International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 5, pp 362-376, May-2023

https://doi.org/10.47001/IRJIET/2023.705054

Data Mining-Based Transaction Labelling Enhancing Financial Insights through Automated Technique

Praneeth Reddy Amudala Puchakayala

Data scientist, Regions Bank, USA. E-mail: appraneethreddy01@gmail.com

***Abstract* - During the most recent era, it has been suggested that various accounts be maintained that preserve the application of natural language text inputs that are submitted as the financial transactions (debit and credit entries) that are automatically generated are the inquiries and responses to specific queries. Machine learning serves as the basis for AI-powered categorization in the process described above. This categorization method makes use of deep learning, natural language processing procedures, and other techniques to improve its algorithms for conducting the operations. It is necessary to employ data mining techniques to carry out predictive modelling that is founded on data analytics. The process of machine learning involves the existence of a vast array of data kinds. Here are some ways that use natural language processing to carry out transaction processing based on data description. Following that, several real-time applications that are special to the financial sector were addressed, each of which was based on a different technique that was being implemented. As a result, a variety of problems and potential solutions are addressed based on the performance of the various methodologies achieved through the utilization of various performance measures.**

Keywords: Bank Transaction; Machine Learning; Data Analytics; Financial; Prediction.

I. Introduction

Query Response Systems, automatic text summarization, sentiment analysis, topic extraction, text mining, named entity recognition, parts-of-speech tagging, machine translation, relationship extraction, automated question answering, and numerous other applications employ natural language processing (NLP). The authors of the paper have suggested an account that maintains the application of natural language text inputs that are posted as The financial transactions (debit and credit entries) that are automatically generated are the queries and responses to individual queries [1].

Machine learning serves as the foundation of AI-powered categorization. AI also employs deep learning, NLP procedures, and other techniques to enhance its algorithms. This enables the data to be processed efficiently in order to accurately categorize transactions. AI transaction categorization integrates data analysis, natural language processing (NLP), and machine learning algorithms. Transaction data is collected to initiate the procedure. ISO standardized transaction message formats to facilitate global financial communication. In this instance, ISO 20022 serves as the universal standard, while ISO 8583 is implemented for card transactions. These facilitate the classification of specific transactions. The process of AI transaction categorization is intricate; therefore, it is imperative to examine it in detail.

[2] To classify bank transactions utilizing poor supervision, natural language processing, and deep neural network approaches. Heuristics and domain knowledge are employed to train precise transaction classifiers, thereby reducing the necessity for costly and challenging-to-obtain manual annotations. We propose an efficient and scalable end-to-end data pipeline that includes data preprocessing, transaction text embedding, anchoring, label creation, discriminative neural network training, and an overview of the system architecture.

A) Processing and collection of data

The process commences with the accumulation of data, which is subsequently processed by the algorithms into a structured format that AI models can comprehend and learn from. Transaction data is gathered from a variety of sources, including financial applications, payment processors, and institutions. The transactions contain information such as the transaction quantity, date, and description.

- Data Transaction classification:

Being able to offer loans to the appropriate candidates is one of the main issues that financial organizations deal with. Our work relies heavily on labelling bank transactions, which is only getting more difficult as the amount and complexity of bank transactions rise. Classifying bank transactions effectively is difficult, even given the abundance of insights that can be gained from bank data. Traditional approaches, such as manual tagging or rule-based systems, are insufficient in the face of rising transaction volume and complexity. This results in low label quantities and biased datasets of transactions that are easily tagged. Being able to offer loans to

ISSN (online): 2581-3048
Volume 7, Issue 5, pp 362-376, May-2023
https://doi.org/10.47001/IRJIET/2023.705054

the appropriate candidates is one of the main issues that financial organizations deal with. It reviles heavily on labelling bank transactions, which is only getting more difficult as the amount and complexity of bank transactions rise.

The increasing utilization of machine learning algorithms, such as deep neural networks, facilitates the classification of transactions; however, a substantial obstacle persists: the absence of training labels and frequently obscure transaction descriptions that impede the acquisition of more profound insights. A scalable system is therefore required, one that can manage substantial amounts of unlabeled transactional data while maintaining precise classifications.

Based on the poorly supervised, extensible transaction classification method to address the issue. It combines deep neural networks, noise-aware label generative models, and unsupervised transaction text embeddings. Even without labelled data, you can use the probabilistic labels to train strong discriminative models, such as deep neural networks, for transaction classification and achieve results that are on par with or better.

Next, the transaction data undergoes a cleaning process. Information may be absent from certain transactions. The preprocessed data is transformed into a normalized format by algorithms. The category of the transaction is determined by the transaction descriptions. The descriptions are divided into individual sentences or tokens in the case of textual data. the data is transformed into numerical vectors by AI, which enables the AI models to learn from it [3]. The procedure entails the addition of specific information to the transaction data, which fills in the gaps and facilitates identification. It is referred to as data enrichment, and it enables the addition of the necessary information to the original transaction data to enhance the consistency and accuracy of transaction categorization.

B) Model training and implementation

In traditional Machine Learning, AI acquires the data from which to learn, enabling the models to train on it in order to categorize transactions. In order for a classic ML to function, the data is divided into several sets that require labeling: training, validation, and testing. The category can be assigned to a transaction (training set) manually to complete the categorization.

The model is trained on a training set that should contain examples of transactions with appropriate categories. Historical transaction data that was painstakingly categorized can be utilized by models to acquire knowledge. This enables it to identify patterns for particular transaction categories. Next, the model's efficacy is assessed and validated. The model is evaluated for efficacy and, if necessary, enhanced. Upon satisfactory performance, the model may be implemented [4].

In order to retain and expand their consumer base, financial institutions must devise novel strategies. Their customer behavior has evolved, and their product portfolios have diversified over the years shifted from long-term loyalty to online interaction. As a result of the intense competition among banks, there is an increasing demand for the conversion of consumer data, which includes shorttext banking transaction (BT) descriptions, into information that is pertinent for decision-making.

In finance, data mining has been effectively implemented in a variety of ways, including:

- Identification of plausible candidates for loan disbursement.
- Product acceptance.
- Characterization of product segments.
- Analysis of customer attrition and retention.

Nevertheless, the associate editor who coordinated the evaluation of this manuscript and approved it for publication, to the best of our knowledge, was Mohamad Forouzanfar.

The issue of the automatic classification of short-text BT descriptions in accordance with a predetermined set of labels has not yet been adequately addressed [5].

Due to the abundance of publicly accessible digital text sources, automated text classification has emerged as a prominent research area from a broader perspective. There are numerous applications for text classification, including event detection [7], opinion mining [6], and web searching [5]. However, the majority of text classification methods are applicable to lengthy texts. The following are some of the unique characteristics of brief texts:

- Sparsity:

Short messages are typically composed of a limited number of sentences and contain fewer than 150 words. They communicate a minimal amount of information that is effective. Traditional techniques, such as those employed for long texts, are impractical due to the fact that sparsity impacts the quality of brief text semantics. This is due to the difficulty of extracting key features from large feature spaces for precise classification training [8], [9].

- Real-time generation:

In the present day, an immense volume of information is generated in the form of brief messages that are incessantly transmitted. For instance, news comments, conversation and microblog information, and so forth.

They are challenging to capture because they reflect real-time reactions to external events. As a result, short-text classification methods must be exceedingly efficient. Unpredictability: Vocabularies are informal or specific, and short-text terminology is not standardized (in our case, it is related to banking). Words are rarely repeated in a given BT description [6], and only a small number of words are irrelevant. These are two critical factors. The significance of a word cannot be solely determined by its frequency of occurrence in the text. Nevertheless, short messages are less noisy than long texts for the same reasons.

II. Data Mining Techniques

An extensive variety of data mining techniques are employed in the fields of data science and data analytics. The technique you select is contingent upon the nature of your issue, the available data, and the desired results. Predictive modelling is a critical element of data mining and is frequently employed to generate predictions or forecasts that are predicated on historical data patterns. Additionally, you may implement a combination of methodologies to acquire a comprehensive understanding of the data. The following are the top ten data mining techniques:

1. Categorization

Classification is a method that is employed to organize data into predetermined classes or categories by analyzing the features or attributes of the data instances. It entails the training of a model on labeled data and its subsequent application to the prediction of the class labels of new [7], unobserved data instances. Three distinct clusters in blue, orange, and green, separated by coordinates, are depicted in a scatter plot. Machine Learning Mastery is the source of the cluster labels, which are 0, 1, and 2 as in Figure 1.

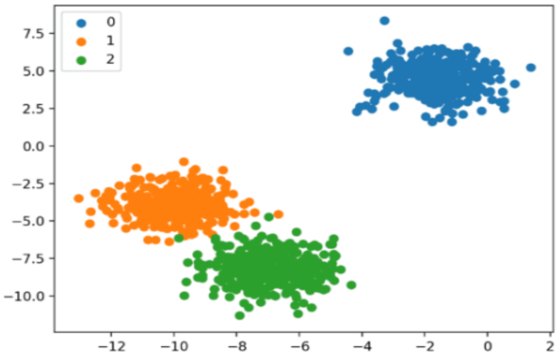


Figure 1: Data Category

2. Regression

The relationship between input variables and a target variable is used to predict numeric or continuous values through regression. It endeavors to identify a mathematical function or model that most closely aligns with the data in order to generate precise predictions. A positive correlation is demonstrated by a scatter diagram with blue data points and a red linear regression line as in Figure 2.

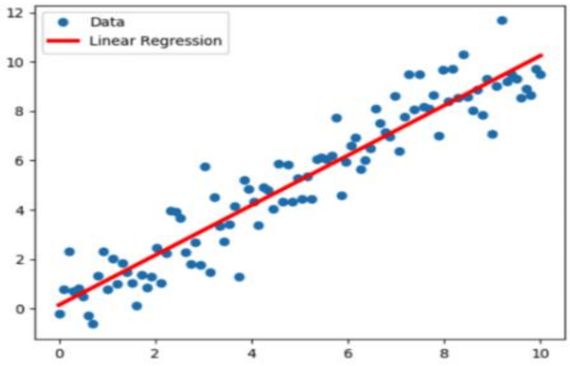


Figure 2: Data Regression

3. Data Clustering

Clustering is a method that is employed to organize data instances that are similar in nature or have similar intrinsic characteristics. It endeavors to identify organic patterns or structures in the data without the use of predetermined classes or identifiers [8]. Two data clusters, shown in red and blue, are separated by a diagonal black line in a scatter plot. The source of the axes is Wikipedia, and they range from 0 to 1.

4. Association Rule

The primary objective of association rule mining is to identify intriguing relationships or patterns among a collection of items in transactional or market basket data. It assists in the identification of frequently co-occurring items and the generation of rules, such as "if X, then Y," to disclose associations between items. The associations between item sets X and Y of a dataset are illustrated in this straightforward Venn diagram as in Figure 3.

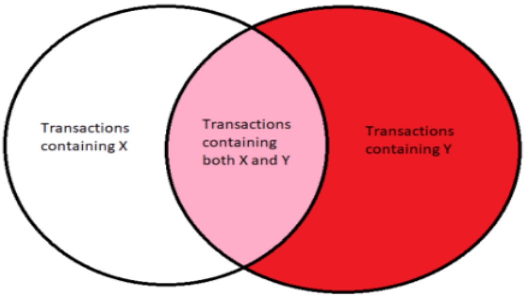

Transactions containing X
Transactions containing both X and Y
Transactions containing Y

Figure 3: Association Rule

III. NLP Techniques for Processing Transaction Descriptions

A) Text Pre-Processing

Text data is the focus of Natural Language Processing (NLP), a subfield of Data Science. In addition to numerical data, text data is widely accessible and employed to analyze and resolve business issues. Nevertheless, it is crucial to process the data prior [9] to utilizing it for analysis or prediction. We conduct text preprocessing to prepare the text data for model construction. It is the initial phase of NLP initiatives. Some of the preprocessing processes include:

- Eliminating punctuation marks such as. and! $( ) * % @
- Eliminating URLs
- Eliminating Stopped sentences
- Lower enclosure
- Tokenization

- Reduction of Noise:

Frequently, text data is contaminated with noise, including punctuation, special characters, and irrelevant symbols. Preprocessing assists in the elimination of these components, thereby facilitating the analysis and cleaning of the text [10].

- Normalization:

The same meaning can be conveyed by various word forms, such as "run," "running," and "ran," despite their distinct appearances. Stemming and lemmatization are preprocessing techniques that contribute to the standardization of these variations.

- Tokenization:

Text data must be broken down into smaller elements, such as words or phrases, in order to facilitate analysis. Tokenization simplifies subsequent processing procedures, such as feature extraction, by dividing text into meaningful units.

- Elimination of Stop words:

Stop words are frequently used words, such as "the," "is," and "and," that have minimal semantic significance. The efficacy of text analysis can be enhanced by removing stop words, as it reduces noise.

- Feature Extraction:

In order to construct machine learning models, preprocessing may entail the extraction of features from text, including word frequencies, n-grams, and word embeddings.

- Dimensionality Reduction:

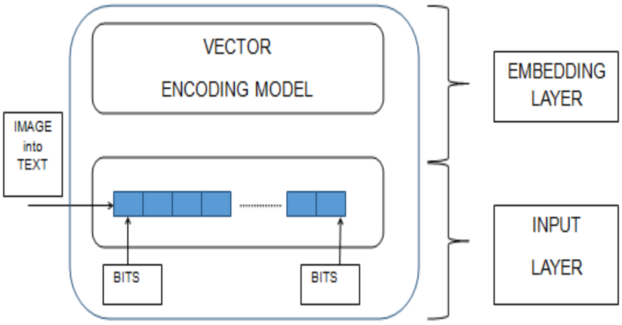
The presence of a vast vocabulary frequently results in text data having a high degree of dimensionality. Dimensionality reduction methods or term frequency-inverse document frequency (TF-IDF) preprocessing techniques may prove advantageous.

B) Lemmatization Stemming

Tokenization is the process of dividing a text into smaller entities, typically words or subwords. This facilitates the conversion of the text into elements that are more easily comprehensible. A component of Speech Tagging [11] – Assigns grammatical elements of speech (nouns, verbs, adjectives, etc.) to each word in a sentence. Demonstrates the syntactic structure of a sentence to the machine learning model.

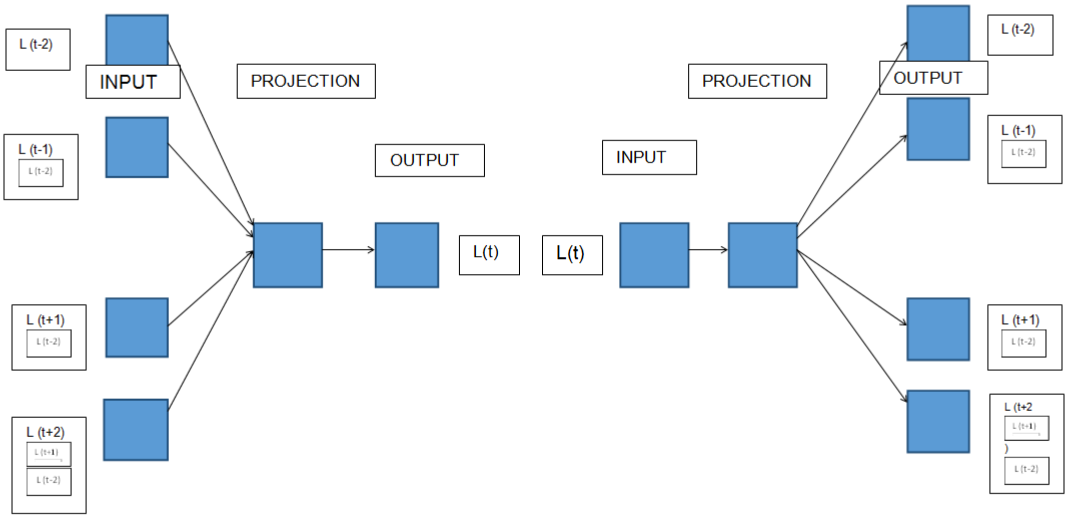
- Named Entity Recognition – Recognizes and categorizes entities in text, including the names of individuals, organizations, and locations. Extracts structured information from unstructured text.
- Parsing and Syntax Analysis – Examines the grammatical structure of sentences to comprehend the relationships between words. Aids in comprehending the significance of sentences and their constituent parts.

C) Vector Encoding Model



Word Embedding (WE) is a representation technique of a text where the words with the same meaning have a similar representation. Recently, there have been several word embedding's widely used in ML and DL models. In the literature, there are many pre-trained WEs that can be categorized into two groups: static representation models and contextual models. Word2vec, GloVe, and FastText are types of static WEs that can convert a text into vectors of meaningful representation.

ISSN (online): 2581-3048
Volume 7, Issue 5, pp 362-376, May-2023
https://doi.org/10.47001/IRJIET/2023.705054



Word2vec works as a language model, which is widely used for many NLP tasks. In general, the word2vec embedding can be obtained using either skip gram or common bag of words (CBOW). The skip-gram model computes the conditional probability of a word by predicting the surrounding context words given the central target word.

$$W_i(X,Y) = a_i + b_i X + C_i Y$$

The CBOW does the opposite of skip-gram, by computing the conditional probability of a target word give the context words surrounding it across a window of size k.

$$CBOW = \sum_{i=1}^{L}\sum_{j=1}^{L} Fun(X_{ij}) * ((L_i^t * L_j) + Base_i + Base_j - \log(X_{ij})^2$$

$$ModCBOW = \frac{1}{\lceil Vector \rceil} \sum_{t=1}^{|Vector|} \sum_{j=t-c, j<>t}^{t+c} \log[Prob(L_i \,|L_t)$$

Algorithm 1: Modified BCOW Model

```
Input: Context words
Output: Middle word
If window size = 5
{
    Input Word = 2
    Future Word = 2
    Output Word = 1 middle word
}
If (Skip-gram = TRUE)
{
   Input = 1 middle word
   Output = 2 Future words && 2 History words
}
If (Window size = Increases)
{
   Quality of the model = Increases;
Else
   Computational Complexity = Increases;
}
```

A) Text Classification

- Database

We have created a dataset by utilizing publicly available datasets for a variety of categories, including Food, Music, Games, Watch Next, and Politics [12].

- Neural Networks

The fundamental concept of a neural network is to create a computer that simulates a multitude of densely interconnected brain cells. This allows the computer to learn, recognize patterns, and make decisions in a manner that is reminiscent of a human. Neural networks are the standard representation of the brain, consisting of neurons that are connected to one another to form a network. The operation of a neural network is simple: input variables are provided (e.g., an image if the neural network is intended to determine the contents of an image), and the output is returned after a series of calculations. For instance, the word "cat" should be returned if an image of a cat is provided [13].

In a typical neural network, a series of layers is used to arrange a variety of artificial neurons, ranging from a few dozen to hundreds of thousands or even millions. Each layer is connected to the layers on either side. The initial layer is where inputs are entered, and it is intended to receive a variety of information from the external world that the network will endeavor to comprehend, identify, or otherwise process. Other neurons are situated on the opposing side of the network and provide feedback on the network's response to the information it has acquired. This layer is referred to as the output layer.

The artificial brain is primarily composed of one or more layers of concealed units, which are situated between the input and output layers [14]. These strata are referred to as "hidden layers." The majority of neural networks are fully connected, which means that each concealed unit and output unit are connected to every unit in the layers on either side. Weight is a numerical value that denotes the connections between units and can be either positive or negative. The greater the weight, the greater the influence of one unit on another.

While a straightforward neural network for straightforward problem solving may consist of only four layers, as demonstrated here, it may also include numerous distinct layers between the input and the output. A deep neural network (DNN) is a more intricate structure that is commonly employed to address much more complex problems. In theory, a DNN has the capacity to convert any type of input into any type of output. However, the disadvantage is that it requires a significantly greater amount of training. In contrast to a simpler network, which may require only hundreds or thousands of examples, a DNN must "see" millions or billions of examples.

Initially, the neuron accumulates the value of each neuron from the preceding stratum to which it is connected. The neuron in the aforementioned figure is connected to three neurons from the previous column, as indicated by the three inputs (x1, x2, x3).The connection between the two neurons is determined by another variable called "weight" (w1, w2, w3), which is multiplied by this value prior to being added. The only values that will be altered during the learning process are the weights of each neuron connection.

Furthermore, it is feasible to incorporate a bias value into the total value that is computed. It is not a value that originates from a specific neuron and is selected prior to the learning phase; however, it may prove advantageous to the network.The neuron ultimately applies a function known as the "activation function" to the obtained value after all of those summations. Sigmoid, Tanh, RELU, Leaky RELU, and Softmax are examples of activation functions. The neuron is now prepared to transmit its new value to other neurons [14].

There are two methods by which information is transmitted through a neural network. The input units feed patterns of information into the network when it is learning (being trained) or operating ordinarily (after being trained). This triggers the layers of hidden units, which in turn reach the output units. A feedforward network is the term used to describe this prevalent design. Not all elements are always capable of firing. The weights of the connections that each unit travels along are multiplied by the inputs it receives from the units to its left. In the simplest form of network, each unit accumulates all of the inputs it receives. If the sum exceeds a specific threshold value, the unit "fires" and activates the units to which it is connected.

For a neural network to acquire knowledge, feedback must be present. Typically, neural networks acquire knowledge through a feedback process known as backpropagation. This entails comparing the output of a network with the output it was intended to produce and utilizing the discrepancy to adjust the weights of the connections between the units in the network. The process proceeds from the output units through the hidden units to the input units, i.e., going backward. Backpropagation results in the network learning over time, thereby reducing the discrepancy between the intended and actual output to the point where they are identical. Consequently, the network is able to resolve the situation as intended [15].

IV. Combining Data Mining and NLP

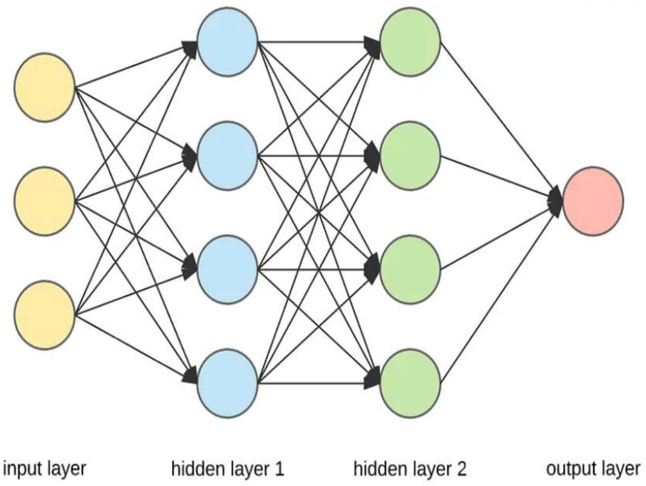


Figure 4: Multi Perceptron Model

When discussing NLP techniques, data mining, and data types such as structured and unstructured, the goal is to extract meaningful insights from both structured and unstructured data through data mining techniques. Unstructured data typically necessitates more advanced NLP processing to convert it into a format that is suitable for analysis [16].

Important factors to Organized the data:

- **Structured Data:**

This type of data is structured in a well-defined manner, similar to a database table with columns and rows, which facilitates its accessibility for conventional data mining methods.

- **Unstructured Data:**

This data is often presented as text, images, or audio, and lacks a predefined structure. It necessitates the use of natural language processing (NLP) methods to extract meaningful information.

Unstructured Data NLP Techniques:

- Tokenization: Tokenization is the process of dividing text into individual words or phrases.
- Stemming/Lemmatization: The process of reducing words to their fundamental form.
- Part of Speech (POS) Tagging: The process of determining the grammatical function of each word (noun, verb, adjective).
- Named Entity Recognition (NER): The process of identifying and categorising entities, such as individuals, locations, and organisations.
- Sentiment Analysis: The process of determining the sentiment (positive, negative, or neutral) that is conveyed in a text.
- Text Classification: The process of classifying text documents according to their subject matter.

Basic feature extraction techniques in NLP are employed to analyse the similarities between text segments. Natural Language Processing (NLP) is a field of computer science and machine learning that involves the training of computers to process a substantial volume of human (natural) language data. The capacity of computers to comprehend human discourse is known as natural language processing (NLP) [17]. The necessity of feature extraction techniques Machine Learning algorithms use a set of features that have already been described in the training data to learn how to make output for the test data. However, the primary issue with language processing is that machine learning algorithms are unable to directly interact with the raw text. Therefore, in order to transform text into a matrix (or vector) of features, we require feature extraction techniques. Several of the most prevalent techniques for feature extraction include:

- Bag of Words
- TF-IDF

A) Bag of Words

The bag of words model is employed in natural language processing and information retrieval tasks to represent text and extract features. It represents a text document as a collection of its words, preserving the frequency of words while disregarding grammar and word order. This representation is beneficial for tasks such as text classification, document similarity, and text clustering.

Bag-of-One of the most fundamental methods for converting identifiers into a collection of features is the use of words. The BoW model is employed in document classification, where each word is utilised as a feature to train the classifier [18]. For instance, in a sentiment analysis task that is based on reviews, the presence of words such as "excellent" and "fabulous" suggests a positive review, while words such as "poor" and "annoying" indicate a negative review. Three stages are involved in the development of a BoW model:

- Text-preprocessing is the initial phase, which entails the conversion of the entire text to lowercase characters.
- Eliminating all superfluous symbols and punctuation.
- A vocabulary consisting of all distinct words from the corpus is the second phase.

B) TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure that is employed in the context of natural language processing and information retrieval. It illustrates the significance of a single word within a document in relation to the entirety of the corpus. The fundamental concept is that a word that is frequently used in a document but rarely in the entire corpus is more informative than a word that is frequently used in both the document and the corpus.

TF-IDF is employed for the following purposes:

a) Information retrieval and text retrieval systems.
b) Text categorisation and document classification.
c) Summarisation of Text.
d) Text data feature extraction in machine learning algorithms.

TF-IDF is an acronym for term frequency-inverse document frequency. It emphasises a particular concern that may not be as prevalent in our corpus but is of significant significance. The TF–IFD value increases in proportion to the number of times a word appears in the document and decreases in proportion to the number of documents in the corpus that contain the specific word. It is divided into two sub-components:

- Term Frequency (TF)
- Inverse Document Frequency (IDF)

A) Clustering in Data Mining

After a brief introduction to the field of data mining, we will now focus on the specific technique of interest, clustering analysis. Clustering analysis, which is frequently abbreviated as "clustering," is a data mining method that unites data points that share comparable characteristics. This approach facilitates the identification of patterns, structures, and relationships in extensive datasets, thereby facilitating the acquisition of more insightful information and the formulation of more informed decisions [19].

Clustering's versatility and significance are underscored by its extensive application in a variety of disciplines, including bioinformatics, market research, image processing, and social network analysis. Consider the act of sorting through a basket of a variety of fruits. Clustering would enable you to automatically separate the kiwis from the bananas and the apples from the oranges.

B) Mechanism by clustering in data mining

Clustering analysis is not a mystical process; rather, it is a well-defined method that groups data points that are similar. Clustering in data mining comprises the subsequent procedures:

- Data Collection and Preparation:

The initial phase of the clustering procedure involves the collection and preprocessing of data. Preprocessing, which is also known as data cleaning, may entail the management of missing values, anomalies, and inconsistencies to guarantee the accuracy of the analysis. This procedure guarantees that the data is pure and appropriate for clustering.

- Feature Selection:

After the data is prepared for clustering, it is necessary to identify the pertinent features or attributes that most accurately represent the data. These features will be employed to quantify the similarity between data points. Consider the possibility of concentrating on the objects' colour, size, and material rather than their brand or origin, which may not be pertinent for grouping. It is important to note that these features should be pertinent to the analysis.

- Normalization/Standardization:

If features have varying scales (e.g., weight versus height), normalisation or standardisation may be required to establish a comparable scale. This guarantees that the similarity measure is not dominated by a single feature [19].

- Similarity or Distance Measurement:

It is essential to establish the method by which the similarity or distance between data elements is measured. This metric quantifies the degree of similarity between two data elements by analysing their features.

  - Euclidean distance (straight-line distance) and Manhattan distance (sum of absolute differences) are two commonly used distance metrics for numerical data.
  - Similarity measures such as Jaccard similarity (ratio of shared features) may be implemented for categorical data.

- Clustering Algorithm Selection:

The subsequent phase in data mining clustering is to determine the most suitable clustering algorithm (clustering technique) for the data and the desired outcome. Cluster Process/Formation: Implement the clustering/mining algorithm, which iterates through the data and assigns data points to clusters based on the similarity measure. Common algorithms include K-means, hierarchical clustering, and DBSCAN (we will cover these algorithms/clustering methods in a subsequent section) [20].

  - This procedure entails the iterative relocation or updating of cluster assignments to optimise inter-cluster distances and minimise intra-cluster distances.
  - The quantity of sub-steps varies among different clustering algorithms. For instance, K-Means clustering may require the recompilation of centroids after each assignment until the clusters achieve stability.

- Evaluation & Validation (Optional):

Although not always necessary, assessing the quality of the clusters can be beneficial in the process of eliminating those that are low-quality. Silhouette analysis is a technique that quantifies the degree of separation between the clusters. Validation guarantees the reliability of the results and the significance of the clusters.

- Interpretation and Analysis:

Conduct an analysis of the clusters that have been generated in order to derive significant insights and conclusions. This phase entails comprehending the attributes of each cluster and their correlation with the issue at hand. It

ISSN (online): 2581-3048
Volume 7, Issue 5, pp 362-376, May-2023
https://doi.org/10.47001/IRJIET/2023.705054

also entails the utilisation of visualisation techniques, such as scatter plots, dendrograms, or heat maps, to represent the clusters and enhance the interpretability of the results.

C) Various Clustering Methods in Data Mining

Clustering methods in data mining are methods that are employed to organise a collection of objects in a manner that ensures that objects within the same group (or cluster) are more similar than those in other groups. Various clustering techniques are appropriate for a variety of data and applications [21].

Data mining provides a diverse array of clustering techniques, each with its own advantages and disadvantages. The six primary clustering techniques in data mining will be the subject of this section:

- Methods of Partitioning
- Algorithms and Methods for Hierarchical Clustering
- Methods Based on Density
- Methods Based on Models
- Clustering Methods Based on Grids
- Methods that are based on constraints

V. Real Time Applications

NLP has a variety of finance-specific applications, such as sentiment analysis, sentiment analysis, auditing and accounting, and portfolio selection as discussed in Figure 5.

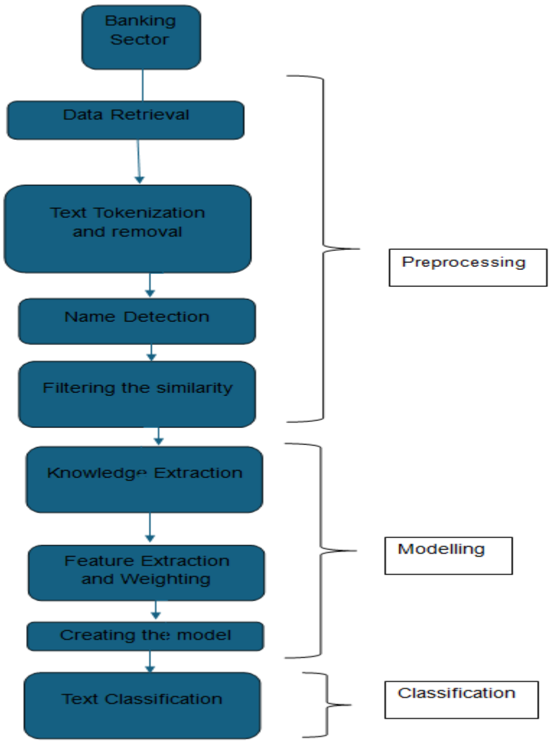


Figure 5: System Stages

1. Fraud Detection:

A credit risk assessment is a method by which banks can determine the likelihood of a successful loan payment. The payment capacity is typically determined by analysing past loan payment history data and expenditure patterns. However, this information is unavailable in numerous instances, particularly for individuals who are impoverished. Poverty is the reason that nearly half of the global population does not utilise financial services, according to an estimate [22].

NLP is present to resolve this issue. Credit risk is evaluated through the utilisation of numerous data points in NLP techniques. For example, NLP can assess entrepreneurial mindset and attitude in business loans. In the same vein, it has the ability to identify data that is not coherent and subject it to further examination. In addition, NLP can be employed to incorporate the subtle aspects of the loan process, such as the emotions of the lender and borrower [23].

Typically, companies extract a significant amount of data from personal loan documents and incorporate it into credit risk models for additional analysis. Despite the fact that the information collected is beneficial for evaluating credit risk, errors in data extraction can result in inaccurate assessments. In these circumstances, named entity recognition (NER), an NLP technique, is advantageous. NER assists in the extraction of pertinent entities from the loan agreement, such as the date, location, and details of the parties involved.

→ Discover how the NLP social graph technique can assist clinical research organisations in the successful analysis of clinical trials by evaluating patient databases.

2. Economic sentiment

Information regarding specific equities is essential for successful trading in the stock market. Traders can determine whether to purchase, retain, or sell a stock by utilising this information. In addition to examining quarterly financial statements, it is crucial to be aware of the opinions of analysts regarding the companies in question, and this information can be accessed through social media.

Social media analysis entails the identification of prospective trading opportunities by monitoring the information contained in social media posts. For instance, the stock price may be adversely affected by news of a CEO's resignation, which typically elicits a negative sentiment. However, if the CEO was not performing well, the stock market may respond favourably to the news of his resignation, which could potentially lead to an increase in the stock price. DataMining and Bloomberg are among the organisations that offer this type of information to facilitate trading. For instance, DataMinr has furnished its users with stock-specific alerts and Dell-related news on its terminals, which may have an impact on the market [24].

The financial sentiment analysis is distinct from the routine sentiment analysis. It is distinct in terms of both its domain and its objective. The objective of conventional sentiment analysis is to determine whether the information is inherently positive or not. Nevertheless, the objective of financial sentiment analysis, which is based on NLP, is to determine whether the market will respond to the news and whether the stock price will increase or decrease.

BioBERT, a pre-trained biomedical language representation model for biomedical text mining, has proven to be highly beneficial in the healthcare sector. Currently, researchers are in the process of integrating BERT into the financial sector. FinBERT is one of the frameworks that was created for the financial services sector. FinBERT operates on a dataset that includes financial news from Reuters. A Phrase Bank was implemented to designate sentiment. The collection comprises approximately 4,000 sentences that have been tagged by individuals with a variety of financial or business backgrounds.

In the conventional application of sentiment analysis, a positive statement is indicative of a positive emotion. However, in the Financial Phrase Bank, negative sentiment suggests that the company's stock price may decline as a result of the published news. FinBERT has achieved considerable success, with an accuracy of 0.97 and an F1 of 0.95, which is considerably higher than that of other available tools. The FinBERT library is available on GitHub, and the pertinent data is available. Various applications can be achieved by employing this robust language model for economic sentiment classification [25].

3. Auditing and accounting

Ernst & Young, Deloitte, and PwC are committed to conducting audits of a company's annual performance that are actionable and meaningful. For example, Deloitte has transformed its Audit Command Language into a more effective NLP application. It has employed NLP techniques to evaluate contract documents and long-term procurement agreements, particularly in the context of government data.

Companies are now recognising the significance of NLP in obtaining a substantial advantage in the audit process, particularly after decades of navigating endless daily transactions and invoice-like documents. NLP enables financial professionals to directly identify, focus, and visualise anomalies in the day-to-day transactions. The investigation of irregularities in transactions and their causes requires less time and effort when the appropriate technology is employed. NLP can assist in the identification of substantial potential hazards and potential fraud, such as money laundering. This facilitates the advancement of value-generating activities and their dissemination throughout the organisation.

4. Optimisation and selection of portfolios

Without knowledge of the fundamental distribution generated by stock prices, the primary objective of every investor is to optimise their capital over the long term. Data science, machine learning, and nonparametric statistics can be employed to forecast investment strategies in financial stock markets. The data that has been gathered in the past can be employed to forecast the commencement of a trade period and a portfolio. Investors can allocate their existing capital among the assets that are accessible as a result of this information [26].

Semi-log-optimal portfolio optimisation can be achieved using NLP. Semi-log-optimal portfolio selection is a computational alternative to log-optimal portfolio selection. When environmental factors are uncertain, it facilitates the attainment of the highest feasible development rate. By filtering out desirable and undesirable equities, data envelopment analysis can be employed for portfolio selection.

5. Predictions regarding stock behaviour

The task of predicting time series for financial analysis is complex due to the fluctuating and irregular data, as well as the long-term and seasonal variations that can result in significant errors in the analysis. Nevertheless, the combination of NLP and deep learning surpasses the previous methodologies that were employed to analyse financial time series to a significant extent. The combination of these two technologies effectively manages substantial quantities of information.

Deep learning is not a novel concept. In the past five years, a significant number of deep learning algorithms have begun to outperform humans in a variety of tasks, including medical image analysis and speech recognition. Recurrent neural networks (RNN) are a highly effective method for predicting time series, such as stock prices, in the financial domain [27]. RNNs possess the inherent ability to identify intricate nonlinear relationships that are present in financial time series data and to approximate any nonlinear function with a high degree of accuracy. These methods are viable alternatives to the current conventional techniques for predicting stock indices due to the high level of precision they provide. NLP and deep learning techniques are valuable tools for making stock trading decisions, as well as for predicting the volatility of stock prices and trends.

- **Real time solution based on secure transaction classification:**

Propose a poorly supervised, extensible transaction classification method to address the issue. It combines deep neural networks, noise-aware label generative models, and unsupervised transaction text embeddings. Creating categorization predictions for financial applications from transactional data is its goal. In essence, there are two parts to the model pipeline. The label model and data preparation are the main topics of the first section,

- Discriminative model
- Processing and Aggregation of Data

Data aggregation, the first layer, cleans the transaction text and arranges transactions by customer account. Next, we use NLP to categorize this transaction text. By extracting the time series of transaction amounts and considering the time patterns of the series, we can use group-wise time series information and gain important insights into customer behaviour, which ultimately helps us categorize transactions accurately. Our goal is to use feature engineering, data combining, and language preprocessing to turn the unprocessed, frequently noisy transaction descriptions into useful information.

The model's second and third layers provide probabilistic labels and weak labels, respectively. Weak labels, or approximation labels, are produced here. Heuristics, crowdsourcing, and supplementary data are some of the sources from which these labels originate. Ultimately, the model is put into practice in a system that can process transaction data from several sources, oversee the model's training and inference, and then provide predictions to applications further down the line [34].

Thus, there are numerous uses for accurate bank transaction classification in the financial sector. Consumer-focused personalized goods such as financial coaching, subscription alerts, pre-incident alerts, reward programs, and personalized credit product matching can be built on the knowledge it generates. Traditional credit scores, which are solely based on prior credit product usage, can give banks a delayed picture of a customer's financial health. On the other hand, bank transactions can provide accurate and detailed information about bank balances, user spending patterns, and problematic financial practices like overdrafts. Such data could be utilized to increase credit accessibility for new customers and small firms with no credit history, so opening up new credit options and improving traditional credit ratings by giving comprehensible and real-time financial measures.

VI. Challenges and Solutions

In order to optimise sparsity, numerous models implement the mixture-of-experts (MoE) principle, which involves routing computation through smaller subnetworks rather than transmitting the input through the entire model. The following works are pertinent to this line: Switch Transformer, which substitute the feed-forward layers in transformers with MoE layers. The size of transformers is increased by compressing and optimising the use of MoE. In general, MoE models have been demonstrated to exhibit robust performance in a variety of NLP tasks, all while minimising the overall resource consumption (Section 8). For example, GLaM utilised only a fraction of GPT-3's energy consumption (with the assistance of hardware-based optimisation), achieves a fivefold decrease in training costs. Nevertheless, MoE models have also demonstrated training instabilities in practice and may necessitate architecture-specific implementation.

An extension of the Adaptively Sparse Transformer allows for a more efficient attention mechanism by identifying the sparsity pattern returned by entmax attention—a sparse alternative to (dense) softmax attention. Lastly, modularity can also induce sparsity by encapsulating task-specific parameters. The effectiveness of the pre-trained model on downstream tasks can be influenced by the task selection. Left-to-right language models, including GPT and PaLM, are trained with the causal language modelling (CLM) objective, which entails predicting the subsequent token based on a context. The BERT task is a masked language model (MLM) task that entails the input of randomly masked tokens [28].

Diverse masking strategies have been examined in order to optimise the utilisation of available data. By masking only objects and content terms, rather than random tokens or masking a greater number of token, task performance has been enhanced and the available data has been utilised more efficiently. ELECTRA and DeBERTa utilised replaced token detection (RTD), an objective that employs a small generator model to replace input tokens and converges more rapidly to superior performance. The MLM and RTD objectives are restricted to the use of single token replacements. This is resolved by T5 and BART, which employ a denoising sequence-to-sequence objective to pre-train an encoder-decoder model. This enables the decoder to forecast a span of tokens for masked positions. In practice, this enables the training of shortened sequences without compromising task performance, thereby reducing training costs.

Natural language processing1 (NLP) techniques are a diverse array of methodologies that have been employed to extract valuable information from unstructured text. NLP may

ISSN (online): 2581-3048
Volume 7, Issue 5, pp 362-376, May-2023
https://doi.org/10.47001/IRJIET/2023.705054

be a significant factor in the primary task of a radiologist, which is to convert information in medical images into a text format (e.g., a radiology report). Radiology reports are a significant source of medical information and a form of annotation for the corresponding medical images. Nevertheless, their unstructured, free-text nature frequently complicates the process of converting them into a computer-friendly format. In an effort to provide greater structure to text, language standards such as RadLex and SNOMED CT were established (1,2). The vast quantity of free-text information has facilitated the development of a significant niche for NLP applications in radiology through additional parsing and structuring [29] and [30].

Traditional NLP systems in radiology were developed by employing a grammatical rule system to provide structure to free narrative text (3,4). Nevertheless, rule-based approaches necessitate a significant amount of effort to establish, as they are unable to accommodate variations in individual and institutional practices and necessitate the creation of domain-specific tasks. In the past few years, machine learning methods have become increasingly prevalent due to their lack of the need for manual rule-engining (5,6). Recurrent neural network (9,10) and convolutional neural network (7,8) methods have been extensively employed in text classification tasks, including the identification of pneumonia (12,13) or pulmonary embolism (7,11) within a radiology report [31- 33].

Many machine learning-based NLP models require encoding words into a numerical vector representation known as a word embedding (6). GloVe (Global Vectors for Word Representation) (14) and word2vec (15) are two widely used methods for generating word embeddings that have been implemented in radiology NLP models (7,16). Initialising the embedding layer with medical domain data has been found to boost performance (17,18); nonetheless, radiology NLP models have frequently relied on embeddings trained on a wide corpus or learnt from scratch for the specific training set (7,11,19).

We believe that embeddings trained on a radiology-specific corpus can capture underlying medical semantics, which can then be used to improve a model's performance on a radiology NLP challenge. A Creative Commons license (CC BY-NC-SA 3.0) is in place for Radiopaedia.org (20), a resource that is readily accessible and can serve as a radiology corpus. In this investigation, we create a collection of GloVe word embeddings that have been trained on Radiopaedia. We then compare these embeddings to a collection of pre-trained embeddings on a multi-label text classification task and an analogy completion task.

VII. Real World Applications based on ML methodologies

Fraud-related losses and damages are plaguing the financial services sector as well as other companies that deal with financial transactions. Financial fraudsters had a very successful year in 2016. Customers who fell victim to fraud reached a record 15.4 million in the US alone, a 16 percent increase over 2015. In 2016, fraudsters stole almost $6 billion from banks. A move to the digital sphere creates new avenues for the dissemination of financial services. Additionally, it made the atmosphere for scammers rich. Now, obtaining a person's account password can be all that is required to steal money, but in the past, fraudsters had to fake client IDs. Both digital and physical environments have an impact on customer loyalty and conversions. Brick and mortar financial institutions take more than 40 days to identify fraud, according to Javelin Strategy & Research. Banks that offer online payment services are similarly impacted by fraud. For example, 20% of consumers switch banks after falling victim to fraud.

Finding these underlying correlations between user behaviour and the probability of fraudulent activities is made possible by machine learning, which enables the development of algorithms that handle massive datasets with numerous variables. Faster data processing and less manual labour are two more advantages of machine learning systems over rule-based ones. To monitor and handle variables including transaction size, location, time, device, and purchase information, MasterCard, for instance, combined machine learning and artificial intelligence. Every operation evaluates account behaviour, and the system gives a real-time assessment of whether a transaction is fraudulent. The project's goal is to lower the quantity of fictitious merchant payment denials. According to a recent study, misleading reductions cost retailers $118 billion annually, while customers lose roughly $9 billion. In the financial services industry, it is the most common area for fraud. Therefore, preventing fraud is a strategic objective for the banking and payments sectors.

A fine-tuned machine learning system can identify up to 95 percent of all fraud and reduce the expense of manual reconciliations, which currently account for 25 percent of fraud expenditures, according to the fintech company Feedzai. According to Capgemini, machine learning and analytics-based fraud detection solutions increase detection accuracy by 90% and reduce fraud investigation time by 70%. The advantages of using machine learning to anti-fraud systems are demonstrated by these data. Every night between 9 and 10 p.m., a customer goes to a particular grocery. It's close to the client's residence. The amount of the payout ranges from $10

ISSN (online): 2581-3048
Volume 7, Issue 5, pp 362-376, May-2023
https://doi.org/10.47001/IRJIET/2023.705054

to $40. Additionally, the client drives to a petrol station every two days.

When a $40 transaction takes place in a pub in a different section of town, the algorithm will flag this activity as suspicious and raise the likelihood of fraud. The system will issue a verification request to the card owner to verify this transaction. When examining behaviour, descriptive statistics such as averages, standard deviations, and high/low values are highly helpful. These measures make it possible to compare individual transactions to intra-group or personal benchmarks. Large standard deviations in payments give the impression that they are suspect. Sending an account owner a request for verification in the event of such aberrations is therefore a smart practice.

The type of problem, dataset size, available resources, etc., all influence the machine learning approach that is best. Using many models is an excellent way to increase accuracy and expedite assessment. PayPal, for instance, uses linear models to differentiate ambiguous transactions from regular ones by direct assessment. A set of three models, including a neural network, a deep neural network, and a linear model, are then applied to any transactions that appear suspicious. The three then cast their votes to determine which outcome is more accurate.

These days, antifraud systems should be able to:

- Identify fraud instantly.
- Increase the reliability of data.
- Examine user behaviour.
- Reveal hidden relationships.

Although machine learning algorithms can provide these attributes, there are two significant disadvantages to be mindful of. They still need certain characteristics of rule-based engines, such as verifying the legal restrictions on cash transactions, and they still need big, meticulously prepared datasets for training. Additionally, to create intricate and reliable ensemble algorithms, machine learning solutions typically call for a high level of data science expertise. This makes it very difficult for small and medium-sized businesses to employ the internal talent leveraging strategy. Deep domain and technological expertise are needed for the task. Hiring outside data science specialists is standard procedure. Compared to starting from scratch with an internal data science team, data engineering and consulting services speed up development and cost less.

VIII. Performance Metrics

The performance levels of the proposed model are calculated by calculating the precision levels, recall levels, accuracy, sensitivity and specificity levels. TP represents True Positive, TN represents True Negative, FP is the False Positive and FN is indicated as False Negative. The calculations are performed as

$$Precision = \frac{TP}{TP + FP}$$

Recall is defined as the percentage of instances categorised as belonging to a certain class divided by the total number of examples in that class.

$$\text{Re}call = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%,$$

$$sensitivity(Sn) = \frac{TP}{TP + FN} \times 100\%,$$

$$specificit\,y(Sp) = \frac{TN}{TN + FP} \times 100\%.$$

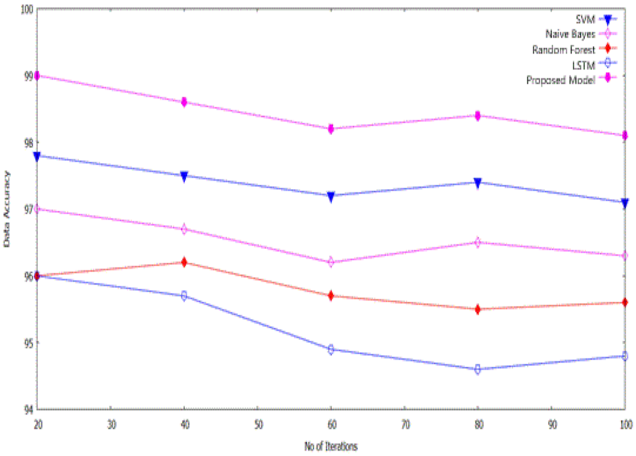


Figure 6: No of Iteratiosn Vs. Data Accuracy

In the Figure 6, Data accruacy is determined based on the variation in the number of iteration from 20 to 100. The proposed model calculates the determined value of 98% as compared to existing model such as, SVM, naïve Bayes, Random forest and LSTM. The Data sensitivity is calculated from the figure 7 based on the variation from 20 to 100 number of iterations.

From the figure 8 and 9 make to determine data specitivity and recall based on the variation in the number of iteration from 20 to 100. The proposed model calculates the determined value of data specitiviity 99.2% and 97.8% values compared to existing model such as, SVM, naïve Bayes, Random forest and LSTM.

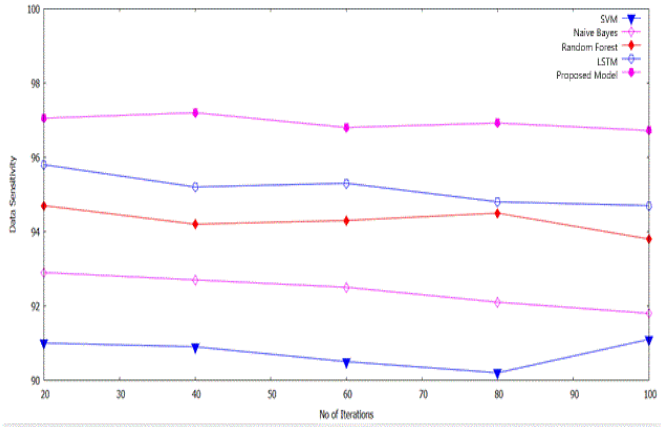


Figure 7: No of Iteratiosn Vs. Data Sensitivity

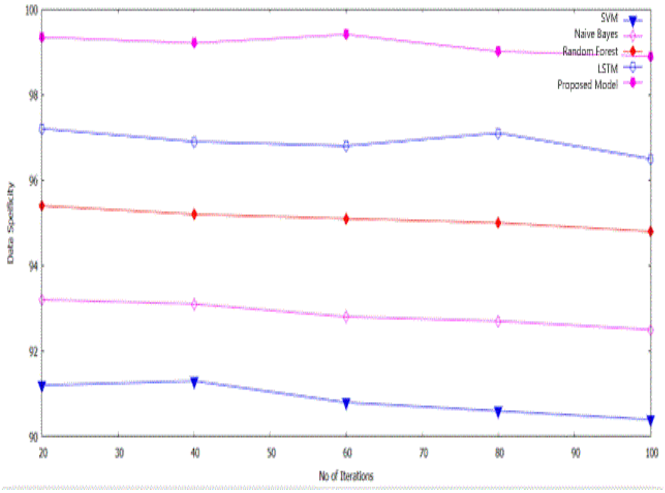


Figure 8: No of Iteratiosn Vs. Data Sepcifivity

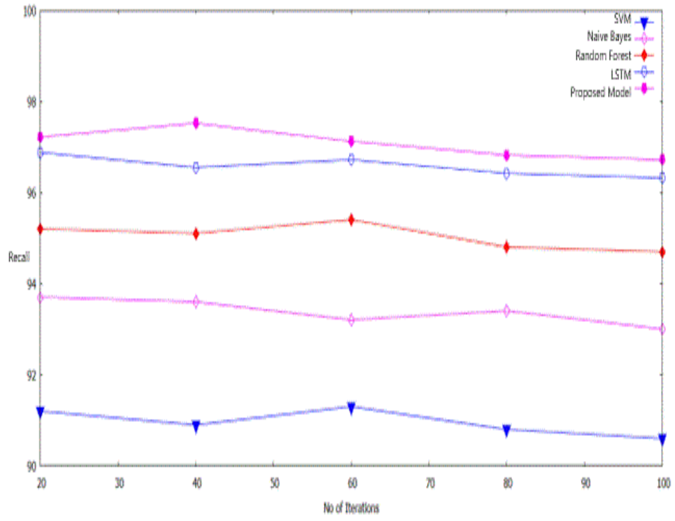


Figure 9: No of Iteratiosn Vs. Recall

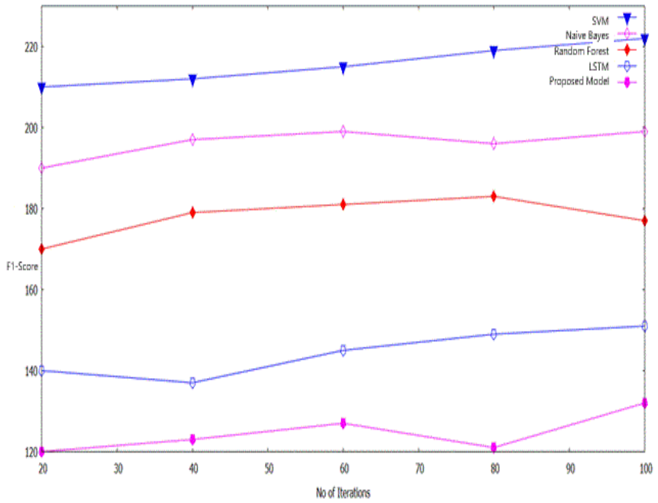


Figure 10: No of Iteratiosn Vs. F1-Score

Form the Figure 10, F1 Score is determined based on the variation from 20 to 100 on number of iteration. The proposed model calculates the determined value of F1 Score 210 as values compared to existing model such as, SVM, naïve Bayes, Random forest and LSTM.

IX. Conclusion

Multiple accounts were kept up to date by utilizing machine learning techniques, which function as AI-powered learning. These techniques were applied to the financial sector, which included both debit and credit entries. It enhances the efficiency with which the model is carried out by deploying a variety of activities. Within this context, many fields like as data mining and natural language processing are utilized to process a huge array of data to carry out prediction modelling. Challenges encountered with applying a variety of approaches, such as machine learning, data mining, and natural language processing. The mixture-of-experts (MoE) principle, which involves routing computation through smaller subnetworks rather than broadcasting the input through the modelling, is implemented by several models to address the many methods that are available to optimize sparsity. Consequently, the levels of performance of the suggested model are determined by computing the levels of precision, recall, accuracy, sensitivity, and specificity simultaneously.

REFERENCES

[1] Ahmad, Wasi, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2020. A transformer-based approach for source code summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4998–5007. Association for Computational Linguistics.

[2] Allamanis, Miltiadis, and Marc Brockschmidt. 2017. Smartpaste: Learning to adapt source code. *arXiv preprint* arXiv:1705.07867.

[3] Allamanis, Miltiadis, Marc Brockschmidt, and Mahmoud Khademi. 2018. Learning to represent programs with graphs. *International Conference on Learning Representations*.

[4] Allen, Frances E. 1970. Control flow analysis. *ACM Sigplan Notices*, 5(7):1–19.

[5] Austin, Jacob, Augustus Odena, Maxwell Nye, Maarten Bosma, et al. 2021. Program synthesis with large language models. *CoRR, abs*/2108.07732.

[6] Banerjee, Satanjeev, and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.

[7] Bunel, Rudy, Matthew Hausknecht, Jacob Devlin, Rishabh Singh, and Pushmeet Kohli. 2018. Leveraging grammar and reinforcement learning for neural program synthesis. *International Conference on Learning Representations*.

[8] Clement, Colin B., Dawn Drain, Jonathan Timcheck, Alexey Svyatkovskiy, and Neel Sundaresan. 2020. Pymt5: Multi-mode translation of natural language and Python code with transformers. *arXiv preprint arXiv*:2010.03150.

[9] Feng, Zhangyin, Daya Guo, Duyu Tang, et al. 2020. CodeBERT: A pre-trained model for programming and natural languages. *Findings of the Association for Computational Linguistics: EMNLP* 2020, 1536–1547.

[10] Ferrante, Jeanne, Karl J. Ottenstein, and Joe D. Warren. 1987. The program dependence graph and its use in optimization. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 9(3):319–349.

[11] Goyal, Vishrav, et al. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

[12] Guo, Daya, Shuo Ren, Shuai Lu, et al. 2020. Graphcodebert: Pre-training code representations with data flow. *International Conference on Learning Representations.*

[13] Hindle, Abram, Earl T. Barr, Mark Gabel, Zhendong Su, and Premkumar Devanbu. 2016. On the naturalness of software. *Communications of the ACM*, 59(5):122–131.

[14] Hovsepyan, Aram, Riccardo Scandariato, Wouter Joosen, and James Walden. 2012. Software vulnerability prediction using text analysis techniques. *Proceedings of the 4th international workshop on Security measurements and metrics*, 7–10.

[15] Hu, Xing, Ge Li, Xin Xia, et al. 2018. Deep code comment generation. *2018 IEEE/ACM 26th International Conference on Program Comprehension (ICPC)*, 200–20010.

[16] Husain, Hamel, Ho-Hsiang Wu, Tiferet Gazit, et al. 2019. Code search net challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:*1909.09436.

[17] Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[18] Gupta, Kavi, Peter Ebert Christensen, Xinyun Chen, and Dawn Song. 2020. Synthesize, execute and debug: Learning to repair for neural program synthesis. *Advances in Neural Information Processing Systems*, 33:17685–17695.

Citation of this Article:

Praneeth Reddy Amudala Puchakayala, "Data Mining-Based Transaction Labelling Enhancing Financial Insights through Automated Technique" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 5, pp 362-376, May 2023. https://doi.org/10.47001/IRJIET/2023.705054

****