracy and precision analysis of the social media data analytics using the deep learning (SMDA-DL) framework using dataset. As shown in Eq. (1) objective of the deep learning algorithm has been expressed for precision analysis. The simulation is carried out by considering the existing methods, RF [15], SVM [15], LR [17], MNB [22], and DT [23], compared with the proposed social media data analytics using the deep learning (SMDA-DL) framework. This study of social media data is reviewed and displayed to show its accuracy and precision. Because of the emoji analysis function, the presented framework offers the best results.

Table 1 shows the performance examination of the social media data analytics using the deep learning (SMDA-DL) framework based on the dataset [10]. The simulation tool runs the Yelp dataset through the current and planned frameworks. Tables like F score and Recall are used to record simulation results like this.; because of the inheritance of the senti-emoji analysis and big data analytics method, the proposed SMDA-DL



(a)



(b)

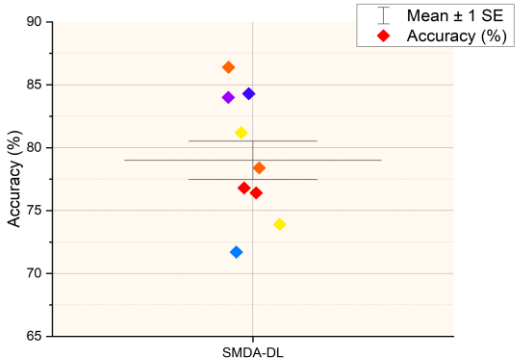
Figure. 6 SMDA-DL framework: (a) Negative feedback analysis, and (b) Positive feedback analysis

framework has the increased performance with other models.

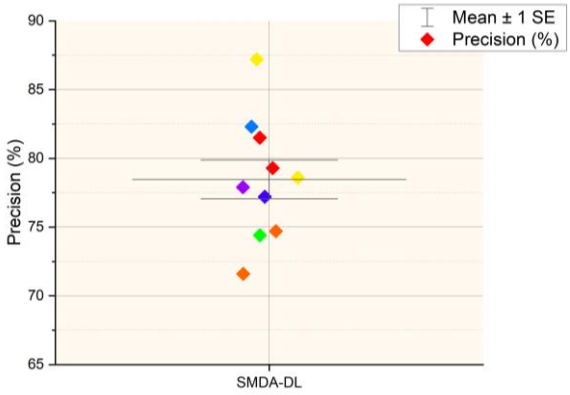
Fig. 6 (a) and (b) depict the negative feedback and positive feedback analysis of the social media data analytics using the deep learning (SMDA-DL) framework based on the dataset [10]. As calculated in eq. (3.1, 3.2, 3.3) encoded symbol, the weight of the symbol and sentimental matrix has been computed for the positive feedback. The social media data is examined based on the Yelp dataset. The terms used the most to provide information about the product that has been bought or used are studied, and the use level of those words is displayed on a scale of 10 in the above figures. The findings indicate that the system that has been presented analyzes all of the complicated data and generates very excellent results with the assistance of senti-emoji analysis and the approach of big data analytics. As shown in Table 1 and Table 2 the effectiveness of proposed model shows a higher result with the metric namely accuracy, precision, recall and f1-score as 98.7%, 99.7%, 98.4% and 98% compared to other existing techniques taken from the research survey analyzed,

Table 2. Performance analysis of SMDA-DL framework

Method	Accuracy (%)	Precision (%)
SMDA-DL	98.7	99.7
RF	81.2	79.5
SVM	76.3	86.2
LR	93.2	95.8
MNB	56.2	69.8
DT	85.9	91.5



(a)



(b)

Figure 7 SMDA-DL framework: (a) Accuracy analysis and (b) Precision analysis of SMDA-DL framework

the second highest performance is LR with accuracy as 93.2% with comparison data is shown in Fig.8

Table 2 shows the performance analysis of the social media data analytics using the deep learning (SMDA-DL) framework based on the dataset [10]. An accurate and precise simulation result assessment of both the conceptual methodology and existing systems, such as naïve Bayes, decision trees, SVMs, and randomized forests for the computational analysis. Emoji analytics and big data analytics are both ways that contribute to the conceptual methodology for SMDA-DL, which the outcome suggests delivers superior quality findings.

Fig. 7 (a) and (b) show the accuracy analysis and the precision analysis of the proposed social media

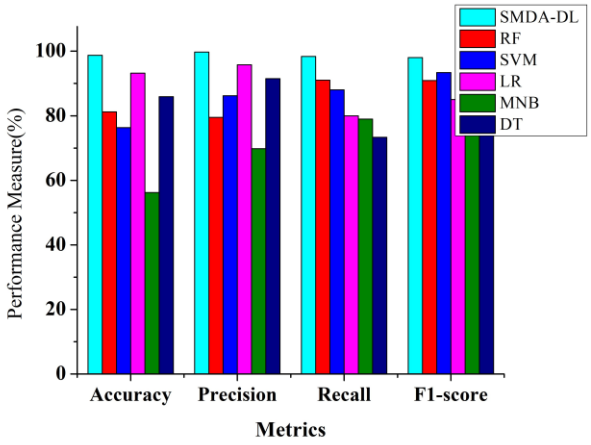


Figure. 8 Performance measure analysis

data analytics using the deep learning (SMDA-DL) framework based on the dataset [10]. As demonstrated in Eq. (7) gradient descent has been evaluated to address the accuracy of the proposed method. The number of simulation iterations for the study ranges from 1 to 10. The accuracy and precision of the suggested framework are examined, and their results are displayed in the figures shown above. The findings demonstrate that the performance of the proposed framework is not correlated with the number of iterations performed to ensure accuracy and precision. Therefore, the simulation analysis may use any number of iterations the user chooses. The proposed social media data analytics using the deep learning (SMDA-DL) framework is considered and instigated. The outcomes demonstrate that the proposed SMDA-DL framework has the highest performance because of the big data analytics method and senti-emoji classification method.

The effectiveness of the SMDA-DL performance is compared with the existing approaches like RF [15], SVM [15], LR [17], MNB [22], and DT [23], in terms of evaluation metrics namely accuracy, precision, recall and f1-score is shown in Fig. 8.

Comparison of the proposed model with the existing approaches are shown in the Fig. 5 (a) and (b) in terms of accuracy analysis and precision metric. Fig. 5 (a) details the comparison of proposed SMDA-DL approach with the existing algorithms to evaluate the accuracy and produces analysis as RF (81.2), SVM (76.3), LR (93.2), MNB (56.2), and DT (85.9) comparatively the SMDA-DL produced higher accuracy as 98.7% and performs well in the way of emoji and big data analytics. Similarly, Fig. 5 (b) produces RF (79.5), SVM (86.2), LR (95.8), MNB (69.8), and DT (91.5) comparatively the SMDA-DL produced higher accuracy as 99.7%. Fig. 8 illustrates the comparison study of proposed algorithm with existing approaches and proves the efficacy using the analysis of emoji function from social media.

5. Conclusion and future scope

The Social media data analytics using the deep learning (SMDA-DL) framework has the essential features such as: (i) it's a two-stage process; the first is associated with analyzing the product ready and determining the best DLM for data processing, while the other is involved with creating a BD architecture capable of doing analysis and visualization activities. (ii) Compared to other structures, the first stage components have been shortened to obtain outcomes in the quickest possible by lowering complications and enabling the incorporation of new areas and methods. (iii) The DLM is tied to specific deep learning techniques. It improves the framework's information by allowing current DL systems to be used to compare outcomes. The data accessible in this research is conducted to the Yelp database, and the findings displayed are restricted to the first phase. Not all of the strategies used in the first stage are readily adaptable to advanced analytics methodologies. Because the deep learning available for textual data in BD and non-BD contexts does not completely correlate, future research can focus on applying the outcomes of the first phase to the latter using BD-specific DL methods and methodologies. The proposed framework yields good simulation results in an accuracy ratio of 98.7%, a precision ratio of 99.7%, a recall ratio of 98.4% and an F score of 98%. Future studies should be free to look into improving the architecture using modern deep learning approaches. Furthermore, it is suggested to broaden the study by examining the distinctions among the two computational models (big data and non-big data) to determine why these variations occasionally limit the employment of the same methodologies.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, first, second and third authors; methodology, first, second and third authors; software, first and third authors; validation, first and second, authors; formal analysis, first and second authors; investigation, third author; resources, second and third authors; data curation, third author; writing original draft preparation, third author; writing review and editing, first, second and third authors;

visualization, first and second author; supervision, first author; project administration, third author; funding acquisition, first, second and third authors.

References

[1] M. Abiad, Kadry, S. Ionescu, and A. A. Niculescu "Customers' perception of telecommunication services", *FAIMA Business & Management Journal*, Vol. 7, No. 2, pp. 51-62, 2019.

[2] A. Alazab, S. Bevinakoppa, and A. Khraisat, "Maximising competitive advantage on E-business websites: A data mining approach", In: *Proc. of 2018 IEEE Conference on Big Data and Analytics (ICBDA)*, pp. 111-116, 2018.

[3] M. H. Ali and M. F. Zolkipli, "Review on hybrid extreme learning machine and genetic algorithm to work as intrusion detection system in cloud computing", *ARPN Journal of Engineering and Applied Sciences*, Vol. 11, pp. 460-464, 2016.

[4] V. E. S. C. Shin and Y. Cho, "Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city", *Building Research & Information*, Vol. 49, No. 1, pp. 127-143, 2021.

[5] G. Manogaran, M. Alazab, V. Saravanan, B. S. Rawal, P. M. Shakeel, R. Sundarasekar, and C. E. M. Marin, "Machine learning assisted information management scheme in service concentrated IoT", *IEEE Transactions on Industrial Informatics*, Vol. 17, No. 4, pp. 2871-2879, 2020.

[6] M. I. A. Mashhadani, K. M. Hussein and E. T. Khudir, "Sentiment analysis using optimized feature sets in different facebook/twitter dataset domains using big data", *Iraqi Journal for Computer Science and Mathematics*, Vol. 3, No. 1, pp. 64-70, 2022.

[7] T. N. Nguyen, S. Zeadally, and A. B. Vuduthala, "Cyber-physical cloud manufacturing systems with digital twins", *IEEE Internet Computing*, Vol. 26, No. 3, pp. 15-21, 2021.

[8] L. Murry, R. Kumar, T. Tuithung, and P. M. Shakeel, "A local decision making technique for reliable service discovery using D2D communications in disaster recovery networks", *Peer-to-Peer Networking and Applications*, Vol. 13, pp. 1131-1141, 2020.

[9] A. K. Luhach, S. K. Dwivedi, and C. K. Jha, "Implementing the logical security framework for E-commerce based on service-oriented architecture", In: *Proc. of International Conference on ICT for Sustainable Development: ICT4SD 2015*, Springer Singapore, Vol. 2, pp. 1-13.

[10] Y. Zhao, Y. Yu, P. M. Shakeel, and C. E. M. Marin, "Research on operational research-based financial model based on e-commerce platform", *Information Systems & e-Business Management*, 2021.

[11] A. Alazab, S. Bevinakoppa, and A. Khraisat. "Maximising competitive advantage on E-business websites: A data mining approach", In: *Proc. of 2018 IEEE Conference on Big Data and Analytics (ICBDA)*, pp. 111-116, 2018.

[12] S. Dhote, C. Vichoray, R. Pais, S. Baskar, and P. M. Shakeel, "Hybrid geometric sampling and AdaBoost based deep learning approach for data imbalance in E-commerce", *Electronic Commerce Research*, Vol. 20, pp. 259-274, 2020.

[13] M. Balaanand, N. Karthikeyan, and S. Karthik, "Envisioning social media information for big data using big vision schemes in wireless environment", *Wireless Personal Communications*, Vol. 109, No. 2, pp. 777-796, 2019.

[14] N. Shah, D. Willick, and V. Mago, "A framework for social media data analytics using Elasticsearch and Kibana", *Wireless Networks*, pp. 1-9, 2022.

[15] T. Kanan, A. Mughaid, R. A. Shalabi, M. A. Ayyoub, M. Elbes, and O. Sadaqa, "Business intelligence using deep learning techniques for social media contents", *Cluster Computing*, Vol. 26, No. 2, pp. 1285-1296, 2023.

[16] X. Liu, H. Shin, and A. C. Burns, "Examining the impact of luxury brand's social media marketing on customer engagement: Using big data analytics and natural language processing", *Journal of Business Research*, Vol. 125, pp. 815-826, 2021.

[17] I. M. Haile and Y. Qu, "Mitigating Risk in Financial Industry by Analyzing Social-Media with Machine Learning Technology", *European Journal of Electrical Engineering and Computer Science*, Vol. 6, No. 2, pp. 33-37, 2022.

[18] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics–Challenges in topic discovery, data collection, and data preparation", *International Journal of Information Management*, Vol. 39, pp. 156-168, 2018.

[19] J. Q. Dong and C. H. C. H. Yang, "Business value of big data analytics: A systems-theoretic approach and empirical test", *Information & Management*, Vol. 57, No. 1, pp. 103-124, 2020.

[20] J. Pringle and S. Fritz, "The university brand and social media: Using data analytics to assess

brand authenticity", *Journal of Marketing for Higher Education*, Vol. 29, No. 1, pp. 19-44, 2019.

[21] A. Subroto and A. Apriyana, "Cyber risk prediction through social media big data analytics and statistical machine learning", *Journal of Big Data*, Vol. 6, No. 1, p. 50, 2019.

[22] A. Alsanad, "An Improved Arabic Sentiment Analysis Approach using Optimized Multinomial Naïve Bayes Classifier", *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 8, 2022.

[23] D. P. Ramadhani, P. M. Wijaya, and A. Alamsyah, "Credit Scoring Model Construction Based on LinkedIn Social Media Data", In: *Proc. of European International Conference on Industrial Engineering and Operations Management*.