Applications of Artificial Intelligence for Information Diffusion Prediction: Regression-based Key Features Models

Majed Algarni*, Mohamed Maher Ben Ismail
King Saud University, College of Computer and Information Sciences
Department of Computer Science, Saudi Arabia

Abstract—**Information diffusion prediction is essential in marketing, advertising, and public health. Public health officials may avoid disease outbreaks, and businesses can optimize marketing campaigns and target audiences. Information diffusion prediction helps identify influential nodes in social networks, enabling targeted interventions to spread positive messages or counter misinformation. Organizations can make informed decisions and improve society by analyzing information propagation patterns. This research study investigates the prediction of information diffusion on social media platforms using a diverse set of features and advanced machine learning and deep learning models. We explore the impact of network structure, early retweet dynamics, and tweet content on social media, provided by the publicly available dataset Weibo, a social network like Twitter. By applying the training of the models on set of features separately, we observed different performances. The Random Forest model using all features achieved an R-squared of 76.690%. The Random Forest (RF) model focusing on the following network structure achieved an R-squared of 90.773%. The RF model analyzing the retweeting network structure achieved an R-squared of 98.161%.**

Keywords—Information diffusion; social media data; machine learning; deep learning

I. INTRODUCTION

In our current digital age, the number of Internet users has increased, which has been accompanied by an increase in the number of users of social media. This was reflected significantly in the speed of exchanging and sharing information through social media platforms. Information is divided by its nature into useful and harmful, directly affecting society in economic, political and commercial terms [1]. The dissemination of this information can be studied and predicted in terms of and is affected by many variables, such as the timing of the publication [2], as well as the content of the publication, and the characteristics of the participating users [3,4]. A comprehensive understanding of the dynamics of information spread on social media platforms can provide valuable insights in various domains, including marketing, disaster management, and community detection. The dissemination of information via social media is a multifaceted process that relies not only on the substance of the communicated content but also on the method of transmission. The comprehension of the mechanisms by which information is conveyed through these networks is progressively imperative due to the escalating migration of various aspects of our existence to online platforms. The spread of information through social media sites like Twitter and Facebook has greatly impacted. The study of how information spreads on these platforms is complex and constantly changing, which is why it's attracted so much attention from researchers. By understanding how information spreads, we can predict how popular information will become, maximize the influence spread, and monitor how cascades develop. On Twitter, users can "retweet" or share a tweet with their followers, which helps information spread even faster. The Popularity of a tweet is measured by its content and how many times it's been retweeted. By understanding how information spreads on social media, platforms can better monitor and guide the spread of information in different scenarios like online advertising, viral marketing, and fake news detection. Several approaches to predicting information diffusion exist but machine learning and deep learning techniques [5, 6] have recently shown much promise.

In this study, the authors proposed Machine learning and deep learning models to improve our forecasts using a publicly available dataset provided by authors in [7]. This dataset was collected from Weibo, a social network like Twitter. The authors applied the models to features such as following network structure, retweeting network structure, early-stage popularity dynamics, and tweet content. Our study offers significant contributions to the domain of predicting information diffusion on social media platforms.

1) Comprehensive examination: Our study delves into the dissemination of information on social media platforms through the utilization of an extensive array of features and sophisticated machine learning and deep learning models.

2) An examination of essential factors for comprehension: In this study, we investigate the influence of network structure, early retweet dynamics, and tweet content on information diffusion. Through our analysis, we aim to gain valuable insights into the underlying factors that contribute to information propagation.

3) Model performance: The models trained using distinct sets of features in order to evaluate their performance. The RF model based on all features demonstrated an R-squared value of 76.690%, whereas the model that specifically targeted the indicated network structure exhibited an R-squared value of 90.773%. The RF model utilized for analyzing the retweeting

*Corresponding Author.

network structure exhibited a high R-squared value of 98.161%.

4) Real-world applications: It refers to the practical utilization of knowledge, theories, and concepts in various fields. These applications involve The results of our study have the potential to assist the organizations in enhancing their decision-making processes, contribute to societal improvement by facilitating comprehension of information diffusion, and enable the identification of influential individuals within social networks for the purpose of implementing targeted interventions. The aforementioned findings carry significant ramifications for the fields of marketing, advertising, and public health. The remains sections of this research article are as follows: a comprehensive review of related work in Section II, followed by the methodology in Section III. Section IV delves into evaluation metrics. Experimental results, discussion, and conclusion in Section V, Section VI and Section VII respectively.

II. Related Work

A. *Background of the Information Diffusion*

Sharing content on social media platforms like Twitter and Facebook spreads information. Social media is about sharing content. Businesses, advertisers, and marketers should address this important issue. They may effectively communicate with their target audience, strategically advertise their products and services, and remain competitive. Social media is crucial for marketing. Companies can grow their market and build client relationships [8]. Commercial organizations can improve their business strategies, increasing their expertise in efficiently spreading their message and news, and leverage social media platforms to amplify their brand by researching information dissemination that helps promote their products and services further in this section. In information diffusion, there are three main elements Sender, Receiver, and Medium, as shown in Fig. 1. The sender, whether an individual or a group, initiates the dispersion process. They are the ones who are in charge of spreading the information that needs to be spread. The receiver, or group of receivers, denotes the individuals who have received the information that has been spread. The process of dispersion generally has a wider scope of influence, as evidenced by the fact that the number of receivers usually be more than the number of senders. The medium is the way information about diffusion travels from sender to receiver. There are many communication channels for information spreading such as TV, newspapers, and social media platforms like Twitter. Additionally, personal connections play a dynamic role in communication. Furthermore, factors such as airborne diseases can also impact the way information is disseminated. Understanding diffusion is crucial process. It shows us how information spreads and changes behavior in a system. Researchers can study the sender-receiver relationship and the medium used. This helps them learn about diffusion and find ways to manage and control it.

B. *Literature Review*

Twitter has become a powerful tool for spreading information. People can share information with followers and beyond through retweets, posts, and hashtags. Various research studies have observed what makes people retweet and how information spreads on these networks. Factors can include the tweet's content, the source of the information, emotional appeal, timing, and social influence. Researchers found Twitter as a way to study diffusion patterns and how different factors affect information spread. But knowing how information spreads efficiently on Twitter is still a challenge. Machine learning and deep learning techniques are being used to analyse and predict diffusion patterns on Twitter. These techniques can find hidden patterns in large datasets. They can help us understand what drives information to spread.

The Research domain of information diffusion on Twitter has a variety of approaches. Study conducted by authors [9] aimed to determine the characteristics that predict the level of engagement a tweet receives. The findings suggest that tweets with positive sentiment and positive arousal receive more retweets and favourites, although the effect size is small. Predicting information diffusion has been another focal point of research. A study by authors in [10] suggest that modelled tweet popularity as the number of retweets and developed machine learning models predict the same, achieving an accuracy of up to 60% and an F1 score of 67%. Another researcher proposed approach to predict information diffusion based on user-based, time-based, and content-based features, resulting in a model that improves the F-measure by about 5% compared to the state-of-the-art Masud et al. Authors in [5] in their study Twitter hate speech based-on topics. The researchers analyze a massive collection of tweets, retweets, user activity logs, and follower networks. They also collect online news stories. The authors provide feature-rich methods for predicting hashtag-related hate speech. The best model scored 0.65 macro F1. RETINA, a neural network for Twitter retweet prediction, is also presented. This design has a macro F1-score of 0.85. Their study sheds light on how Twitter users start hate speech and spread it through retweets. Authors in study [11], introduce an approach for predicting the spread of information prompted by a Twitter user's tweet. They leverage six user features, like number of followers and tweet frequency, along with 80 linguistic and psychological aspects provided by the Linguistic Inquiry and Word Count (LIWC) software. Their approach comprises a module that formulates regular basic tweet propagation patterns and a classifier that anticipates the tweet pattern associated with a specific user tweet. The study uses Tree-Shaped Tweet Cascades for the dataset, achieving significant accuracy in predictions, notably an F-measure of 0.89 with the JRIP model.

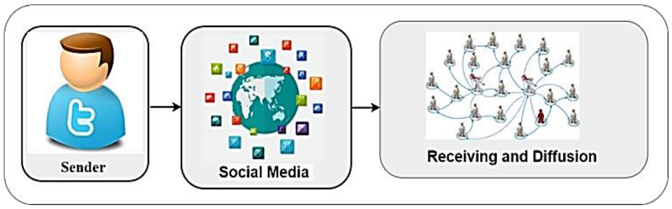


Fig. 1. Demonstrating the information diffusion process on social media.

The propagation on Twitter is the current problem, with a particular emphasis on influence and prediction. The authors of [3] examine Twitter's information diffusion model, which sees the spread of information on the platform as a problem with several variables over time. It focuses on tweet quantity, emotional tone, and influence. Time series clustering reveals Twitter information transmission patterns. The study clusters hashtags with similar patterns using time series clustering methods. This study forecasts three-dimensional parameters using Autoregressive Integrated Moving Average and LSTM linear and non-linear time series models. LSTM models attained an accuracy of 80%.

Previous studies have directed their attention towards more granular facets of information dissemination on the Twitter platform. The study proposed an information diffusion model that conceptualized information diffusion as a problem of multivariate time series analysis. This model specifically addressed the variables of tweet volume, tweet sentiment, and tweet influence [12]. In their study, the authors introduced an algorithm named SentiDiff, which integrates textual data and sentiment diffusion patterns to effectively forecast sentiment polarities conveyed in Twitter messages. The algorithm that has been suggested demonstrates improvements in PR-AUC, ranging from 5.09% to 8.38% for classification tasks, in comparison to existing sentiment analysis algorithms that rely on textual information [12, 13].

Other approaches are based on the anticipated graph information diffusion node activations; the Topological Recurrent Neural Network (Topo-LSTM) was introduced in [14]. Diffusion topologies are used to characterize the cascade structure. This DAG model depicts the cascade. The Topo-LSTM is a diffusion-prediction-specific LSTM architecture. A MAP increases from 20.1% to 56.6%. The researchers found that using dynamic directed acyclic graphs (DAGs) with the innovative data model and Topo-LSTM architecture improves diffusion structure representation.

Authors in [15] presented a model that predicts how information flows on Twitter. The model learns the probability of influence between users. It considers both time-based and structural aspects of influence spread. The effectiveness of the suggested models is examined on two datasets, Darwin and MelCup17. The TDD-CP model achieved a balanced precision of 94.64% and a recall of 95.9% on the Darwin dataset. The MelCup17 dataset achieved a balanced precision of 95.13% and a recall of 98.2Authors in the study [16] developed a Twitter information spread model. To improve predictions, the model has used custom-weighted word characteristics. A Custom Weighted Word Embedding is proposed to measure content diffusion via retweets. Twitter postings are used to extract lexical units, build a matrix with word sequences, and apply specific weights based on the sentence's presence index. Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) are used to predict information dissemination and improve accuracy and training time. The CWWE framework improves the deep learning framework model accuracy, according to experiments. The researchers collected 230,000 tweets from over 45,000 users over six months. Model accuracy increased from 53% to 80%. The authors in [32] introduced the LARM (Lifetime Aware Regression Model), a ground-breaking method for tackling the problem of online content popularity long-term prediction in dynamic YouTube networks. One way that LARM sets itself apart is by considering content longevity as a significant aspect, which helps to mitigate the drawbacks that come with a large amount of historical data and inappropriate model assumptions. The model exhibits flexibility to different observation intervals and is fitted using a forecast lifetime metric that is derived from early-accessible variables. The varied lives of video content are accommodated by specialized regression models. Based on two YouTube datasets, the experimental findings demonstrated the significant advantage of LARM, with prediction error reductions of up to 20% and 18%, respectively.

Despite the considerable body of research that has been conducted on the phenomenon of information diffusion on the social media platform Twitter, there remains a notable gap in the existing literature that necessitates further investigation and exploration. The authors in [34] explored the intricacies of information dissemination on social media, focusing on the Sina microblog controversy around the L group Double 11 fraudulent advertising incidents. Using a strong data analytics methodology that combines time series regression and data mining, the study reveals the key variables influencing the dissemination of information. User activity, emotional shifts, and media attention have been recognized as important drivers, with sentiment polarity and reposting being critical factors in various dissemination phases.

The area of utilizing the number of retweets and retweeted counts, in combination with advanced machine learning techniques, is currently lacking in comprehensive exploration. The comprehension and anticipation of diverse aspects of information diffusion, encompassing the factors that impact the dissemination of information and the dynamics of diffusion patterns, constitute significant areas of scholarly inquiry that merit consideration. This study addresses the identified research gap by proposing using machine learning and deep learning models. Through utilizing these models, significant insights can be obtained regarding the intricacies of information dissemination on the social media platform Twitter. This research endeavor aims to augment our comprehension of the mechanisms through which information propagates on social media platforms. This pursuit aims to establish a basis for enhanced prediction and awareness of the intricate dynamics of disseminating information.

III. Methodology

This study proposed a methodology for predicting information diffusion in social media based on retweets and retweeted counts. Deep learning and machine learning techniques are used to predict information dissemination. The methodology involves data preparation and training models on large-scale social media datasets. The choice of the features and the proposed framework used in our study was a carefully considered process aimed at comprehensively understanding social media information diffusion. The elected features, such as Following Network Structure (FNet), Retweeting Network Structure (RNet), and Early Popularity (early), were strategically picked to capture dissimilar dimensions of the diffusion phenomenon. The proposed framework of the information diffusion prediction is demonstrated in Fig. 2.

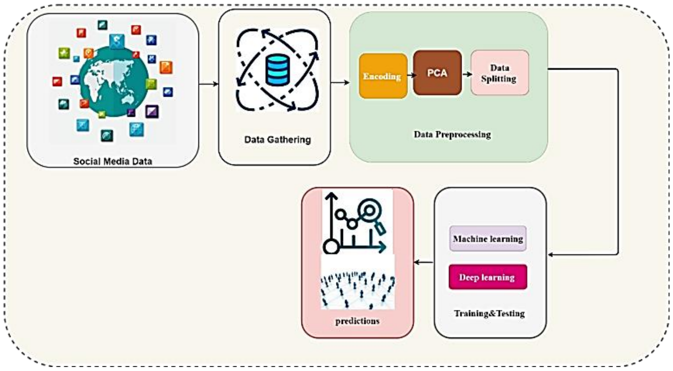


Fig. 2. The proposed framework for information diffusion prediction.

A. Data Source

Sina Weibo is a popular social network site like Twitter. We used a large dataset from Sina Weibo [7] for our research. The dataset provides by DataCastle. It has 30,010 original tweets. Each tweet has its own ID. The ID has information like the user's ID, the time the tweet was posted, and the tweet's content. The tweets were posted between January 1, 2015 and May 20, 2016. This dataset was developed by authors in [7]. It is publicly available online. The dataset is large, so we can study how information spreads on Weibo. We can learn more about predicting information cascade scales based on analyzing set of features.

B. Feature Extraction

The dataset developers categorized data features into four groups: following network structure, retweeting network structure, early-stage popularity dynamics, and tweet content. We aim to capture various aspects of information diffusion dynamics by extracting and incorporating these features.

1) Following network structure: The following network structure plays a crucial role in information diffusion. The constructs a network (G_F) based on the relationships among users who follow each other. Five network structural features are derived from measuring user influence in this network, including out_degree_F, in_degree_F, all_degree_F, bi_degree_F, and pagerank_F, as described in Table I [7].

2) Retweeting network structure: Retweeting is a key mechanism for information diffusion. The builds a retweeting network (G_R) using retweet data from the first 60 minutes. Similar to the following network, by extracts five network structural features to quantify user influence in the retweeting network: out_degree_R, in_degree_R, all_degree_R, bi_degree_R, and pagerank_R, as described in Table I [7].

3) Early retweet dynamics: The early-stage retweet time series contains valuable temporal information. We construct four types of temporal features to capture trends and fluctuations in cascade sizes over time. These features include cascade (cumulative cascade sizes in one minute intervals), burstiness (measuring the burst of popularity), stability (assessing the stability of Popularity), and release_time (hour of original tweet release), as described in Table I [7].

4) Tweet content: The initial tweets' content also plays a role in how the information spreads. To get feature vectors out of the tweet content, the dataset developers [18] used the word frequency-inverse document frequency algorithm. In addition, they get topic distributions using the latent Dirichlet allocation (LDA) topic model. Contentgory (themes category), wordlength (text length), URL presence, hashtag presence, photo presence, and mention presence are the six features retrieved from tweet content, as described in Table I [7].

TABLE I. ILLUSTRATES THE FOUR CATEGORIES OF FEATURES USED IN THE PREDICTION OF CASCADES

Features	Description	Representation
Following Network Structure (FNet)	The count of individuals who are following the user who made the post.	out_degree_F [7,17,18,19,,20].
	The count of individuals whom the posting user follows.	in_degree_ [7,17,18,19,,20].
	The count of individuals who either follow or are followed by the posting user.	all_degree_F [7].
Following Network Structure (FNet)	The count of individuals who both follow and are followed by the posting user.	bi_degree_F [7]
	PageRank centrality in G_F	pagerank_F [7]
Retweeting Network Structure (RNet)	The count of individuals who retweet the tweets of the posting user	out_degree_R [7]
	The count of individuals who have been retweeted by the posting user	in_degree_R [7]
	The count of individuals who either retweet or are retweeted by the posting user	all_degree_R
	The count of individuals who both retweet and are retweeted by the posting user	bi_degree_R
	PageRank centrality in G_R	pagerank_R [33]
Early Popularity (early)	the sizes of cascades during the observation period, divided into intervals of 1 minute	cascade
	The measure of how quickly and intensely popularity increases in the early period	burstiness
	The measure of how consistent and stable the popularity remains in the early period	stability
	The timestamp indicating when the original tweet was posted	Release_time

C. Data Pre-processing

A rule-based methodology has been employed to label the dataset to encode the variable 'all_degree2', which signifies the number of retweets and retweeted counts. This study aimed to minimize the level of variability in the chosen variable and enhance the ease of analyzing diffusion patterns. The encoding procedure entailed the categorization of values into three distinct groups: 'low diffusion' (set as '10'), 'medium diffusion' (set as '20'), and 'high diffusion' (set as '30'). The encoding scheme offers a distinct representation of the levels of diffusion linked to the number of retweets and retweeted counts.

Assigning discrete levels to the variable 'all_degree2' enhances comprehensibility and facilitates the analysis of diffusion patterns in the dataset. Fig. 3 represents the retweet counts in the dataset.

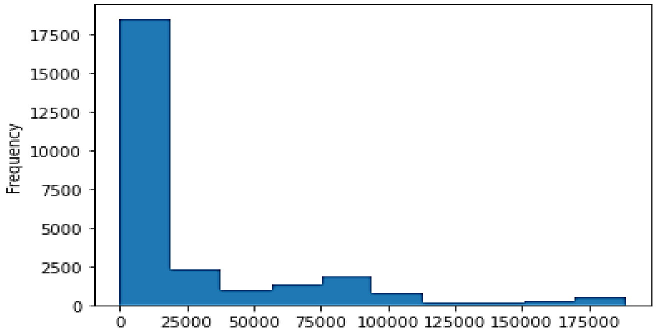


Fig. 3. Representing of the dataset retweet count analysis.

Fig. 4 provides a visual representation of the encoded values, thereby offering valuable insights into the distribution patterns of diffusion levels. This methodology streamlines the inspection of information propagation dynamics and deepens our comprehension of the dataset samples.

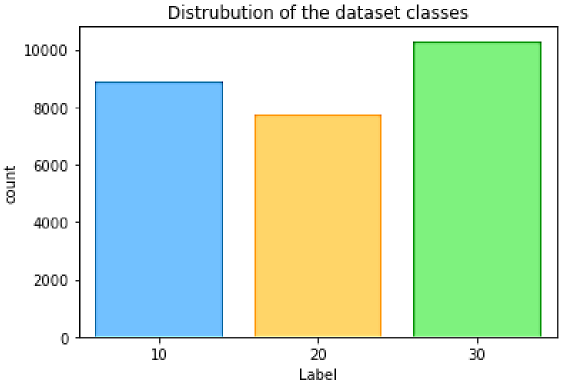


Fig. 4. Show the distribution of the dataset classes.

D. Data Normalization

In this study, the standard scalar is used on features for predicting information spread on social media. It makes the features have a similar scale values. This helps some machine learning algorithms that are sensitive to the scale of input data. This step makes all features contribute equally to the prediction task at hand. No feature dominates the learning process because of its original scale. The standard scaler calculates the mean and standard deviation of each feature. Then it transforms the values based on these statistics. The scaled features have a mean of zero and a standard deviation of one [21]. Using the standard scaler on this study’s features enhances the model's comparability and interpretability. This leads to more accurate predictions and analysis.

E. Dimensionality Reduction

The Principal Component Analysis (PCA) uses to remove unimportant elements from high-dimensional data to ease interpretation. Dimensionality reduction shortens data analysis and interpretation without sacrificing essential features. A PCA finds the dataset's most variable eigenvectors. PCA selects the most significant eigenvectors while retaining the most variation by reducing data dimensionality. The PCA projects data onto eigenvectors for visualization task. This helps identify data patterns and simplifies multidimensional data. It facilitates complex data, focuses on critical issues, and aids interpretation [22].

F. Data Split

In research, it is common practice to split the data into training and testing sets, usually using 80:20 ratios. The training set is used to train the model, while the testing set is used to evaluate the model's performance on test data. This split allows researchers to assess how well their models generalize to new, unseen data and make reliable results based on their research findings.

G. Prediction Models

This subsection details of machine learning and deep learning model architecture to analyze and model information diffusion patterns in social media networks.

1) Machine learning models: Machine Learning is a subfield within the realm of artificial intelligence that facilitates the acquisition of knowledge and improvement of systems through experience, thereby enhancing their performance without the need for explicit programming [23]. Regression is a widely employed machine learning methodology that aims to forecast continuous outcomes by utilizing statistical techniques to establish the association between independent and dependent variables.

a) Linear regression model: Linear Regression (LR) is a Statistical analysis method. This analysis considers several independent factors and one dependent variable and the relations between them [24]. The LR technique can help one understand how social media spreads information. In study of RL analysis for information dissemination, dependent variable is information dispersion or an analogous indicator like retweets or post spread. Independent variables include content features, network design, and other external factors that may affect content transmission. Fitting a linear regression model to the data estimates the coefficients for each independent variable. They control the variable. This model assumes a linear relationship between characteristics and diffusion strength. Linear regression helps us understand how features affect diffusion levels. Positive coefficients mean an increase in the feature leads to an increase in diffusion. Negative coefficients mean the opposite. The intercept shows the baseline diffusion when all features are zero, giving us insight into the content's inherent diffusion capacity.

b) Random forest regressor model: The Random Forest Regressor (RFR), a machine learning technique, is well-suited for modelling information diffusion. It predicts information dissemination by averaging the results of several decision trees. This prediction aggregation reduces variance, improving the model's test set generalization [25]. The RFR model can handle diverse, noisy data because it resists outliers. This is one of the RFR main advantages. It can also discover non-linear data linkages, which helps it produce accurate forecasts even in challenging settings. It distributes errors evenly across classes to handle unbalanced datasets. The RFR also helps assess the model's numerous variables' relative importance. It assigns significance scores to qualities, indicating their relevance to the prediction task. This knowledge can help researchers comprehend information spread factors.

2) Deep learning models: CNNs and RNNs can be used to model how information spreads through social networks and communication channels. CNNs can identify important elements or patterns that spread information by extracting relevant textual or visual input features. The ANN, BilSTM [26], and GRUs models can sequential data well for temporal information diffusion analysis. These architectures can identify critical diffusion nodes and anticipate future spread by capturing relationships and patterns in interaction or event sequences. Deep learning approaches can help to predict, analyse, and intervene in information diffusion by revealing its mechanisms and dynamics.

a) The ANN Model: In this study, ANN Regression has been proposed for information diffusion prediction in online social media. This sequential regression neural network model architecture consists of the input, output and hidden layers [27]. The input data passed to two hidden layers with 64 units with rectified linear unit (ReLU) activation function. Those layers can capture non-linear relationships and complex patterns present in the data. The model's output is a single unit for predicting the continuous variables related to information diffusion. The model Parameter describes in Table II.

TABLE II. THE ANN USED PARAMETERS

Parameter	Dilates
Input Shape	(4, 1)
Input Dimensions	Input shape
Hidden Layers	2
Output Units	1
Optimizer	Adam
Epochs	100

The mean squared error (MSE) has been used to measure the average squared difference between predicted and actual values. The Adam optimizer dynamically adjusts the learning rate based on the gradients' first and second moments to update the model's weights. MSE and MAE are used to evaluate model performance. These metrics reveal the model's accuracy and precision, helping researchers assess its ability to predict information diffusion patterns in online social media.

b) CNN-BiLSTM Model: This paper introduces a neural network model designed to predict information diffusion on online social media platforms. The model was constructed utilizing the Keras library and incorporates a range of components designed to effectively capture complex patterns and dynamics associated with information propagation. The architecture has proposed adheres to a sequential structure, wherein each component has been meticulously considered. The initial stage involves utilizing one-dimensional convolutional layer (Conv1D) comprising ten filters and a kernel size of 3. The function of this particular layer is to detect and analyze localized patterns and connections within the given input data. This process aids in identifying and extracting significant features associated with the diffusion of information. To avoid the overfitting issue, a dropout layer is incorporated into the model architecture, wherein a random selection of 20% of the input units are dropped or deactivated during the training process. The utilization of this regularization technique facilitates the process of generalization and diminishes dependence on particular features, thereby enhancing the resilience of diffusion predictions. Subsequently, a bidirectional Long Short-Term Memory (LSTM) layer [26], consisting of 10 units, is incorporated. Long Short-Term Memory (LSTM) networks demonstrate exceptional performance in modeling sequential data and effectively capturing temporal dependencies. The model's bidirectional characteristic enables it to consider both preceding and subsequent time steps, thereby facilitating a holistic comprehension of diffusion dynamics.

Two dense layers follow the LSTM layer. The first dense layer consists of 10 units with the rectified linear unit (ReLU) activation function, enabling the extraction of higher-level features. The second dense layer, with a single unit, generates the final diffusion predictions. To prepare the output for the dense layer, a flatten layer is added, reshaping the preceding layer's output into a one-dimensional vector. The model Parameter describes in the Table III. The evaluation metrics, such as MSE and mean absolute error (MAE), were used to assess the model's performance in accurately predicting the diffusion patterns.

TABLE III. SUMMARIZING OF THE CNN-BILSTM PARAMETERS

Parameter	Dilates
Input Shape	(4, 1)
Convolutional Layer	Filters: 10, Kernel Size: 3
Padding	Same
Activation Function	ReLU
Dropout Rate	0.2
Bidirectional LSTM Layer	Units: 10, Dropout: 0.3, Recurrent Dropout: 0.3
Dense Layer 1	Units: 10, Activation Function: ReLU
Flatten Layer	-
Dense Layer 2	Units: 1
Loss Function	Mean Squared Error (MSE)

c) CNN-GRU Model: In this study, the CNN-GRU deep learning model architecture uses to predict information diffusion in social media platforms. We define a sequential model to build linear information diffusion. The model consists of one-dimensional convolutional layer employs 10 filters and a kernel size of 3, while utilising the Rectified Linear Unit (ReLU) activation function. This layer is responsible for extracting pertinent features and patterns from the input data to comprehend the information diffusion processes. During training, the dropout layer is employed to randomly deactivate 20% of the input units to mitigate the overfitting risk. Regularization techniques enhance generalization capabilities of prediction models by reducing feature dependence, thereby increasing their robustness. A Gated Recurrent Unit (GRU) layer follows, consisting of 10 units. This recurrent layer captures temporal dependencies in sequential data, allowing the model to capture the dynamics of information diffusion over time. The model architecture also includes two dense layers. The first dense layer has 10 units with the ReLU activation function, extracting higher-level features. The second dense layer, with a single unit and no activation function, generates the final predictions. The evaluation metrics, such as MSE and mean absolute error (MAE), were used to assess the model's performance in accurately predicting the diffusion patterns. Table IV provides the summarization of the CNN-GRU used parameters.

TABLE IV. THE CNN-GRU MODEL PARAMETERS

Parameter	Dilates
Input shape	(4,1)
Conv1D	Filters: 10, Kernel Size: 3, Padding: 'same'
Activation	ReLU
Dropout	Rate: 0.2
GRU	Units: 10, Dropout: 0.3, Recurrent Dropout: 0.3
Dense	Units: 10, Activation: ReLU
Flatten	-
Dense	Units: 1
Optimizer	Adam

IV. EVALUATION METRICS

In this study, the researchers used various fundamental measurement metrics to evaluation the models' performance, including Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R2), and Root Mean Squared Error (RMSE). These metrics function as significant indicators of the accuracy and predictive capabilities of the models, allowing for a precise evaluation and comparison of their performance.

A. Mean Squared Error (MSE)

The Regression analysis uses MSE to evaluate prediction models. The MSE method calculates the averages of the squared deviations between predicted and actual values. Squaring large deviations makes the MSE more sensitive to outliers [28]. The RF model performs best with a low MSE. The following formula can calculate the MSE [28].

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1}$$

B. Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is a commonly employed evaluation metric especially for regression analysis. It calculates the square root of the mean of the squared differences between the predicted and actual values [31]. Root Mean Square Error (RMSE) is a statistical metric that quantifies the typical magnitude of errors and indicates the standard deviation of the residuals. This technique proves to be highly advantageous in scenarios where substantial errors notably influence the model's overall performance. Smaller root mean square error (RMSE) values indicate enhanced model accuracy and a stronger alignment with the observed data. The following formula can calculate the RMSE [31].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y_i} - y_i)^2} \tag{2}$$

C. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is an alternate evaluation metric frequently used in regression analysis. It calculates the average absolute discrepancy between forecasted and actual values in a dataset. This discrepancy evaluates the accuracy of forecasts. Unlike Mean Squared Error (MSE), which squares errors before averaging, MAE doesn't square errors [29]. This makes it less influenced by outliers or extreme values. It provides an indication of the average error magnitude in predictions. Smaller MAE values signify improved model performance, similar to MSE. A value of zero denotes perfect correspondence between forecasted and actual values. MAE is particularly helpful where the focus is on absolute error magnitude rather than squared discrepancies. This is because MAE considers absolute error magnitude. Eq. (3), calculates the MAE metric [29].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

D. R-squared (R2)

R-squared, or coefficient of determination, is a statistical measure assessing the goodness-of-fit of a regression model. It represents the proportion of variance in dependent variables explained by the model's independent variables [30]. The R-squared value ranges from 0 to 1. 1 indicates a perfect fit where the model presents all variability in data. 0 indicates model doesn't show any variability. R-squared can be interpreted as the percentage of variance in dependent variables accounted for by independent variables in the model. Eq. (4) calculates the R-squared [30].

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{4}$$

V. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the experimental results of our study, conducted on a laptop equipped with an 8th generation Intel Core i7 processor, 16GB of RAM, and an NVIDIA GeForce GTX GPU with 8GB. We evaluated the performance of our models using three metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared.

A. Models Performance Based on All Feature

The RF model has the best performance among these for modelling information diffusion on social media using All Features, as described in Table V. Based on R-squared value, a statistical measure representing a proportion of variance for dependent variables explained by independent variables in the regression model, the RandomForest model appears best performing. It has the highest R-squared value of 0.767, explaining approximately 76.7% of the variance in the dependent variable. Moreover, for MSE and MAE measures of prediction error, the RF model has the lowest scores on the test set. This signifies the smallest prediction errors among models. Therefore, based on provided metrics (R-squared, MSE, and

MAE), the RF model models is provided the best results among used models.

B. Models Performance Based on Early Retweet Dynamics Features

The ANN model has the best performance for modelling information diffusion on social media using early retweet dynamics features. Based on the R-squared value described in Table VI, the highest R-squared value was achieved by the ANN model, which indicates that it explains approximately 24.30% of the variance in the dependent variable. It is essential to consider that this value is relatively low, which indicates that there is only a limited amount of goodness of fit to the data. Every model could be improved by performing additional tuning or using additional or different features.

TABLE V. RESULTS OF INFORMATION DIFFUSION BASED ON ALL FEATURES

Model	MSE	RMSE	MAE	R-squared (%)
Linear Regression	46.364	6.811	5.864	33.995
CNN-BiLSTM	29.610	5.440	4.482	57.846
CNN-GRU	24.400	4.939	3.953	65.264
ANN	18.679	4.323	3.237	68.021
RF	16.374	4.048	2.599	76.690

TABLE VI. RESULTS OF INFORMATION DIFFUSION BASED ON EARLY RETWEET DYNAMICS FEATURES

Model	MSE	RMSE	MAE	R-squared (%)
Linear Regression	53.459	7.314	5.932	8.48
CNN-BiLSTM	43.708	6.609	5.740	18.32
Random Forest	46.608	6.826	5.729	20.21
CNN-GRU	44.626	6.678	5.810	23.60
ANN	44.212	6.645	5.776	24.309

C. Models Performance Based on Following Network Structure Features

The Random Forest model has the best performance among these for modelling information diffusion on social media using the Following network structure features. Based on R-squared value, Random Forest model is the best performing among these for models, as described in Table VII. It has highest R-squared value of 0.907, explaining approximately 90.7% of the variance in dependent variable. RandomForest model also has lowest MSE and MAE values on the test set, implying smallest prediction errors among models.

D. Models Performance Based on Retweeting Network Structure

Based on R-squared value, Linear Regression model is best performing among these for modelling information diffusion on social media using retweeting network structure features, as described in Table VIII. It has highest R-squared value of 0.982, explaining approximately 98.2% of the variance in dependent variable.

Although RF model has slightly higher MSE and MAE values on test set compared to Linear Regression, its R-squared value is also very close, making it a strong competitor.

E. Models Performance Based on Tweet Content Features

The ANN model achieved a higher level of effectiveness in representing the spread of information on social media platforms using tweet content features, as evidenced by the R-squared value. The R-squared of the ANN model is 0.055, suggesting that it explains approximately 5.5% of the variability observed in the dependent variable. Nevertheless, it is crucial to acknowledge that all models exhibit low or negative R-squared values that suggests inadequate alignment with the observed data. Table IX below summarizes the results using Tweet content features.

TABLE VII. RESULTS OF INFORMATION DIFFUSION BASED ON THE FOLLOWING NETWORK STRUCTURE FEATURES

Model	MSE	RMSE	MAE	R-squared (%)
Linear Regression	46.669	6.836	5.155	20.1
CNN-GRU	25.957	5.095	4.086	55.6
CNN-BiLSTM	22.762	4.767	3.970	57.5
ANN	20.031	4.475	3.549	65.7
RF	5.389	2.322	0.999	90.8

TABLE VIII. THE RESULTS OF INFORMATION DIFFUSION BASED ON RETWEETING NETWORK STRUCTURE

Model	MSE	RMSE	MAE	R-squared (%)
CNN-BiLSTM	5.846	2.416	1.392	80.1
CNN-GRU	11.601	3.405	2.280	80.1
ANN	5.253	2.293	1.486	91.0
Linear Regression	1.097	1.048	0.237	98.1
Random Forest	1.074	1.036	0.233	98.2

TABLE IX. THE RESULTS OF INFORMATION DIFFUSION BASED ON TWEET CONTENT FEATURES

Model	MSE	RMSE	MAE	R-squared (%)
Linear Regression	68.593	8.289	6.708	-0.174
Random Forest	68.548	8.284	6.707	-0.174
CNN-BiLSTM	56.095	7.491	6.289	0.040
CNN-GRU	55.754	7.468	6.329	0.045
ANN	54.834	7.404	6.257	6.124

V. DISCUSSION

Our findings have significant implications for predicting the spread of information through social media. The models were evaluated based on their ability to capture the intricacies of the information diffusion process. Their performance was measured using various feature sets such as all features, early retweet dynamics, Following network structure, Retweeting network structure and Tweet content. The RF model performance is higher than other models. The model's MSE was 16.374, RMSE was 4.049, MAE was 2.599, and R-squared was 76.690. This result indicates that more accurate predictions of the extent of information dispersion can be achieved by including a wide variety of features. The RF model reduces variation and improves generalization performance by averaging predictions from multiple decision trees.

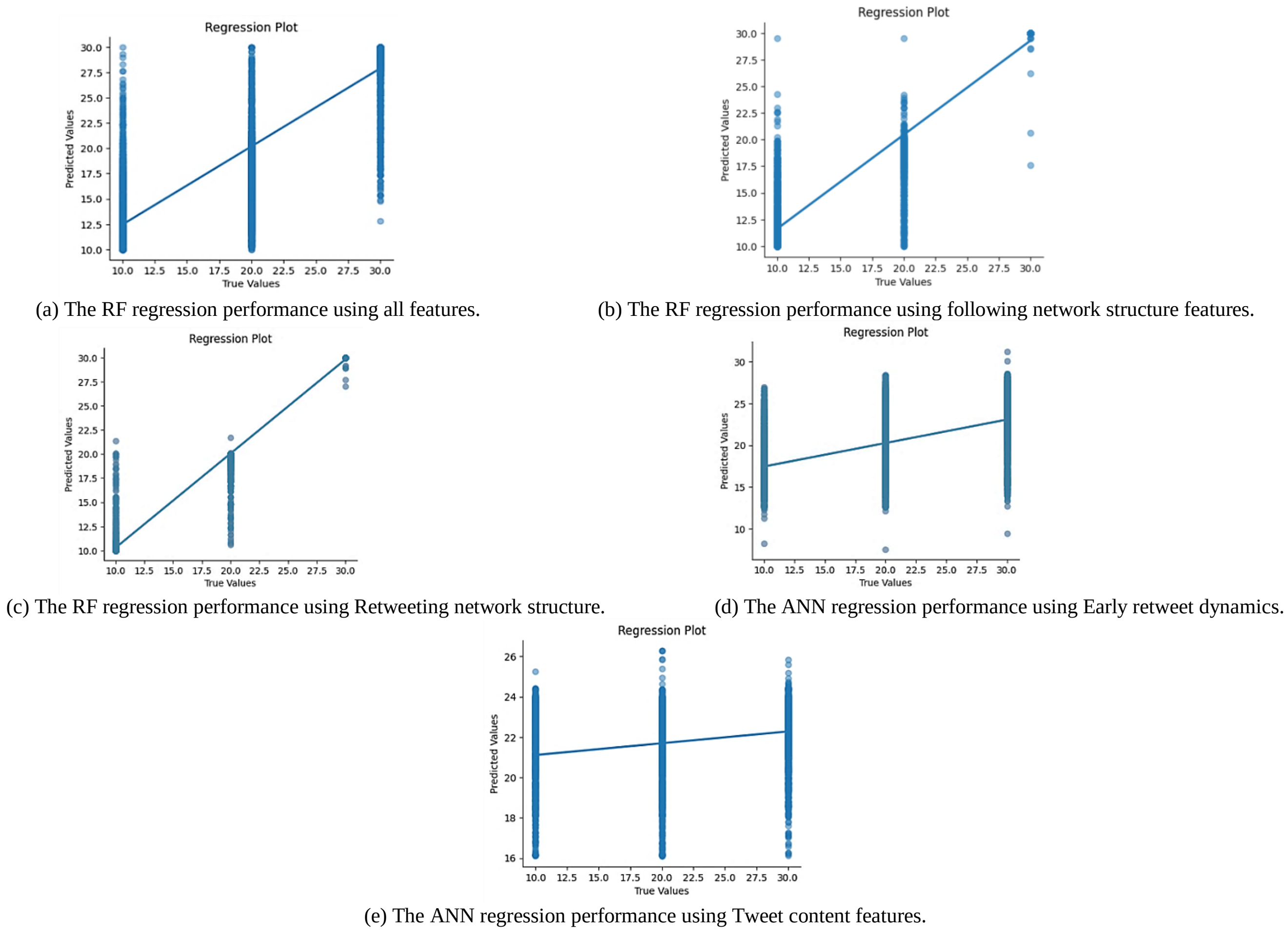
We also explored how tweet content features play a role in prediction. The ANN model was run on these features, and the results showed that the MSE, RMSE, MAE, and R^2 values were as follows: 44.212, 6.645, 5.776, and 24.309, respectively. The model performance was not good when we concentrated most of our attention on early retweet dynamics. Consequently, it would appears that the dynamics of early-stage popularity may not be adequate to accurately predict the extent of information cascades, even though they do play a part in the process of information dissemination.

The impact of network structural features on prediction performance was then examined. The MSE, RMSE, MAE, and R-squared for a RF model trained on the following network architecture were 5.389, 2.321, 0.999, and 90.773, respectively. This shows how the extent of information cascades heavily depends on the level of user influence in the subsequent network. Similar success was seen with a RF model that was trained using variables extracted from the retweeting network; this model achieved an excellent MSE of 1.074, RMSE of 1.036, MAE of 0.233, and an astounding R-squared of 98.161. These findings underline the significance of taking into account the topology of the retweeting network when estimating the rate of information diffusion in a Twitter stream.

Finally, we investigated the significance of tweet content features in the context of prediction. The ANN model, which was exclusively trained using the content of tweets, demonstrated performance metrics including a mean squared error (MSE) of 54.834 , root mean squared error (RMSE) of 7.404, mean absolute error (MAE) of 6.257, and a relatively low R-squared value of 6.124. This implies that the sole consideration of tweet content may not be adequate in accurately predicting the sizes of cascades. Including network structure and the dynamics of early-stage popularity appears essential to understand the intricacies of information diffusion comprehensively.

This study emphasizes the significance of considering various factors in predicting the scale of information diffusion on social media. The optimal predictive performance is achieved by considering a combination of network structure features, early-stage popularity dynamics, and tweet content. The RF models' efficacy in capturing network structure's impact has been demonstrated. In contrast, the ANN model offers valuable insights into the significance of early-stage popularity dynamics.

Subsequent investigations may priorities examining supplementary attributes and integrating more sophisticated machine learning and deep learning methodologies to enhance the precision of predicting information dissemination on social media platforms. Fig. 5 shows the regression plots for the best models performance for information diffusion.



(a) The RF regression performance using all features.

(b) The RF regression performance using following network structure features.

(c) The RF regression performance using Retweeting network structure.

(d) The ANN regression performance using Early retweet dynamics.

(e) The ANN regression performance using Tweet content features.

Fig. 5. Shows the combination of various regression models performance plots for predicting information diffusion.

Table X summaries the best results obtained from our experimental work carried out in this study.

TABLE X. PRESENTING THE BEST RESULTS OF OVERALL

Features Used	**Best Model**	**MSE**	**RMSE**	**MAE**	**R-squared (%)**
All features	Random Forest	16.374	4.049	2.599	76.690
Early retweet dynamics	ANN	44.212	6.645	5.776	24.309
Following network structure	Random Forest	5.389	2.321	0.999	90.8
Retweeting network structure	Random Forest	1.074	1.036	0.233	98.2
Tweet content	ANN	54.834	7.404	6.257	6.124

As shown in above cited X table, the best results gained for information diffusion on social media by the ANN and RF models. The results emphases on important findings with various characteristics that perform differently in terms of models, and some features have better predictive ability and accuracy when it comes to comprehending the dynamics of information dissemination on social media platforms. The R-squared values provide information about how well the models represent the diffusion process's variability. These results advance our knowledge of information distribution inside social media networks in a more complex way.

Fig. 6 presents the visualization of the best results for the proposed models.

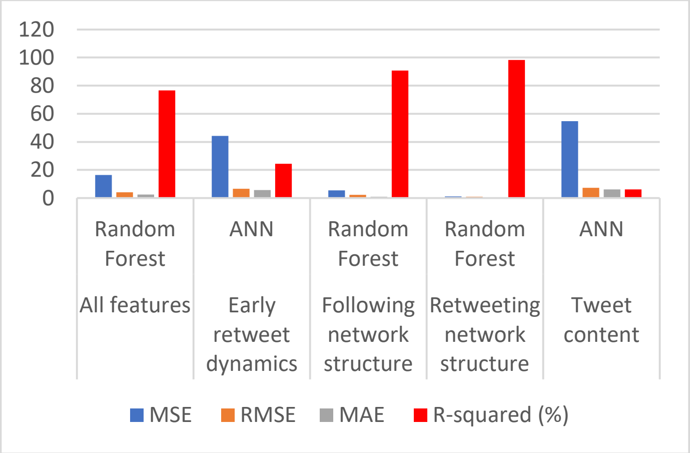


Fig. 6. Graphical representation shows the best models performance and results.

In our research study, we compared how well our models predicts the spread of information based on retweets and retweet counts versus another study that used XGBoost modeling. We aim methodically to evaluate our model's performance with exist ones. We observed the root mean squared error (RMSE) - a standard way of measuring how closely regression models predict values. We investigated the RMSE for different features, like, all features, early retweet dynamics, Following network, Retweeting network, Tweet content. Authors in study [7] used XGBoost model, which had RMSE ranging from 134 to 237 for the different features. Our models using RF and neural networks performed much better, with RMSE from 1.036 to 7.404 as described in Table XI. These results demonstrate our models predict information diffusion from retweets and retweet counts far more accurately than the XGBoost models in the other study. Our models matched the real dynamics of information spreading much closer, as shown by the lower RMSE values.

TABLE XI. A COMPARISON BETWEEN THE RESULTS OF OUR APPROACH WITH EXISTING ONES

Features	**Result from Study [7] (XGBoost)**	**Our Study (RF / ANN)**
All features	196	4.049
Early retweet dynamics	237	6.645
Following network	151	2.321
Retweeting network	134	1.036
Tweet content	205	7.404

VI. CONCLUSION

This study predicted social media information diffusion based on the number of retweets and retweets counts. The deep learning and machine learning models were trained on Weibo, a social network platform similar to Twitter. Those models used various features such as all features, early retweet dynamics, following network structure, Retweeting network structure, and Tweet content. The best result was achieved by RF 98.161 using Retweeting network structure features. The results of our study provide insights into mechanisms of information dissemination on social media platforms and provide guidance for predicting potential influence of forthcoming news events. Optimal predictive performance is attained by considering interplay of network structure, early popularity dynamics, and features inherent to tweet content .Accurate information diffusion prediction helps businesses improve their marketing strategies, public health officials respond to disease outbreaks, and identify social network influencers. The present study has potential to establish a fundamental basis for future investigations integrating cutting-edge machine learning and deep learning techniques to explore supplementary variables and other social media features that can enhance precision of predictions of information diffusion on online platforms. Enhancements in capacity to forecast extent of information dissemination hold significant potential for various domains, including advertising, crisis management, and examination of online social dynamics.

REFERENCES

[1] Nam, Y. W., Son, I. S., & Lee, D. W. “The impact of message characteristics on online viral diffusion in online social media services: The case of Twitter”. Journal of Intelligence and Information Systems, 17(4), 2011, pp. 75-94.

[2] Kumar, N. “Information Diffusion and Summarization in Social Networks” (Doctoral dissertation, IIT Hyderabad, India), 2019, pp. 2019.

[3] Hatua, A., Nguyen, T. T., & Sung, A. H. “Information diffusion on Twitter: Pattern recognition and prediction of volume, sentiment, and influence”. In Proceedings of the 4th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), 2017, pp. 157-167. doi: 10.1145/3148055.3148078.

[4] Al-Taie, M. Z., & Kadry, S. “Information Diffusion in Social Networks”. In Python for Graph and Network Analysis, 2017, pp. 165. doi: 10.1007/978-3-319-53004-8_8.

[5] Masud, S., et al. “Hate is the New Infodemic: A Topic-aware Modeling of Hate Speech Diffusion on Twitter”. Proceedings of the International Conference on Data Engineering (ICDE), 2020, pp. 504-515. doi: 10.1109/ICDE51399.2021.00050.

[6] Kushwaha, A. K., Kar, A. K., & Ilavarasan, P. V. “Predicting Information Diffusion on Twitter: A Deep Learning Neural Network Model Using Custom Weighted Word Features”. Lecture Notes in Computer Science (LNCS), 2020, pp. 456-468. doi: 10.1007/978-3-030-44999-5_38.

[7] Cao, R. M., Liu, X. F., & Xu, X. K. “Why cannot long-term cascade be predicted? Exploring temporal dynamics in information diffusion processes”. R Soc Open Sci, 8(9), 2021, pp. 2021. doi: 10.1098/RSOS.202245.

[8] Razaque, A., Rizvi, S., Khan, M. J., Almiani, M., & Al Rahayfeh, A. “State-of-art review of information diffusion models and their impact on social network vulnerabilities”. Journal of King Saud University - Computer and Information Sciences, 34(1), 2022, pp. 1275-1294. doi: 10.1016/J.JKSUCI.2019.08.008.

[9] Vargo, C. J. “Brand messages on Twitter: Predicting diffusion with textual characteristics”, (Doctoral dissertation, The University of North Carolina at Chapel Hill), 2014, pp. 2014.

[10] Lytvyniuk, K., Sharma, R., & Jurek-Loughrey, A. “Predicting Information Diffusion in Online Social Platforms: A Twitter Case Study”. Studies in Computational Intelligence, 812, 2018, pp. 405-417. doi: 10.1007/978-3-030-05411-3.

[11] Kafeza, E., Kanavos, A., Makris, C., & Vikatos, P. “Predicting information diffusion patterns in twitter”. IFIP Adv Inf Commun Technol, 436, 2014, pp. 79-89. doi: 10.1007/978-3-662-44654-6_8/COVER.

[12] Wang, L., Niu, J., & Yu, S. “SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis”. IEEE Trans Knowl Data Eng, 32(10), 2020, pp. 2026-2039. doi: 10.1109/TKDE.2019.2913641.

[13] Wang, J., Zheng, V. W., Liu, Z., & Chang, K. C. C. “Topological Recurrent Neural Network for Diffusion Prediction”. Proceedings of the IEEE International Conference on Data Mining (ICDM), 2017, pp. 475-484. doi: 10.1109/ICDM.2017.57.

[14] Sankar, A., Zhang, X., Krishnan, A., & Han, J. “Inf-VAE: A Variational Autoencoder Framework to Integrate Homophily and Influence in Diffusion Prediction”. doi: 10.1145/3336191.

[15] Zhang, Z., Zhao, W., Yang, J., Paris, C., & Nepal, S. “Learning influence probabilities and modelling influence diffusion in Twitter”. The Web Conference 2019 - Companion of the World Wide Web Conference (WWW), 2019, pp. 1087-1094. doi: 10.1145/3308560.3316701.

[16] Kushwaha, A. K., Kar, A. K., & Ilavarasan, P. V. “Predicting Information Diffusion on Twitter: A Deep Learning Neural Network Model Using Custom Weighted Word Features”. 2020, pp. 2020. doi: 10.1007/978-3-030-44999-5_38.

[17] Bich, T., Hoang, N., & Mothe, J. “Predicting information diffusion on Twitter – Analysis of predictive features”. J Comput Sci, 2017, pp. 2017. doi: 10.1016/j.jocs.2017.10.010.

[18] Suh, B., Hong, L., Pirolli, P., & Chi, E. H. “Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network”. Proceedings - SocialCom 2010: 2nd IEEE International Conference on Social Computing, PASSAT 2010: 2nd IEEE International Conference on Privacy, Security, Risk and Trust, 2010, pp. 177-184. doi: 10.1109/SOCIALCOM.2010.33.

[19] Petrovic, S., Osborne, M., & Lavrenko, V. “RT to Win! Predicting Message Propagation in Twitter”. Proceedings of the International AAAI Conference on Web and Social Media, 5(1), 2011, pp. 586-589. doi: 10.1609/ICWSM.V5I1.14149.

[20] Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., & Kustarev, A. “Prediction of retweet cascade size over time”. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), 2012, pp. 2335-2338.

[21] Quackenbush, J. “Microarray data normalization and transformation”. Nature genetics, 32(4), 2002, pp. 496-501. doi: 10.1038/NG1032.

[22] Bro, R., & Smilde, A. K. “Principal component analysis”. Analytical Methods, 6(9), 2014, pp. 2812-2831. doi: 10.1039/C3AY41907J.

[23] Mahesh, B. “Machine learning algorithms-a review”. International Journal of Science and Research (IJSR), 9(1), 2020, pp. 381-386.

[24] Aalen, O. O. “A linear regression model for the analysis of life times”. Stat Med, 8(8), 1989, pp. 907-925. doi: 10.1002/SIM.4780080803.

[25] Grömping, U. “Variable Importance Assessment in Regression: Linear Regression versus Random Forest”. Vol. 63, no. 4, 2012, pp. 308-319. doi: 10.1198/TAST.2009.08199.

[26] Rhanoui, M., Mikram, M., Yousfi, S., & Barzali, S. “A CNN-BiLSTM Model for Document-Level Sentiment Analysis”. Machine Learning and Knowledge Extraction 2019, 1(3), 2019, pp. 832-847. doi: 10.3390/MAKE1030048.

[27] Jain, A. K., Mao, J., & Mohiuddin, K. M. “Artificial neural networks: A tutorial”. Computer (Long Beach Calif), 29(3), 1996, pp. 31-44. doi: 10.1109/2.485891.

[28] Das, K., Jiang, J., & Rao, J. N. K. “Mean squared error of empirical predictor”. Vol. 32, no. 2, 2004, pp. 818-840. doi: 10.1214/009053604000000201.

[29] Chai, T., & Draxler, R. R. “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature”. Geoscientific model development, 7(3), 2014, pp. 1247-1250.

[30] Cameron, A. C., & Windmeijer, F. A. G. “R-Squared Measures for Count Data Regression Models With Applications to Health-Care Utilization”. Vol. 14, no. 2, 2012,pp.209-220.doi: 10.1080/07350015.1996.10524648.

[31] Willmott, C. J., & Matsuura, K. “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”. Clim Res, 30(1), 2005, pp. 79-82. doi: 10.3354/CR030079.

[32] Ma, C., Yan, Z., & Chen, C. W. “LARM: A Lifetime Aware Regression Model for Predicting YouTube Video Popularity”. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM), 2017, pp. 467-476.

[33] Hong, L., Dan, O., & Davison, B. D. “Predicting popular messages in Twitter”. Proceedings of the 20th Int. Conf. Companion on World Wide Web - WWW '11, New York, NY: ACM Press, 2011. doi: 10.1145/1963192.1963222.

[34] Yan, Z., Zhou, X., Ren, J., Zhang, Q., & Du, R. (2023). Identifying underlying influential factors in information diffusion process on social media platform: A hybrid approach of data mining and time series regression. Information Processing & Management, 60(5), 103438.