A Modelling Framework of an Intelligent News Recommendation Engine for Financial Analysts

Piyumika Samarasekara[#1], Nadeeka Basnayake[2], Desika Hureekaduwa[3], Beshani Weralupitiya[4].

***Abstract*— Referring to news articles has become a predominant task in the daily routine of each financial analyst, as it helps to make accurate financial insights. However, as there are thousands of news articles generated daily, finding the most relevant articles becomes a time-consuming task. Hence, this study develops a modelling framework of an intelligent news recommendation engine for financial analysts in a particular financial company, which assists in recommending the most appropriate articles according to analysts' preferences in an efficient and effective manner without tedious browsing.**

In this study, the data collection phase of the recommendation engine is accomplished by retrieving news articles from online news websites using the web scraping technique. The response variable, analysts' preference level for each article is manually acquired from a group of financial analysts at a selected financial company. In the analysis phase of the recommendation engine, a classification-based method was utilized where three machine learning models, KNN, SVM, Random Forest, and two deep learning models, LSTM and CNN with Natural Language Processing techniques are experimented to discover the best algorithm. Moreover, the synonym replacement method is employed as a text data augmentation method to address the imbalanced problem in the dataset.

As CNN obtains the highest accuracy in comparison to other applied methods it is chosen as the most suitable approach in the analysis phase. Moreover, this study reveals that DL models perform considerably higher performances in the context of news recommendation engines rather than ML approaches. Lastly, this framework can be adjustable for any financial company with minor modifications.

***Keywords*—Natural Language Processing, Machine Learning, Data Augmentation, News Recommendation engine, Financial Analysts**

I. INTRODUCTION

Financial analysts hold a more coveted and paramount job role because a small mistake of them can cause huge financial losses from reputed companies to individual civilians. Most importantly it can finally affect the financial stability of a whole country. It is mentioned that the financial analyst's job role contains tasks such as identifying market opportunities, evaluating macroeconomic and microeconomic conditions of countries by diving into financial data, and providing guidance about the business decisions and investment options to businesses and individuals respectively. To deliver the best output of mentioned financial activities, financial analysts must be aware of the facts which can affect the companies in their specialized sector such as current developments of certain sectors, regulations, policies, political and economic trends of countries, etc.

A. Importance of a News Recommendation Engine to Financial Analysts

The best way to hold attention to global incidents and to be updated about the daily happenings in the world is to follow news articles. Most especially there are numerous scientific evidence such as [1] and [2] to confirm that there exists a positive relationship between financial news articles and financial activities such as stock market predictions. Not only for that but also for gaining information on some non-financial instances like understanding the structure of the companies and so on, analysts must refer to news articles. Hence referring to the news is an indispensable thing for analysts, and it becomes a mandatory task in their daily routine.

However, the way of disclosing one incident may not be the same in different news websites due to the diversity of writers' views. When it comes to relevancy, analysts have to seek news articles that are related to their specialized sector, and which can devote more meaning to analysts' tasks. For example, both news articles illustrated in figure 1, are related to the Apple Inc. company. The first one is about shutting down the Apple retail stores in the US, and it can be clearly understood that closing stores of a reputed company such as Apple, may cause a shake in the stock market. The second news is about a trailer for Apple TV and that will be liked by the apple products lovers. The most relevant article for financial analysts from those two articles is clearly the first one.

Currently they access to news through online news websites. However, it is known that in every hour there are thousands of news articles generated globally. Because of that, finding out the most relevant articles from those, and visiting each website to read news articles, are tedious and time-consuming.

Even though there are studies which are focused on building news recommendation engines, researchers have put up less emphasis on building news recommendation engines specifically for financial news articles. Among the few, they mainly investigate the general interest when recommending financial news. As different institutes have different agendas, implementing a mechanism to filter out the relevant articles which are financially and non-financially significant for a financial company's interest is very important. However, according to the past literature, researchers have made less focus on building a mechanism

Correspondence: Piyumika Samarasekara (E-mail: piyumika.herath@gmail.com) Received: 16.08.2022 Revised:01.06.2023 Accepted: 22-11-2023

Piyumika Samarasekara and Nadeeka Basnayake are from the Department of Statistics University of Colombo, Sri Lanka. (piyumika.herath@gmail.com, nadeeka@stat.cmb.ac.lk). Desika Hureekaduwa and Beshani Weralupitiya are from the Acuity Knowledge Partners, Sri Lanka. (desika.hureekaduwa@acuitykp.com beshani95@gmail.com

DOI: https://doi.org/10.4038/icter.v16i4.7269

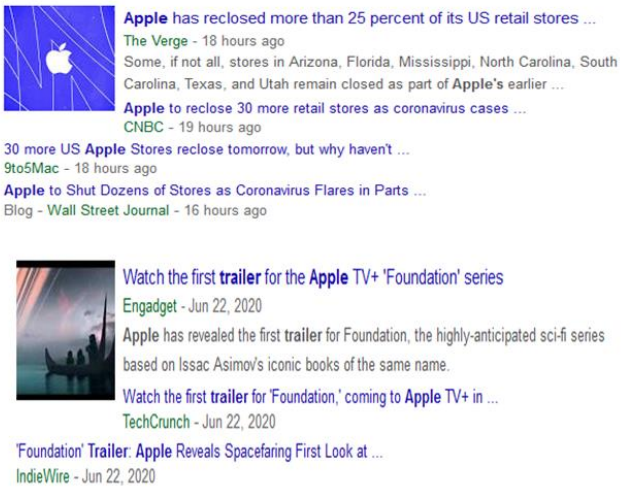

Fig. 1 Importance of a news article to a financial analyst. (Example)

to fulfil that requirement. Hence this research which develops the framework of an intelligent news recommendation engine which can be defined as an intelligent tool that assists to filter out the most suitable news items for financial analysts, according to their preferences plays a very significant role.

This paper is organized as follows: Next section gives an overview of the related works. Section III describes the methodology used in this research. Section IV provides the results obtained from the analysis phase. Section V presents the discussion. In section VI the conclusion is drawn based on the findings of the study and in the final section, Section VII the future research directions are indicated.

II. Related Works

A. Recommendation Engines

In this modern society, due to the thousands of data availability on the internet, people most commonly face the problem of not being able to find the suitable item that they need. Researchers focused on this major issue and came up with the idea of a recommendation engine as the solution. As expressed in previous research, a recommendation engine recommends information to users taking different characteristics and different activities of them into account, [3], [4]. Felfernig illustrated that the two of the major approaches used in the construction process of recommendation engines are content-based filtering approach and collaborative filtering approach [5].

As mentioned earlier, in the studies that developed news recommendation engines also utilized these two approaches vastly in the analysis process. Under the content-based filtering approach, the engine recommends an item to a user that has similar contents to the previously preferred items. According to [6] and [7], in the study [3] researchers presented details about news recommender systems which were employed this approach.

The next major approach that is used in news recommendation engines: collaborative filtering, considers the historically exhibited behaviours of the large set of users and the similarities of those users and recommends an item to a user where other users who have similar characteristics to that user preferred in the past. An engine that facilitated to recommend real time news articles as soon as after gathering online data called NEWSD (News Explorer for Web Streaming Data) used this approach [8]. Again, the research [3], exhibited the studies that used this approach in the analysis process of the news recommender system.

All the above-stated recommendation engines were designed for the general public, and they are not focused only on one particular set of people in society. The study that was done by Wanrong Gu and the team, was one step ahead and it compiled a news recommendation engine specially for microblog users using a hybrid method [9].

It is known that financial analysts are a crucial group of people who benefit greatly from referring to news articles. Both the explained methods, collaborative filtering, and content-based filtering are not suitable for our study as there's a high risk of missing rarely happening but crucial incidents if those two methods are utilized. As previous researchers have less focused on developing a news recommendation engine specifically for financial analysts, this study is going to address that research gap.

A group of researchers built a recommendation engine to recommend accommodations to backpackers using a classification model-based approach [8]. A similar approach is used with text classification techniques and natural language processing techniques in this study to build the financial news recommendation engine.

B. Machine Learning Approaches to Text Classification

Machine learning (ML) and deep learning (DL) models which worked well in the text classification scenario are selected to employ in the analysis process of this study. Machine learning is a subset of artificial intelligence that allows learning from experiences without directly programming. Under the machine learning terminology, there are 3 broad subcategories supervised learning, unsupervised learning, and reinforcement learning. The classification problems where the response is available for all training data come under the supervised learning principle.

Deciding a single best machine learning model that can be applied to every text classification problem is impractical. The best model differs according to the unique characteristics of each dataset such as the type of data, the size of the data, and the number and distribution of classes among records [10]. Hence this study employs different models to the dataset with the hope of obtaining the best algorithm.

An effective non-parametric approach that is widely used in the text classification field is the k-nearest neighbour (kNN) algorithm. This performs well with binary classification and as well as with multiclass classification. The research article, [11] investigated the compatibility of the kNN approach in text classification and came to know that good results can be obtained regardless of the k factor value in the kNN algorithm. Another research was carried out to classify a collection of speeches of two well-known American politicians Barack Obama and Mitt Romney where the goal of the research was to predict the speaker of a new speech. It showed satisfactory results with kNN and according to its findings, kNN is one of the fastest and simplest ML algorithms used in text classification. But [12] illustrated a controversial viewpoint to it. It says, that as the

kNN algorithm considers all the features, in distance computation to deploy the classification process, kNN takes higher computational time.

Support Vector Machine (SVM) which was introduced into the text classification field by Joachims in 1998 [13], is considered as one of the state-of-the-art algorithms in text classification [14]. As SVM has the ability of handling high dimensional data, it performs very effective and promising results in text classification [15]. SVM was used to classify documents into categories such as sports, business and entertainment and obtained high performance [16]. Ravi Kumar and Anil Kumar developed a new approach to address for text classification by combining Principal Component Analysis with SVM and came to know that higher accuracies can be obtained by using SVM than many other algorithms [17].

The random forest algorithm is another prominent model utilized in text classification tasks. As the structure of the recommendation engine has a high ability of dealing with high dimensional data researchers tend to utilize this with TFIDF and BOW approaches. 30 real world datasets were analysed using a random forest based ensemble method gave a satisfactory result compared to other machine learning models [18]. Another multiclass classification task was done by adding weightages to the features in [19] and [20] and obtained higher accuracies with random forest.

C. Deep Learning Approaches on Text Classification

Deep Learning is a subfield of Machine Learning methods that has captured much attention of researchers in many broad areas such as image processing and natural language processing. At present a trend of using deep learning approaches in text classification can be investigated as they can give precise accuracies in most of the scenarios. Convolutional Neural Network (CNN) is one of the highly popularized Deep Learning models in the text classification field. Research that was done to classify text extracted from electronic instruction manuals, presented CNN works very well on text classification when there are weighted word vectors [21]. Two types of CNN, seq-CNN and bow-CNN were utilized on a text classification problem and came to know that CNN effectively takes the word order of text documents in the classification process [22]. And Multi class problems which have many categories in the response variable, performs well with CNN [23].

Long Short Term Memory (LSTM) algorithm is another Deep Learning approach that works well with short text classifications [24]. Due to the specific structure of the LSTM model, it is capable of carefully identifying the hidden patterns in the data. Hence many research works such as [25] and [26] can be found out where the LSTM outperforms the other classifiers. Our study employs these ML and DL models to answer the above-mentioned research gap.

III. METHODOLOGY

It is mentioned that recommendation engines consist of three main components.

- Data collection
- Recommendation Engine
- User Interface [27]

This research focused on first two components which are considered as the heart of the recommendation engine.

A. Dataset Description

Due to the lack of already collected datasets that are capable of helping to gain the intended outcome of this research, a new, appropriate dataset was constructed by retrieving online information through web scraping technique.

Under this data collection phase, 6 online news websites that are widely used by financial analysts are selected. As mentioned under scope and the limitation section, this study limits only to the technology sector. Hence news articles from technology sector of the selected websites are grabbed using web scraping technique.

The variables exercised in the different news article classification scenarios in the literature are considered to obtain a basic idea of the variables to be extracted from news articles. The researchers, Bashar Al Asaad and Madalina Erascu encountered news article title, date of publication, author's name and the content of the article to a fake news detection problem [28]. Sherif Saad explored another fake news detection approach with the use of variables such as news article text, type, title and date [29]. Another research which was aimed to classify financial news articles as significant and non-significant, attributed variables such as the headline of the article, text content, published date, to obtain the intended outcome [30]. Based on those facts and especially considering the financial analysts' suggestions, to continue with this research below variables are selected and extracted through web scraping. These variables are respectively marked in the figure 2.

1) Title
2) Short Description on the home page
3) Published Date
4) Authors name
5) Content of the article
6) Keywords

In the context of data harvesting from the World Wide Web, the recently popularized method, "web scraping" which is also known as web crawling or web data extraction is the best and effective method. Web scraping is solely extracting the data from websites [31]. In line with a known programming language such as R or Python, the web scraping process helps to decrease the human involvement in data acquisition process as less as possible and assists to accomplish the required task with less human effort by providing a user-friendly environment in much automation procedure.

The way of appearing the content in the webpages is not comfortable enough for purposes such as knowledge extraction or analysis approaches. Hence, to grab the content in web pages into more manageable and structured formats, to fulfil analysis purposes, web scraping is applied. Python programming language is used to fulfill this task.

Then the scraped articles are provided to the group of financial analysts who specialized in the technology sector to obtain the response variable, "relevance level of each article to the financial analysts". And it is taken as a categorical variable with 3 categories where articles in each of the three categories are defined as,

- High relevance articles,

- Middle relevance articles
- Low relevance articles.

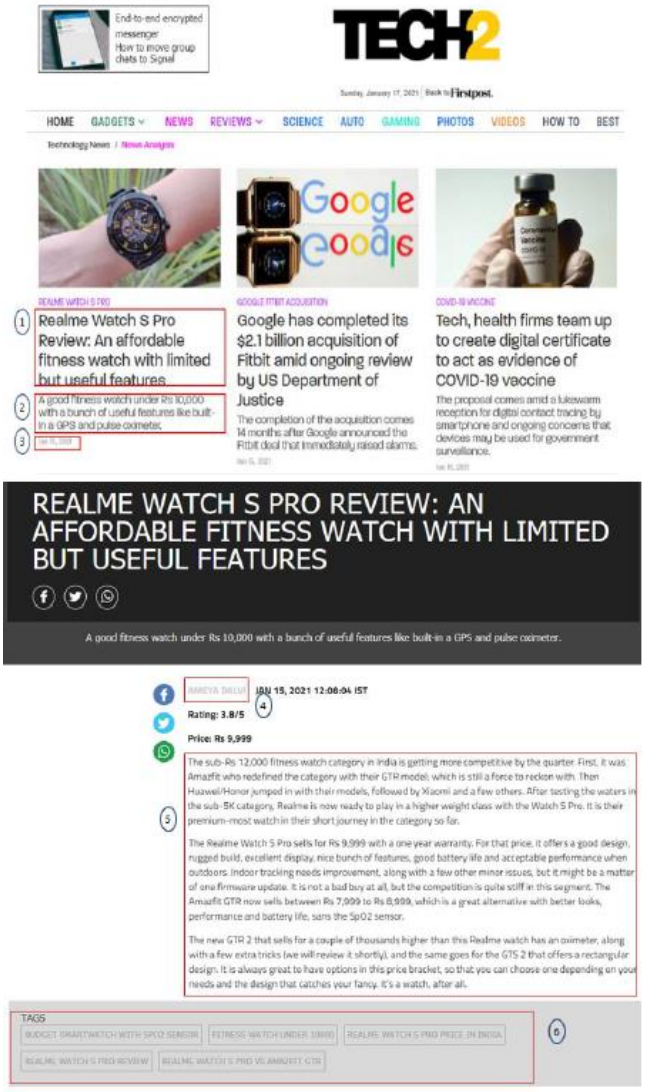

Fig. 2 Illustration of the scraped variable [38]

The group of analysts individually provided the response variable according to their preference level and then the final response for each article is obtained by taking the mode of the individual responses.

As mentioned in the related work section, collaborative filtering and content-based filtering approaches are the vastly seen approaches in building recommendation engines. But this research uses a relatively different approach, "classification model-based approach" in the developing phase of the recommendation engine.

The analysis phase of the recommendation engine comes under the recommendation engine component. Here, NLP (Natural Language processing) techniques are utilized in the analysis phase along with the machine learning and deep learning techniques.

B. Data Pre-processing

Since the intended outcome of this research does not depend on the time effect, the variable, published date was not considered in the analysis part. Furthermore, as the same content in the short description variable could be found again inside in the full content of the article and since it was missing in a considerably high number of articles, that variable was also not considered in the analysis part. So, the final list of variables fed to the algorithms are given in the TABLE 1.

TABLE I

VARIABLES APPLIED TO MODELS

Predictor Variables	Response Variable
Title	Rate – the relevance level of the article (ordinal categorical variable)
Content	
Author_name	
Keywords	
Website_name	

Then, one of the mandatory steps in the NLP task, text data pre-processing is carried out to bring text into more comfortable formats by removing unnecessary complicated formats of texts. The techniques used under this are as follows.

1) Lower-case

This is one of the effective text pre-processing steps that is useful in most of the NLP scenarios to keep the consistency of expected output. As the name illustrated, under this step all the text data are lowercased. The different variations of the same word can be visualized (eg: Launch, launch) due to the capitalization. In the feature extraction step these variations may take as different words and it may lead to creating misleading information when it feeds to ML algorithms.

2) Tokenization

This is a fundamental and common step in data pre-processing tasks. In order to understand the context of the data in the given piece of texts, machine learning algorithms required those texts to be broken down into small units. Tokenization is the method of splitting the texts into smaller units called tokens, which gives an easy understanding of the context to ML algorithms.

3) Stop word removal

Stop words are the common, frequently used words in a language and they are considered insignificant in analysing approaches. Mostly, stop words consist of articles, prepositions and conjunctions which are used to connect the sentences, or they are helped to keep the sentence structure. Some examples for the stop words in the English language are 'the', 'a', 'about', 'above', 'after', 'again', 'against' , 'all' , 'am' , 'an' , 'and' , 'any' , 'are' , 'aren't' , 'as' , 'at' , 'be' , 'because' , 'been'. By removing these insignificant and low informative words the ML algorithms can focus on important words and this helps to reduce the high dimensionality.

4) Lemmatization

Lemmatization returns the base form, which is also called a lemma, of a given word by conducting a proper vocabulary and morphological analysis. This helps to carry out the classification procedure effectively by reducing the unnecessarily increasing attributes. For example, the word "better" is mapped to its root form "good". And the words "launching", "launched", "launches" are mapped to its base form "launch".

C. Data Augmentation

As there exists imbalanced problem in the dataset, after splitting data into text and training sets, text data augmentation techniques are utilized on the training set to solve that problem [32]. Under the text data augmentation, the category with the highest number of records among the categories of the response variable in the training set is considered and that category is equalized through generating synthetic data to the other remaining categories in the training set. Only the synonym replacement technique is utilized as the text data augmentation method in this thesis.

D. Feature Extraction

Machine learning algorithms are not capable of understanding the texts, as they are. Hence in the tasks such as text classification, transforming texts into numerical values in a way that ML algorithms can understand them is a must.

1) Bag of Word (BOW) method

The BOW method is a very simple and flexible algorithm to extract features from a text corpus. This method represents a text by considering whether that word appears in the given text or not. It is called 'Bag' because it does not take any information about the order of the word when constructing the numerical vectors to represent text data. Each unique word in the text corpus is taken as a feature and then checks the presence or absence of that word in each text. [33].

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

TF-IDF is a feature extraction method that gives numerical values to texts by reflecting the importance or relevance of a word to a document in a collection of documents. TF-IDF is a product of two sub parts. This gives much more meaning to the numerical values of each word than the BOW method by adding a weightage to it [34].

$TF(w_i,D_j)$ =
Number of times the word w_i occurs in D_j / Total number of words in the D_j

(1)

$IDF(w_i)$ =
Total number of documents in the corpus / Number of documents contain in the given word w_i

(2)

$$TFIDF = TF * IDF \quad (3)$$

where i is the number of unique words in a given text corpus, j is the number of documents in the corpus

3) N-gram approach

Above mentioned two approaches do not consider the order of the word. To address that issue, the n-gram approach is utilized with the above two methods. Instead of considering one word in constructing the features, n numbers of adjacent words are considered. Then the similar procedures that are mentioned above can be manipulated with the newly obtained feature set.

E. Machine Learning Models

To accomplish the goal of classifying new news articles into relevant categories, supervised machine learning algorithms have to be applied. In this phase, the output of the feature extraction process; the numerical vectors, are fed into machine learning algorithms. Since there is no one best ML algorithm is defined for text classification, in this research several algorithms are applied to the dataset with the hope of discovering the most suitable algorithm which gives the highest accuracy. The models that are selected to apply to the dataset are the ones that give the highest accuracies in different text classification situations in previously done researches.

The selected ML algorithms, KNN (K Nearest Neighbours), SVM (Support Vector Machine), Random Forest, applied on the dataset with the two feature extraction methods BOW and TFIDF approaches.

1) KNN (K Nearest Neighbours)

KNN is a non-parametric algorithm that is widely applicable in text classification scenarios. KNN is considered as an instance-based learning or lazy learning algorithm because the generalization of training data is delayed until the query is made to the algorithm. This simply means, in the training process KNN algorithm only stores the data in multi-dimensional space without learning from it and it starts to classify objects whenever it receives the test data. Because of this nature, researchers recommend applying KNN in recommendation systems as they have continuously updating training sets.

KNN uses 'feature similarity' to classify a new observation into an existing category. When assigning a category to an unknown observation, KNN examines the k nearest neighbour points to that unknown observation in the multi-dimensional feature space. The class that can be mostly seen within the k nearest neighbours is assigned to the new observation. Nearest neighbours are calculated by using distances such as Euclidean, Manhattan, or Hamming distances. The most used method: Euclidean distance is used in this research. The most favourable k value can be identified by inspecting the error plot or accuracy plot with the use of cross-validation. The main steps of KNN can be illustrated below.

- The distance between the test data point and each of the data points in the training dataset is calculated with the help of Euclidean distance.
- Based on the distance values, the training data points are ranked and the k nearest data points are selected by looking at the ranks.
- The frequent class among the k nearest neighbours is assigned to the new data point.

Even though KNN is used widely due to its simplicity and easy interpretability, there are some disadvantages of using this can be investigated in practical situations.

As KNN stores all the training data points in multi-dimensional feature space, it requires high memory storage. And also, since it calculates distances between the new data point and every data point in the training data set, this method is computationally expensive and time-consuming. It is said that KNN is sensitive to the scale of the data. Hence it is necessary to eliminate unnecessary features in the dataset prior to applying the algorithm.

2) SVM (Support Vector Machine)

SVM is one of the robust prediction methods in machine learning algorithms that produces significant accuracies with less computational power in text classification problems [35]. The objective of SVM in classification is to discover the optimal hyper-plane in N-dimensional space (N-Number of features) that distinctly separates the data points into correct class labels.

In its most simple type SVM is used to separate data, into two groups. When it comes to multiclass classification, the same procedure that follows in the binary classification is utilized, after breaking down the multi-classification problem into multiple binary classifications.

SVM maps the data points into high dimensional feature space to gain an optimal hyper plane. Hyper planes can be considered as decision boundaries that classify data points into two distinct categories. The data points on either side of the hyper plane can be classified into one group. Support vectors that fall closer to the hyper plane are the data points that help in finding the optimal hyper plane. The optimal hyper plane maximizes the margin from both classes of training data and maximizes the distance between the nearest points of each class. When data points show complex patterns, to separate the points into distinct classes, different kernel functions such as linear, sigmoid, polynomial, rbf are used.

In most machine learning algorithms, when the number of features increase, the amount of calculation and required storage capacity increase. But as SVM uses only support vectors for calculations, it excels in handling high dimensional inputs, and it does not face the curse of dimensionality problem.

SVM takes much time when dealing with large datasets mainly due to the kernel function calculations and finding the optimal hyper plane in higher dimensions. And SVM does not perform well with overlapping classes.

3) Random Forest

Random forest is a machine learning algorithm that can be applied for both regression and classification problems and this research uses the random forest model for classification purposes. In NLP tasks, especially in text categorization scenarios, researchers have utilized random forest as they exhibit prominent performances with high dimensional data. Due to its algorithmic simplicity, this becomes a popular machine learning algorithm among the text classification research community [19].

Random forest is an ensemble learning algorithm that leverages the power of multiple decision trees to obtain error free predictions. Each of the decision trees in the random forest model depends on the randomly sampled vector from the dataset. In the training process, at each node to select the best splitting variable, m numbers of variables are randomly selected from the p number of all predictor variables. (Where $m<p$) This helps to reduce the correlation among trees and finally, it gives a lower variance. As it selects a random sample of predictor variables, datasets with larger dimensions behave efficiently with the random forest model. As this research applies random forest for the classification task, classification trees are ensemble together and construct the random forest and, in the tree growing phase impurity measurements such as ‘Gini index’ or ‘cross entropy’, are used to evaluate the splitting variable and the splitting value. After each classification tree completes the prediction, the majority vote among the predicted results is taken as the final class. To obtain the optimal model, there are many hyper parameters to optimize in the random forest model and those are described in the implementation of the random forest section.

F. Deep Learning Models

The deep learning approach, a subset of machine learning context, is greatly employed in NLP tasks due to its surprisingly high performances. These deep learning models which are based on artificial neural networks (ANN) are capable of uncovering hidden, complex patterns from unstructured data through a self-adaptive algorithm. As the model learns the patterns laid in data itself, to produce better results large datasets are required. Artificial neural networks are inspired by the way the human brain works. ANN consists of 3 layers: input layer, hidden layer, and output layer. Hence, with the hope of achieving higher accuracies, two deep learning models, LSTM (Long Short Term Memory Model) and CNN (Convolutional Neural Network) are applied to the dataset.

1) LSTM

The Recurrent Neural Network (RNN) forms a directed circle between neurons which enables to keep an internal memory by taking the output of a neuron as an input of the same neuron in the next instance. With a larger sequence of words RNN models always victimize to the problem of vanishing gradient which causes forgetting long term dependencies. The gradient that is used to update the weights of NNs shrinks as it back propagates through time and the gradient value gets extremely small which cause to keep only a short term memory. This is called vanishing gradient problem. This research uses the content of the news articles, and they consist of large sequences of data. Hence RNN is not appropriate in this context.

A generalization of the RNN model, a Long Short Term Memory (LSTM) network directly addresses this issue and provides specialization on remembering information for extended periods [36]. As the LSTM network is trained using Back propagation Through Time (BPTT) it helps to reduce the vanishing gradient problem.

In LSTM model, after initiating the sequential model as the first layer, the embedding layer is added to the model. In the training process embedding layer itself learn the patterns and characteristics hidden in the text data and constructed an indigenous word vocabulary for the dataset where each of the tokens is represented as a numerical vector in the vocabulary. The size of that numerical vector can be changed according to the requirements by changing the parameter embedding_dim in the embedding layer. Then a dropout layer is added to prevent any overfitting. Then a flatten layer, a dense layer with ReLu activation function, again a dropout layer and finally dense layer with softmax activation function are added to the model respectively. Flatten layer is used to map the output of the LSTM feature into a single column that is connected to a dense layer. As there exist three categories in the dataset, three neurons are included in the final dense layer.

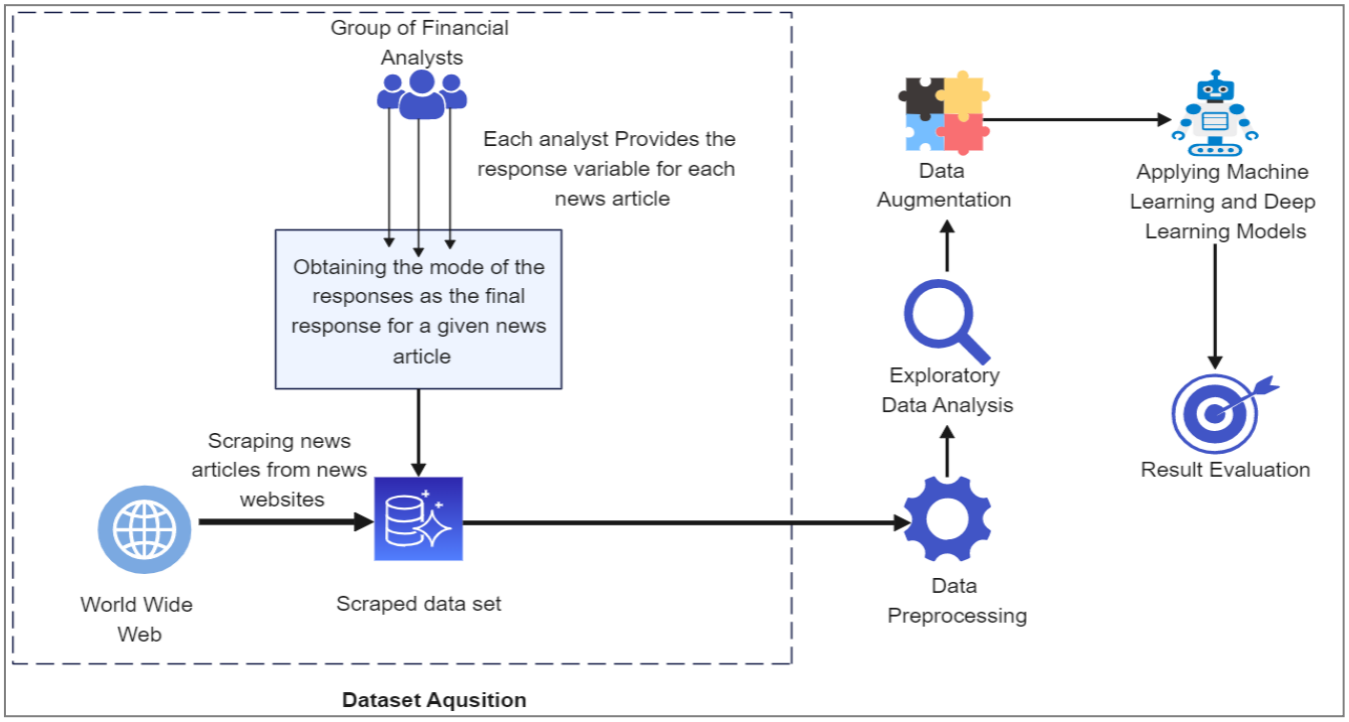

Fig. 3 Implementation process diagram

2) CNN

This is considered as the state-of-the-art model in text classification. There is much evidence to prove that fact as researchers obtained breakthrough results in NLP tasks with the CNN model. In the beginning, CNN was mainly developed for the image classification community. But after the researcher Yoon Kim exhibited the valuable discovery by showing the fact that CNN can be utilized successfully in text classification scenarios, the research community focus their attention on experimenting CNN with different NLP approaches [37].

Fig. 4 Implementation process diagram

Constructing the CNN is also started by initiating the sequential model. Same as in the LSTM model the first layer is the embedding layer which performs the same task as described in the LSTM section. Along with the embedding layer, Conv1D layer, GlobalMaxPooling1D layer, Dropout layer, and Dense layer are added to the model respectively. The objective of including the dropout layer is the avoiding over fitting. GlobalMaxPooling layer is primarily included to reduce dimensions of convolutional layers feature map output. The final dense layer contains 3 neurons as there are three categories in the dataset.

As both models are constructed using the sequential approach, Adam optimization algorithm is used to optimize the deep neural network model. As this study is based on a multiclass classification problem, the loss function; categorical cross entropy is employed to quantify the deep learning model error.

Optimal parameters for all the applied models are selected through grid search.

G. Model Evaluation

Since the dataset utilized in this research is an imbalanced dataset, using only the accuracy for evaluating model performances is not adequate. This happens because the overwhelming number of examples from the majority category will overpower the number of examples in the minority class. Therefore, to assess the performances of the models with the imbalanced dataset, the measurements such as precision and recall are used together with accuracy. When there are multiple classes the precision and recall are calculated separately for each class. The process that follows in implementing the recommendation engine is illustrated in figure 3.

The main intention of following the above-mentioned procedure is to solve two research questions. The primary research question is to identify the most suitable machine learning or deep learning model for the analysis part of the recommendation engine. The secondary research question is to utilize a descriptive analysis to identify the most prominent keywords in each of the categories.

IV. Results

The word cloud provides a novelty way to represent textual data. It visualizes textual data in varying sizes depending on the frequency of appearing each word. In order to achieve the secondary objective, word clouds are drawn. There are few words which appear in all three categories with the same probability. Therefore, those words are removed, and word clouds are constructed to each relevance category separately, to check whether there exist any distinct predominant words in categories. Figures 4, 5, 6 presented the word clouds for ‘Low’, ‘Middle’ and High relevancy categories separately. Even though it is difficult to identify dominant words in categories clearly, it is noticeable that the ‘Middle’, and ‘High’ categories contain some high-ended company names such as ‘apple’, ‘amazon’, ‘microsoft’ with larger frequencies with comparison to ‘Low’ relevance category.

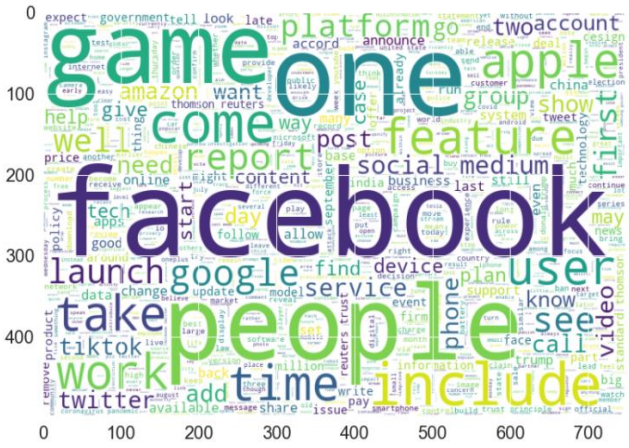

Fig. 5 Word cloud for the 'Low Relevance' Category

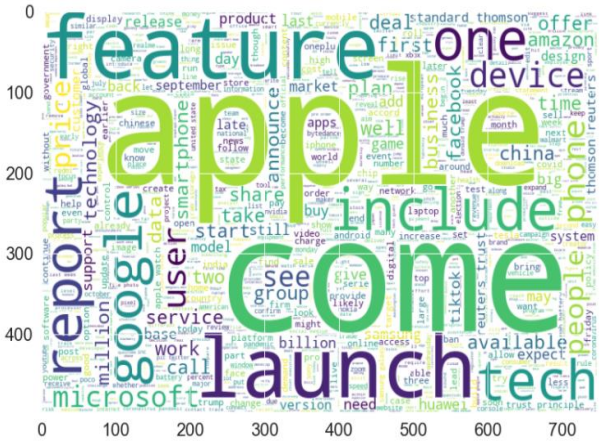

Fig. 7 Word cloud for the 'High Relevance' Category

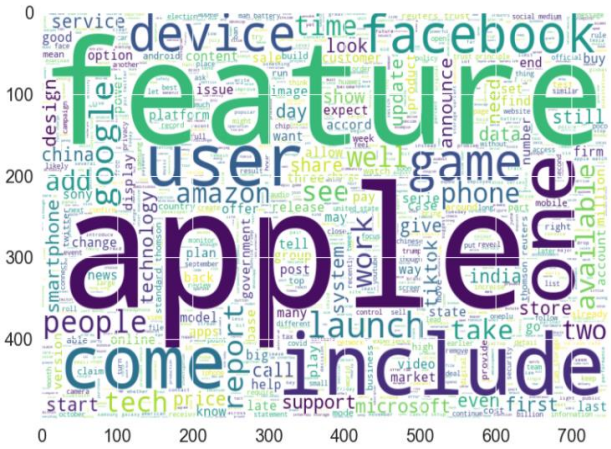

Fig. 6 Word cloud for the 'Middle Relevance' Category

With the hope of finding the solution for the primary research question, advanced analysis techniques are used with the balanced dataset. The obtained results of the dataset with each of the mentioned algorithms were compared using different matrices such as accuracy, precision, and recall. As there's not much deviation between precision and recall of each model, the accuracies obtained from each algorithm are shown in TABLE II.

It can be clearly seen that performance of CNN model outperforms the other models. The learning curves of the best model are shown in figure 7 and the precision and recall of each class of the best model are depicted in TABLE III.

TABLE II

PERFORMANCES OF EACH ALGORITHM WITH THE DATASET

Machine learning algorithm	KNN with BOW	SVM with BOW	Random Forest with BOW	KNN with TFIDF	SVM with TFIDF	Random Forest with TFIDF	CNN	LSTM
Accuracy	54%	54%	56.2%	54.5%	56%	56.6%	64.3%	62%

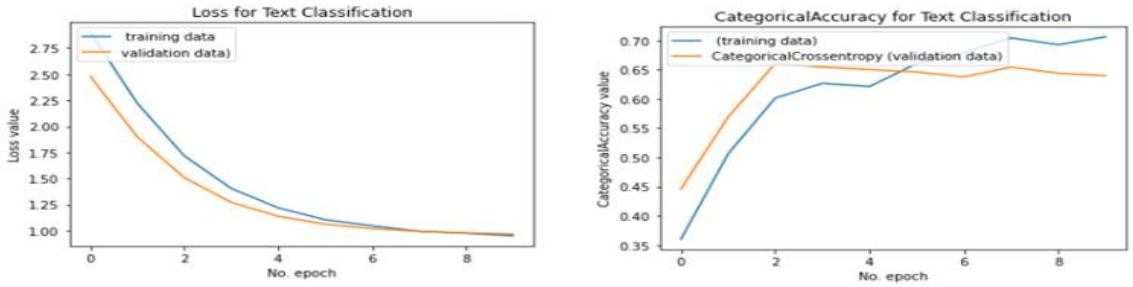

Fig. 8 Learning curves of the best model (CNN)

TABLE III

CLASSIFICATION REPORT OF THE BEST MODEL (CNN)

	Precision	Recall	Accuracy
Low relevance	0.71	0.61	64.3%
Middle relevance	0.61	0.68	
High relevance	0.46	0.54	

V. Discussion

According to the achieved results DL models indicate higher performances than the ML models. Furthermore, according to TABLE I all the three ML models accuracies attained with TFIDF approach are higher than the accuracies with BOW approach. It can be clearly understood because when calculating the TFIDF values it adds some weightage to each word in the corpus and hence it grabs more information about texts with comparison to the BOW [28].

Though, this study applied a dropout layer in CNN model with the hope of avoiding over fitting, it seems that it is not adequate to prevent over fitting. Hence going for other remedies to overcome the problem is necessary. The research [25] described another way to avoid over fitting called weight decay. It is commonly done in the form of L2 regularization (Ridge regression) which adds a penalty for big weights to the cost function of the neural network. This method can be utilized in the future to overcome the existing issue.

A. Limitations of the Study

The response variable of this study which is the relevance level of the news article is marked manually by the financial analysts. But, due to the time constraint, even though having a large dataset is an adornment to researchers because it leads to extract more insights, obtaining a larger dataset becomes a challenging aspect here. So, this limitation directly affects to the accuracy level of the algorithms. As future works, it's suggested to obtain a larger dataset to have considerably good accuracies.

It is true that if there exist a considerably large number of analysts to mark the response variable then it helps to increase the generalizable ability of the research to the whole population of financial analysts. However, again due to the time constraint, it was not an easy task. A group of analysts from a well renowned financial company in Sri Lanka, created the response variable of this research. Therefore, as the title suggested this research was mainly aimed towards the responses of the analysts of that company.

Even though there exist many sectors that financial analysts focus on, within the given time it is not possible to pay attention to all those sectors. Because of that, this study limits only to the technology sector. The scope of this research relies on the financial analysts in the selected well renowned financial company and on technology sector news articles. When expanding this research, it's better to consider a broad area to obtain more generalized solutions.

VI. Conclusion

In this real-world dataset, the CNN algorithm outperforms the LSTM, KNN, SVM and Random Forest. Hence, to develop the framework of the recommendation engine the CNN classifier can be applied. Deep neural networks: CNN and LSTM are more likely to detect hidden patterns laid in the utilized dataset of this study with comparison to machine learning models; KNN, SVM, and Random Forest. In conclusion, the CNN model with the balanced training set gives satisfactory performances. Hence the main objective of this thesis; developing a modelling framework of a news recommendation engine which filters out the most relevant news articles for financial analysts who mainly focus on technology sector, can be achieved by using web scraping technique in the data collection phase and the CNN algorithm in the analysis phase.

VII. Future Works

1) Developing the User Interface component for the constructed recommendation engine of this study.

2) DL models always require larger datasets to provide high performances. Hence, obtaining more manually annotated data for the text corpus, and achieving better and more accurate insights with deep learning models.

3) Extending the scope of the study without being limited to one company and developing more generalizable recommendation engine that can provide service to whole financial analysts' community.

4) Utilizing more advanced algorithms such as ensemble deep neural networks to obtain better predictive performances.

Acknowledgement

I extend my extreme gratitude towards my research supervisor, Dr. Nadeeka Basnayake for her guidance, assistance and support rendered in successfully concluding this study. Further I would like to express my sincere gratitude to my co supervisors Ms. Desika Hureekaduwa and Ms. Beshani Weralupitiya from Acuity Knowledge Partners, Sri Lanka for their continuous guidance, assistance and support extended to me during the entire process. Especially, I would like to thank Mr. Rohan Fernando and his financial analysts' team from Acuity Knowledge Partners for the immense support they provided by manually marking the relevance level of each article.

References

[1] G. Gidófalvi, "Using News Articles to Predict Stock Price Movements," no. March 2015, 2019.

[2] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, 2009, doi: 10.1145/1462198.1462204.

[3] C. Feng, M. Khan, A. U. Rahman, and A. Ahmad, "News Recommendation Systems-Accomplishments, Challenges Future Directions," *IEEE Access*, vol. 8, pp. 16702–16725, 2020, doi: 10.1109/ACCESS.2020.2967792.

[4] B. Fortuna, C. Fortuna, and D. Mladenić, "Real-time news recommender system," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6323 LNAI, no. PART 3, pp. 583–586, 2010, doi: 10.1007/978-3-642-15939-

8_38.
[5] A. Felfernig, M. Jeran, G. Ninaus, F. Reinfrank, S. Reiterer, and M. Stettinger, “Basic approaches in recommendation systems,” *Recomm. Syst. Softw. Eng.*, pp. 15–37, 2014, doi: 10.1007/978-3-642-45135-5_2.
[6] H. L. Borges and A. C. Lorena, “A survey on recommender systems for news data,” *Stud. Comput. Intell.*, vol. 260, pp. 129–151, 2010, doi: 10.1007/978-3-642-04584-4_6.
[7] A. H. Parizi, M. Kazemifard, and M. Asghari, “Emonews: An emotional news recommender system,” *J. Digit. Inf. Manag.*, vol. 14, no. 6, pp. 392–402, 2016.
[8] U. Mohiuddin, H. Ahmed, and M. Ismail, “NEWSD: A Realtime News Classification Engine for Web Streaming Data,” no. Racs 2015, pp. 61–66, 2016, doi: 10.2991/racs-15.2016.10.
[9] W. Gu, S. Dong, Z. Zeng, and J. He, “An effective news recommendation method for microblog user,” *Sci. World J.*, vol. 2014, 2014, doi: 10.1155/2014/907515.
[10] M. F. Garcia-constantino, “On The Use Of Text Classification Methods For Text Summarisation,” no. July, 2013.
[11] B. Trstenjak, S. Mikac, and D. Donko, “KNN with TF-IDF Based Framework for Text Categorization,” vol. 69, pp. 1356–1364, 2014, doi: 10.1016/j.proeng.2014.03.129.
[12] B. Baharudin, L. H. Lee, and K. Khan, “A Review of Machine Learning Algorithms for Text-Documents Classification,” *J. Adv. Inf. Technol.*, vol. 1, no. 1, 2010, doi: 10.4304/jait.1.1.4-20.
[13] T. Joachims, “Text Categorization with Support Vector Machines : Learning with Many Relevant Features,” pp. 2–7, 1998.
[14] F. Van Meeuwen, “Multi-label text classification of news articles for ASDMedia,” no. August, 2013.
[15] M. Thangaraj, “T EXT C LASSIFICATION T ECHNIQUES : A L ITERATURE R EVIEW,” vol. 13, pp. 117–135, 2018.
[16] S. Mayor and B. Pant, “Document Classification Using Support Vector Machine,” vol. 4, no. 04, pp. 1741–1745, 2012.
[17] A. R. Kumar and G. A. Kumar, “Support Vector Machine for Text Categorization using Principle Component Analysis in Data Mining,” no. 5, pp. 3164–3167, 2020, doi: 10.35940/ijrte.D7350.018520.
[18] M. Z. Islam, J. Liu, J. Li, L. Liu, and W. Kang, “A semantics aware random forest for text classification,” *Int. Conf. Inf. Knowl. Manag. Proc.*, no. November, pp. 1061–1070, 2019, doi: 10.1145/3357384.3357891.
[19] B. Xu, X. Guo, Y. Ye, and J. Cheng, “An improved random forest classifier for text categorization,” *J. Comput.*, vol. 7, no. 12, pp. 2913–2920, 2012, doi: 10.4304/jcp.7.12.2913-2920.
[20] D. Liparas, Y. Hacohen-kerner, and A. Moumtzidou, “News Articles Classification Using Random Forests and Weighted Multimodal Features,” *In Proceedings of the 7th Information Retrieval Facility Conference, IRFC 2014*, Copenhagen, Denmark, November 10-12, 2014, pages 63-75. Springer, Cham, 2014.
[21] P. Song, C. Geng, and Z. Li, “Research on Text Classification Based on Convolutional Neural Network,” *Proc. - 2nd Int. Conf. Comput. Network, Electron. Autom. ICCNEA 2019*, pp. 229–232, 2019, doi: 10.1109/ICCNEA.2019.00052.
[22] R. Johnson and T. Zhang, “Effective use of word order for text categorization with convolutional neural networks,” *NAACL HLT 2015 - 2015 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, pp. 103–112, 2015, doi: 10.3115/v1/n15-1011.
[23] J. Liu, W. C. Chang, Y. Wu, and Y. Yang, “Deep learning for extreme multi-label text classification,” *SIGIR 2017 - Proc. 40th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 115–124, 2017, doi: 10.1145/3077136.3080834.
[24] J. Nowak, A. Taspinar, and R. Scherer, “LSTM recurrent neural networks for short text and sentiment classification,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10246 LNAI, pp. 553–562, 2017, doi: 10.1007/978-3-319-59060-8_50.
[25] A. Grünwald and S. Rauf Ahmad, “Applications of Deep Learning in Text Classification for Highly Multiclass Data,” 2019.
[26] S. L. Supervisor, A. Itoo, and D. Science, “Fake News Detection Using Machine Learning Author : Simon Lorent Supervisor : Ashwin Itoo A thesis presented for the degree of Master in Data Science,” 2019.
[27] S. Athalye, “Recommendation System for News Reader,” 2013.
[28] B. Al Asaad, “A Tool for Fake News Detection,” *In 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)* (pp. 39-46). IEEE.
[29] H. Ahmed, I. Traore, and S. Saad, “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques,” no. December, 2017, doi: 10.1007/978-3-319-69155-8.
[30] S. Yıldırım, D. Jothimani, C. Kavaklıo˘glu, and A. Bas¸ar, “Classification of ‘Hot News’ for Financial Forecast Using NLP Techniques.pdf.” 2018.
[31] G. Boeing and P. Waddell, “New Insights into Rental Housing Markets Across the United States: Web Scraping and Analyzing Craigslist Rental Listings,” *SSRN Electron. J.*, no. October, 2018, doi: 10.2139/ssrn.2781297.
[32] M. Ibrahim, M. Torki, and N. El-Makky, “Imbalanced Toxic Comments Classification Using Data Augmentation and Deep Learning,” *Proc. - 17th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2018*, no. November, pp. 875–878, 2019, doi: 10.1109/ICMLA.2018.00141.
[33] George, S. J., & Joseph, V. "Text classification by augmenting bag of words (bow) representation with co-occurrence feature". *In 2019 International Conference on Advances in Computing, Communication and Computing Technologies (ICACCCT)* (pp. 1-6). IEEE. doi: 10.9790/0661-16153438.
[34] S. Qaiser and R. Ali, “Text Mining : Use of TF-IDF to Examine the Relevance of Words to Documents Text." *International Journal of Computer Applications*, 181(1), 25-29 July, 2018, doi: 10.5120/ijca2018917395.
[35] A. Borg, A. Borg, and M. Boldt, “E-mail classification with machine learning and word embeddings for improved customer support,” *Neural Comput. Appl.*, vol. 4, 2020, doi: 10.1007/s00521-020-05058-4.
[36] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[37] Kim, Yoon. "Convolutional Neural Networks for Sentence Classification." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014. doi:10.18653/v1/D14-1181.
[38] A. Dalvi, “REALME WATCH S PRO REVIEW”, Firstpost, January 2021. [Online]. Available: https://www.firstpost.com/tech/news-analysis/realme-watch-s-pro-review-an-affordable-fitness-watch-with-limited-but-useful-features-9204461.html