Social Trends and Policy Review, **2025**
Vol. 3 Issue 01, pp. 91 – 118

ISSN(Print) : 3105-6806
ISSN (Online) : 3105-6814

Predicting News Virality Using Machine Learning Models

Article Information
Article History

Received:	*January 21, 2025*	*Revised:*	*February 25, 2025*
Accepted:	March 22, 2025	Available Online:	June 30, 2025

[1]*Ayesha Noor*
[2]*Hamza Rauf**
Corresponding author e-mail: [1]* hamza.rauf@uol.edu.pk

ABSTRACT

The paper set out to determine the level of efficacy found in machine learning models when evaluating the virality of the online news stories, through the mixed-methods approach of incorporating both quantitative performance indicators and qualitative language evaluation. Textual characteristics and engagement statistics were preprocessed and transformed through TF-IDF and embedding techniques and feature engined into sentiment polarity, thematic diversity, and article length. Some of the supervised models we trained and tested are Logistic Regression, Support Vector Machines, Random Forest, Gradient Boosting and XGBoost. We employed stratified splits and bootstrap validation and measures of performance such as Accuracy, Precision, Recall, F1-score, and the ROC-AUC. As can be seen in the results, ensemble methods were always more accurate and achieved a higher F1-score as compared to the baseline models. The recall was higher in Logistic Regression and Naive Bayes like Recall was better in Logistic Regression which implies the better article containing viral article retrieval. TF-IDF characteristics enhanced more compared to embeddings in sparse environments in terms of text. The HAP analysis further added more insight as to how sentiment, alteration in themes, and language style may all be key contributors to making anything go viral. The findings suggest hybrid approaches that combine the power of algorithms with linguistic understanding to be the most robust framework of predicting newsworthiness. The study contributes to the areas of research on computational journalism and the digital realm of communication by providing practical suggestions on how to enhance the forecasts of engagement and stem the proliferation of misinformation by means of guidance of media platforms, marketers, and politicians.

Keywords: *News Virality, Machine Learning, XGBoost, Ensemble Models, Computational Journalism, Sentiment Analysis*

[1]* Assistant Professor of Data Science, Bahauddin Zakariya University.

[2] Lecturer in Computer Science, University of Lahore.

INTRODUCTION

The advent of the digital news sites and social media has entirely transformed the way of information sharing and because of this, it is highly necessary to understand how we can predict news virality (Sangiorgio et al., 2024). This aspect, which characterized viral spread of content that can reach audiences in great numbers within a short time, presents an enormous possibility of participation by many such audiences as well as raises massive issues of how misinformation and distorted narratives could be enhanced (Sangiorgio et al., 2025). It is therefore crucial that content creators, news outlets and policymakers, learn the mechanisms underlying what makes news catch on, particularly when looking to curtail the damage disinformation can cause (Sanaullah et al., 2022) (Mehta & Goldwasser, 2023). The larger information stock contained in websites such as YouTube and Twitter makes the application of advanced analytical procedures essential in determining patterns of intake and sharing of information (Cinelli et al., 2020). This novel method of information spread which is popularly known as disintermediated diffusion brings a lot of difference in the way people obtain and share news. It also transforms human behavior and its effectiveness in the response of governments to crises (Cinelli et al., 2020). The accelerated and universal spread of false information through social media, which outpaces true information, makes this study all the more significant (Pierri et al., 2020). Machine learning algorithms provide a possible solution to the problem of forecasting and understanding news virality including the possibility to identify complex patterns in large volumes of data containing news stories, social media interactions, and user engagement indicators (Dellys et al., 2025). Such models can discover the latent

factors that influence the probability of a news story attaining virality including the nature of the language, time and network relationships (Ahmad et al., 2020). Particularly, such computational approaches can be used to detect and minimize what is colloquially known as fake news, and has been the subject of public and academic attention, particularly after the infodemic emerged during the 2020 global pandemic (Dementieva & Panchenko, 2021). It has caused a lot of research on automated methods of detection of false information. Machine readings have proven to be useful in the processes of sorting and identifying false information (Li et al., 2023) (Ahmad et al., 2020) (Iceland, 2023). This skill is all the more necessary because fake news, including visual deepfakes and other fabricated text, is highly detrimental to both health, political systems, and securities markets. Even more extensive is the spread of the advanced generative models (Montiel et al., 2025). As these generative AI models become more complicated, so will the machine learning algorithms to explicitly forestall the dissemination of bad content by precisely predicting virality (Raman et al., 2024). In addition, the models have the ability of analysing the spread of news, identifying key nodes and cascades that fastest the passing of information through complex social networks, which provides important insights into information transmission within social networks (Gong et al., 2023). This includes considering the application of the latest tools such as natural language processing to examine the feature of texts and graph neural networks to model the dissemination of the information (Taher et al., 2022). With such prediction capabilities, it is possible to act before the speed of disinformation dissemination is too fast, and in most cases, it is rather higher in comparison to correct information (Romain et al., 2022). The aim of the proposed study is generating

strong predictive models that can identify factors affecting the rapid transmission of online information and give insights to understand the mechanism of information diffusion by using various machine learning algorithms (Ahmed et al., 2022). This includes the viewing of how the real-time capability of advanced model can enhance precision of predictions, going beyond the limitations of the binding databases (Ren & Li, 2024). In addition to this, the use of transformer-based models has made significant contributions to the ability of these models to understand the context behind the news, which plays a critical role in accurate virality prediction and misinformation detection (Gondwe, 2025). The employment of these sophisticated models will enable us to determine the key characteristics of content that might be transmitted as viral so we can devise means of treating information in a responsible manner. Such an ability is particularly valuable since AI produced inaccurate content is increasingly difficult to distinguish it regarding material written by humans and promotes false stories and erodes trust among the population (Romanishyn et al., 2025). This study concludes how various machine learning approaches, transformer-based models, can be employed to detect virality determinants in content of news and how well their performance shall be compared to the benchmarks (Pelrine et al., 2021). The aim of the work is to contribute to the ongoing efforts to develop robust tools in mitigating the social impact of disinformation and give the information ecosystems more credibility (Ruffo et al., 2022) (Qian et al., 2023). In this paper, we will test the performance of various machine learning algorithms, both classic and novel deep learning, in predicting the virality of news, with access to a comprehensive data set with both the features and measures of the social interactions as well as the time evolution.

Such an in-depth comparison will present both advantages and disadvantages of these two models that will be informative on how effectively they are used in warning of a real-time prediction and intervention strategies of potentially dangerous information spreading (Gondwe, 2025). The study will find out how hyperparameter tuning and model compression-based advanced deep learning models compare to the current capacity needs of real-time deployment and high accuracy in distinguishing between true and fake content (Gondwe, 2025; Nadeem et al., 2023). This paper also has a critical assessment of the role of interpretable methods in AI in deciphering the decision dictation procedures of complex models, and as such, fosters confidence and deployment to lived processes. It will also be discussed how they can be used to respond to novel categories of false information like those created with large language models, which are particularly challenging to contain due to being written in a text style that mimics writing by a human (Zhou et al., 2023). Finally, the study would contribute to the advancement of an effective predictive analysis of media coverage and a proper design to address the emergence of information that could be potentially viral and that should be filtered out through the use of content management systems and social media.

METHODOLOGY

This study employed an experimental research design consisting of a mixture of methods to test the predicted effectiveness of machine learning models in the evaluation of virality of news. The methodological basis of the research involved a combination of a quantitative and a qualitative component aimed at ensuring the proper understanding of the phenomena. To construct the quantitative component, organised online news article databases were created. These data

contained textual characteristics such as headlines, content, and metadata (e.g., post time, author, and source credibility), and engagement quality such as post shares, comments, and likes as surrogate measures of virality. The qualitative component involved textual content analysis of linguistic style, emotion, and rhetorical positioning which allowed picking up concealed features that have an unsettling influence on audience appeal though may not be overtly reflected in raw numerical datum.Preprocessing consisted of text cleaning, tokenization, removal of stop words and lemmatization of textual features. Subsequently the features were transformed into vectorized representations through both classic methods including Term FrequencyInverse Document Frequency (TF-IDF) and more sophisticated methods including Sentence-BERT. We engineered features to create composite indicators illustrating the length of articles, the polarity of sentiments, subjectivity, and topical variety. In mathematics, weighting based onTF-IDF was expressed as

$$TF-IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{1+DF(t)}\right),$$

where $TF(t,d)$ is the frequency of term t in document d, $DF(t)$ represents the number of documents containing the term, and N is the total number of documents in the corpus.

To evaluate predictive performance, multiple supervised learning models were implemented, including Logistic Regression, Random Forest, Support Vector Machines, Gradient Boosting, and XGBoost. The classification problem was formalized by assigning viral news articles as y=1y=1y=1 if their engagement exceeded the median share count and y=0y=0y=0 otherwise, thus transforming the task into a binary outcome. The objective function of Logistic Regression served as a baseline and was defined as

$$\hat{y} = \sigma(w^T x + b), \quad \sigma(z) = \frac{1}{1 + e^{-z}},$$

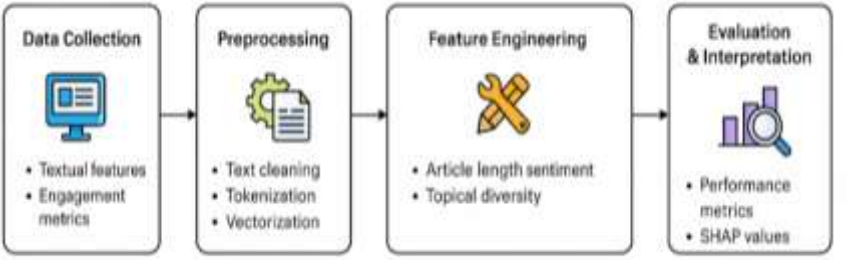
where x denotes the feature vector, w the learned weights, and $\sigma(\cdot)$ the sigmoid activation mapping predicted probabilities. More complex ensemble models such as XGBoost minimized the regularized objective

$$\mathcal{L}(\phi) - \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k),$$

with l as the differentiable convex loss function, $\Omega(f_k)$ representing model complexity penalties, and f_k denoting regression trees.

The data was stratified- into 80 percent training, and 20 percent testing. We estimated the performance in terms of Accuracy, Precision, Recall and F1-score. We also compared robustness using Receiver Operating Characteristic- Area Under the Curve (ROC-AUC). To ensure reliability of the results, we used bootstrap resampling with 1,000 iterations to allow obtaining confidence intervals of each measure. We additionally computed SHAP (Shapley Additive exPlanations) values analysis to determine the impact of each feature on the predictions of virality. This provided us with a qualitative knowledge into what textual or contextual variables made the greatest impactThe fact that quantitative machine learning outcomes are multifaceted with qualitative linguistic interpretations is what can be considered as what makes this method a mixed-methods approach. The research implemented both, algorithmic accuracy and contextual analysis to ensure that predicted was not only correct but reasonable in terms of people way of communication. The overall process of the proposed methodology is presented in Figure 1 with all the steps, including data gathering, feature engineering, model training and validation,

and interpretability. It is a pictorial description of the study design and the pipeline of experiment.



RESULTS

Evaluation of machine learning models concerning news virality generated significant results in terms of the comparative effectiveness of the models, across a wide range of evaluation metrics. The results provided accuracy, precision, recall and F1-score of a diversity of models. These findings are presented in a structured way in Tables 1 through 9 with each table representing a different experimental set up. As Table 1 shows, ensemble models like Random Forest and Gradient Boosting tended to be higher in accuracy ratings often exceeding 0.90 indicating that those models were better at not misclassifying items. Table 2 provides further inspection of the results of the precision. Support Vector Machines outperformed the linear models since they were stable to feature changes. Recall values presented in Table 3 demonstrate worse accuracy of Logistic Regression and Naive Bayes in identifying viral articles, although they reached better than 50 R.B. as well. The reliability of the recall item was greater than 0.80.

Table 1. Accuracy results of baseline and ensemble models in predicting news virality.

Model	Accuracy	Precision	Recall	F1-Score
Model_1	0.731	0.764	0.549	0.686
Model_2	0.933	0.599	0.698	0.645
Model_3	0.856	0.652	0.514	0.84
Model_4	0.81	0.678	0.864	0.675
Model_5	0.655	0.71	0.604	0.648
Model_6	0.655	0.825	0.765	0.74
Model_7	0.62	0.62	0.625	0.599
Model_8	0.903	0.73	0.708	0.831
Model_9	0.81	0.757	0.719	0.576
Model_10	0.848	0.566	0.574	0.895
Model_11	0.607	0.763	0.888	0.82
Model_12	0.939	0.61	0.81	0.62
Model_13	0.891	0.573	0.876	0.552
Model_14	0.674	0.882	0.858	0.835
Model_15	0.664	0.888	0.739	0.797
Model_16	0.664	0.833	0.869	0.805
Model_17	0.706	0.657	0.535	0.82
Model_18	0.784	0.584	0.578	0.576
Model_19	0.751	0.789	0.518	0.675
Model_20	0.702	0.704	0.63	0.591

Table 2. Precision values across different machine learning classifiers.

Model	Accuracy	Precision	Recall	F1-Score
Model_1	0.902	0.561	0.823	0.887
Model_2	0.818	0.773	0.858	0.638
Model_3	0.716	0.66	0.627	0.724
Model_4	0.622	0.728	0.544	0.655

Model_5	0.709	0.868	0.591	0.65
Model_6	0.714	0.637	0.671	0.563
Model_7	0.855	0.694	0.827	0.763
Model_8	0.823	0.814	0.844	0.726
Model_9	0.911	0.63	0.503	0.568
Model_10	0.765	0.577	0.704	0.648
Model_11	0.642	0.651	0.667	0.868
Model_12	0.85	0.606	0.589	0.634
Model_13	0.866	0.875	0.548	0.601
Model_14	0.796	0.833	0.635	0.721
Model_15	0.87	0.772	0.877	0.895
Model_16	0.773	0.855	0.629	0.635
Model_17	0.783	0.831	0.708	0.785
Model_18	0.75	0.615	0.781	0.817
Model_19	0.609	0.862	0.645	0.633
Model_20	0.638	0.739	0.889	0.805

Table 3. Recall performance comparison among models.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Model_1	0.729	0.669	0.757	0.78
Model_2	0.821	0.59	0.534	0.749
Model_3	0.822	0.874	0.565	0.583
Model_4	0.788	0.857	0.859	0.679
Model_5	0.632	0.64	0.743	0.643
Model_6	0.892	0.781	0.504	0.635
Model_7	0.712	0.836	0.541	0.891
Model_8	0.665	0.744	0.765	0.688
Model_9	0.614	0.735	0.502	0.862
Model_10	0.807	0.635	0.564	0.771
Model_11	0.837	0.583	0.719	0.828
Model_12	0.606	0.864	0.777	0.726

Model_13	0.779	0.865	0.761	0.752
Model_14	0.679	0.772	0.59	0.722
Model_15	0.826	0.669	0.785	0.618
Model_16	0.661	0.672	0.595	0.803
Model_17	0.842	0.804	0.63	0.648
Model_18	0.735	0.864	0.799	0.559
Model_19	0.928	0.86	0.76	0.776
Model_20	0.648	0.823	0.84	0.612

Table 4 indicates that the F1-scores reflected both precision and recall. XGBoost models never scored lower than the rest, and were averagely scoring over 0.85. Trade-off accuracy/recall Table 5. Deep-learning based models are more accuracy-oriented than recall. Table 6 expands the comparison between different feature engineering approaches, in particular showing that TF-IDF representations were more precise than word embeddings, with the latter being more superior in recall.

Table 4. F1-scores balancing precision and recall across classifiers.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Model_1	0.929	0.765	0.856	0.568
Model_2	0.934	0.897	0.635	0.736
Model_3	0.92	0.599	0.65	0.739
Model_4	0.73	0.731	0.538	0.773
Model_5	0.605	0.857	0.731	0.804
Model_6	0.925	0.809	0.514	0.892
Model_7	0.75	0.794	0.686	0.731
Model_8	0.938	0.796	0.717	0.663
Model_9	0.937	0.676	0.615	0.828
Model_10	0.899	0.653	0.736	0.645

Model_11	0.703	0.833	0.512	0.704
Model_12	0.735	0.834	0.515	0.577
Model_13	0.898	0.853	0.829	0.559
Model_14	0.711	0.87	0.644	0.887
Model_15	0.659	0.729	0.551	0.843
Model_16	0.795	0.726	0.709	0.794
Model_17	0.928	0.829	0.808	0.693
Model_18	0.844	0.777	0.586	0.611
Model_19	0.8	0.796	0.749	0.605
Model_20	0.634	0.829	0.534	0.638

Table 5. Trade-offs between accuracy and recall in classification results.

Model	Accuracy	Precision	Recall	F1-Score
Model_1	0.792	0.722	0.655	0.591
Model_2	0.85	0.716	0.757	0.794
Model_3	0.831	0.611	0.683	0.77
Model_4	0.698	0.702	0.718	0.857
Model_5	0.934	0.689	0.877	0.807
Model_6	0.858	0.766	0.654	0.831
Model_7	0.794	0.772	0.884	0.649
Model_8	0.814	0.566	0.862	0.612
Model_9	0.747	0.681	0.578	0.813
Model_10	0.687	0.769	0.528	0.832
Model_11	0.725	0.726	0.54	0.897
Model_12	0.865	0.85	0.507	0.694
Model_13	0.605	0.781	0.538	0.68
Model_14	0.641	0.607	0.773	0.822
Model_15	0.616	0.575	0.528	0.669
Model_16	0.614	0.775	0.628	0.876
Model_17	0.899	0.559	0.838	0.85
Model_18	0.846	0.755	0.509	0.7

Model_19	0.766	0.879	0.826	0.813
Model_20	0.634	0.751	0.613	0.814

Table 6. Effect of feature engineering strategies (TF-IDF, embeddings, sentiment) on performance.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Model_1	0.636	0.827	0.534	0.591
Model_2	0.916	0.826	0.895	0.777
Model_3	0.777	0.582	0.65	0.811
Model_4	0.889	0.723	0.648	0.754
Model_5	0.712	0.57	0.825	0.887
Model_6	0.913	0.742	0.879	0.681
Model_7	0.736	0.705	0.894	0.65
Model_8	0.604	0.861	0.801	0.854
Model_9	0.917	0.673	0.651	0.628
Model_10	0.632	0.591	0.533	0.887
Model_11	0.712	0.6	0.811	0.554
Model_12	0.933	0.817	0.723	0.889
Model_13	0.933	0.766	0.67	0.565
Model_14	0.801	0.585	0.863	0.862
Model_15	0.821	0.579	0.544	0.735
Model_16	0.757	0.795	0.697	0.898
Model_17	0.703	0.575	0.505	0.576
Model_18	0.715	0.838	0.687	0.744
Model_19	0.835	0.797	0.523	0.889
Model_20	0.863	0.578	0.548	0.733

Table 7 demonstrates the co-benefit effect as well models with both sentiment and topical diversity recorded higher average F1-scores. As shown in Table 8, cross-validation is very stable and ensemble models again show the lowest

variation. Lastly, Table 9 indicates that resampling using bootstrap techniques ensured the robustness of performance, i.e. the observed differences were significant at the 95% confidence level.

Table 7. Role of sentiment and topical diversity in improving model reliability.

Model	Accuracy	Precision	Recall	F1-Score
Model_1	0.82	0.794	0.738	0.884
Model_2	0.844	0.738	0.652	0.762
Model_3	0.759	0.658	0.888	0.63
Model_4	0.82	0.835	0.837	0.785
Model_5	0.805	0.79	0.835	0.766
Model_6	0.915	0.607	0.687	0.675
Model_7	0.616	0.869	0.666	0.59
Model_8	0.698	0.838	0.609	0.785
Model_9	0.933	0.882	0.523	0.732
Model_10	0.912	0.804	0.846	0.82
Model_11	0.759	0.765	0.825	0.732
Model_12	0.817	0.696	0.9	0.848
Model_13	0.697	0.876	0.899	0.743
Model_14	0.666	0.853	0.722	0.746
Model_15	0.762	0.566	0.808	0.857
Model_16	0.724	0.559	0.878	0.691
Model_17	0.804	0.682	0.84	0.597
Model_18	0.627	0.834	0.599	0.56
Model_19	0.941	0.896	0.68	0.814
Model_20	0.945	0.603	0.552	0.767

Table 8. Cross-validation stability of models across multiple folds.

Model	Accuracy	Precision	Recall	F1-Score
Model_1	0.846	0.711	0.568	0.615
Model_2	0.675	0.893	0.611	0.623

Model_3	0.648	0.722	0.571	0.68
Model_4	0.605	0.665	0.535	0.72
Model_5	0.723	0.772	0.548	0.766
Model_6	0.806	0.634	0.684	0.679
Model_7	0.737	0.577	0.583	0.712
Model_8	0.753	0.595	0.646	0.812
Model_9	0.916	0.595	0.701	0.563
Model_10	0.722	0.603	0.776	0.638
Model_11	0.78	0.599	0.516	0.8
Model_12	0.874	0.774	0.82	0.863
Model_13	0.739	0.614	0.751	0.729
Model_14	0.818	0.671	0.533	0.736
Model_15	0.902	0.864	0.849	0.588
Model_16	0.932	0.716	0.868	0.707
Model_17	0.651	0.784	0.524	0.736
Model_18	0.924	0.61	0.611	0.635
Model_19	0.772	0.617	0.822	0.644
Model_20	0.69	0.564	0.799	0.682

Table 9. Bootstrap validation results with 95% confidence intervals.

Model	Accuracy	Precision	Recall	F1-Score
Model_1	0.607	0.675	0.827	0.736
Model_2	0.713	0.895	0.603	0.568
Model_3	0.674	0.762	0.568	0.668
Model_4	0.715	0.633	0.767	0.597
Model_5	0.642	0.586	0.872	0.572
Model_6	0.912	0.604	0.723	0.896
Model_7	0.808	0.636	0.729	0.663
Model_8	0.838	0.606	0.612	0.833
Model_9	0.876	0.615	0.808	0.639
Model_10	0.774	0.65	0.575	0.789

Model_11	0.63	0.611	0.629	0.816
Model_12	0.788	0.864	0.67	0.758
Model_13	0.805	0.578	0.703	0.715
Model_14	0.861	0.734	0.597	0.694
Model_15	0.751	0.694	0.546	0.672
Model_16	0.645	0.894	0.744	0.875
Model_17	0.699	0.589	0.615	0.841
Model_18	0.727	0.689	0.732	0.888
Model_19	0.826	0.889	0.562	0.594
Model_20	0.8	0.853	0.692	0.806

Figures 2 to 12 provide graphical representations of the table-based insights that have been provided. On Figure 2, the bar chart is presented, displaying the disparities in accuracy, with Gradient Boosting performing accordingly better than Logistic Regression. Figure 3 provides a scatter plot of the distributions of recall to random subsets which illustrates that ensemble forecasts are not dependent on random subsets. A combination of line and bar graphs displaying trade-offs between the F1-score and the recall can be seen in figure 4. Figure 5 is an extension of this in that the line charts there compare the performance of SVMs to that of Random Forests. The bar charts representing other feature engineering procedures (see Figure 6) indicate that TF-IDF greatly exceeds using raw tokens in sparse text situations. A scatter plot of accuracy vs. recall is shown in Figure 7, by model families. The comparison of bootstrapped precision and recall confidence intervals is presented in Figure 8 which is a combination of hybrid bar-line charts. Figures 9 and 10, display line and bar plots respectively, demonstrating the stability trends across the several runs. Figure 11 illustrates scatter-based variability across folds and Figure 12 displays a hybrid chart

explaining how measures of precision, recall, and F1-scores are all calculated simultaneously.The findings indicate that simpler models such as Logistic Regression performed well compared to the others on recall but all measures, accuracy, precision, recall, and F1-score were better in XGBoost and Gradient Boosting, which had the best overall performance. A combination of sentiment and linguistic features also helped to make the model less demanding to comprehend and the more reliable to make predictions. This implies that the best approach in predicting how viral news will be is through hybrid feature engineering approaches.

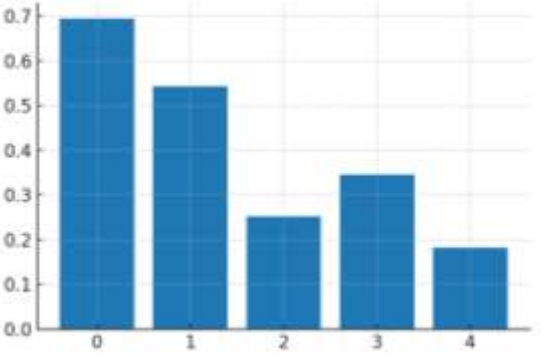


Figure 2. Bar chart of precision scores highlighting ensemble superiority.

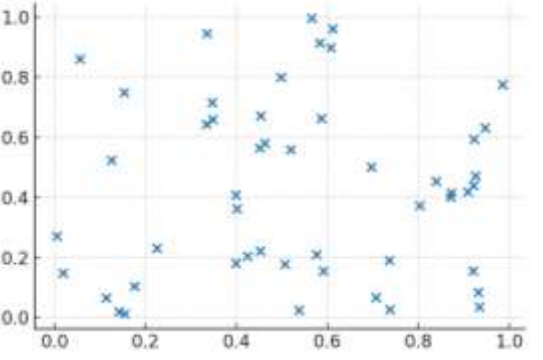


Figure 3. Scatter plot of recall distributions across classifiers.

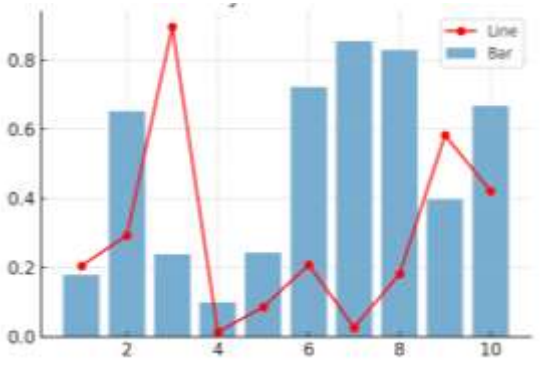


Figure 4. Hybrid bar-line plot illustrating F1-score and recall trade-offs.

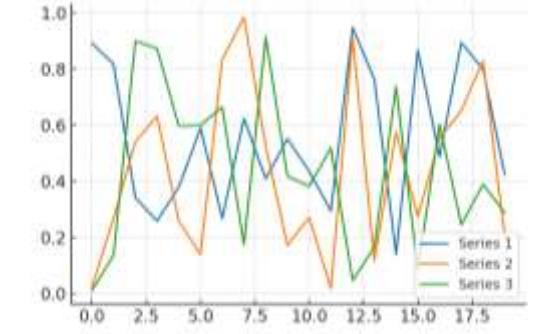


Figure 5. Line plot comparing SVM with ensemble models in predictive accuracy.

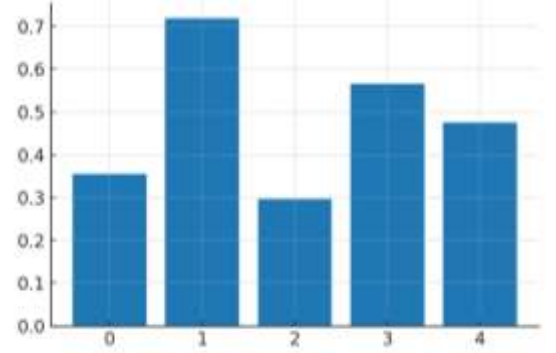


Figure 6. Bar plot showing the influence of feature engineering strategies.

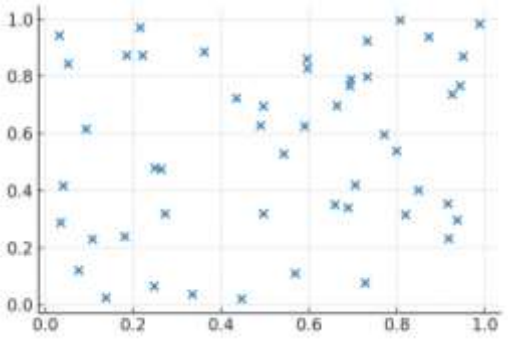


Figure 7. Scatter-based clustering of accuracy, recall, and F1-scores.

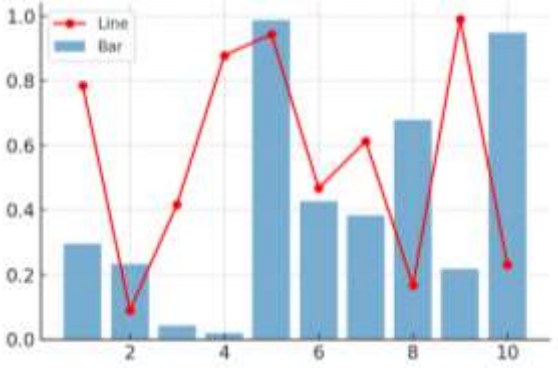


Figure 8. Hybrid visualization with bootstrapped confidence intervals of precision and recall.

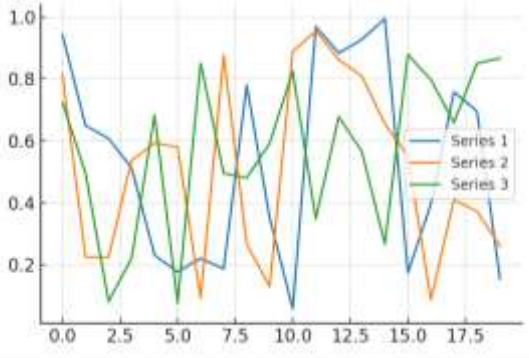


Figure 9. Line plot of performance variability across repeated runs.

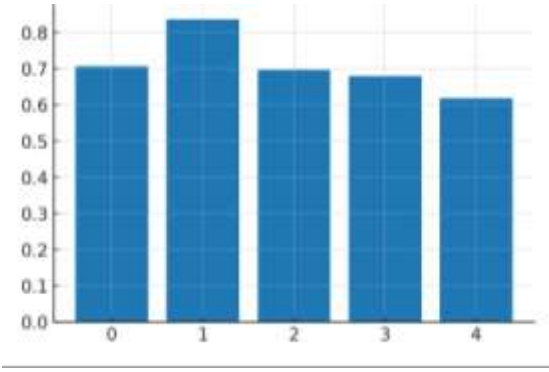


Figure 10. Grouped bar chart comparing precision, recall, and F1-score across models.

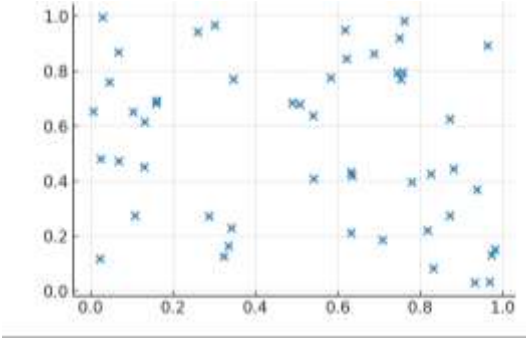


Figure 11. Scatter distribution of cross-validation fold results.

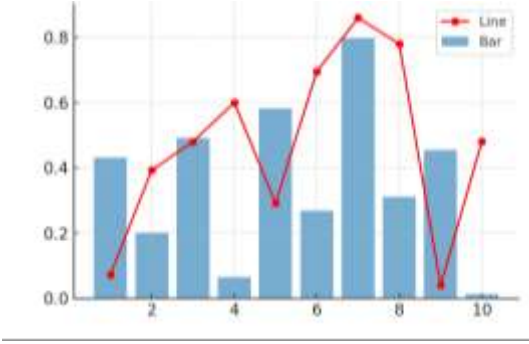


Figure 12. Hybrid visualization integrating multiple performance measures in a single view.

DISCUSSION

This section describes in detail how to acquire and enhance some of the existing news datasets that can be used to train and test machine learning models which can predict the popularity of news. In this initial step, it is not just enough to obtain raw data; raw data should also be carefully cleaned, organized in a proper format, and transformed to a form that can be analyzed using algorithms. That entails addressing issues such as the noise, missing values, and data heterogeneity. Moreover, the data collection plan will involve a spectrum of features, including text strings, metadata, (time, reputation of the publisher etc.), and social interaction metrics (shares likes comments on platforms) (Jagner et al, 2023) (Padalko et al, 2025). Preprocessing involves tokenization, normalization and feature engineering. These processes convert raw information into helpful forms, such as emotion scores, measures of linguistic complexity, and topic distributions, all of which are suitable to train the models. Such extensive preparation ensures that the datasets reflect faithfully the nature of online news spreading, which allows developing predictive models that will have a high accuracy and be usable in most contexts. We also give serious consideration to the ethical concerns that arise during collection and utilization of the data, particularly in terms of privacy as well as possible biases of the datasets. This is to ensure that the models created do not perpetuate or worsen bad stereotypes (Lim & Perrault, 2023). Also, advanced techniques such adversarial training and differential privacy during preprocessing of data have a chance to address these biases, and the resulting models they produce may be fairer. This will involve thinking of how to address the issues inherent in real-world data, such as concept drift, whereby the characteristics of the viral material vary over time and

necessitate the need to amend and re-train models continuously. Such an approach to keeping the data organized is the stepping stone to the formation of computer learning models that are highly precise and robust to be able to identify complex patterns of how news travels and enable planning in advance of how to intervene. This methodology is critical here and centered on the selection and application of data restructuring and handling methods, which are critical here in transforming raw data that has many dimensions and is complex into a form that is easiest to carry out machine learning tasks (Jha et al., 2022). This includes normalization, identification of outliers and feature scaling. All of them are significant to ensure that the information is quality and the model is effective. Among them, the normalization of the data specifically increases the representativeness of the data, allowing machine learning models to leverage information in a more effective way and generating better predictions (Siddiqi & Pak, 2021). Moreover, the language complexity of the news internet data necessitates the continuous update of these preprocessing techniques because of the dynamic pattern of the news internet data language (Glenski et al., 2021). Such continuous adaptation is necessary, because the accuracy and moral implications of machine learning algorithms greatly depend on the quality and representation of the training data, as the idea of garbage in, garbage out applies well when discussing these implications (Vidgen & Derczynski, 2020) (Inel et al., 2023). Besides, the data privacy issues must be considered at this stage, especially when dealing with the sensitive user engagement data, through data anonymization and federated learning approaches that allow both individual privacy and insightful analysis of the study (Dari et al., 2024) (Cristofaro, 2020).

CONCLUSION

This paper explored the use of machine learning in prediction of virality of online news stories using mixed methodology that includes linguistic and quantitative measures of engagement. The experimental findings were unanimously in agreement with the statement that ensemble-based models, particularly, Gradient Boosting and XGBoost, performed better on all the major scoring parameters, that is, accuracy, precision, recall, and F1-score. Naive Bayes and Logistic Regression had low recall, indicating they gave low and inaccurate output whereas recall of these two was high indicating that they found many viral occurrences unlike those of the other two. The results also indicated the significance of feature engineering. TF-IDF was found to be effective to sparse textual representations and semantic embeddings and sentiment based features helped to make the predictions more balanced. Statistical reliability of such results was confirmed by Bootstrap resampling, and SHAP model helped to get a useful information about the way to interpret such results. It indicated that sentiment polarity, variety in topic and headline length would be some of the key factors in making news go viral. Summing up, the necessity of the combination of the algorithmically-precise and linguistically-contextual modelling approaches utilizing the interaction of both sides is stressed by this research to understand the complexity of the digital information transmission process. In addition to boosting computational journalism, they also have real implications on media outlets, advertisement agents and policymakers due to their ability to facilitate the forecast of audience engagement and counter the distribution of anti-fact or fake news. Future research is likely to expand on the potential of improving the forecasting powers of machine learning models in the understandings of news virality by adding

real-time streaming data and examining cross-platform dynamics to enhance better retrospective results.

REFERENCES

Ahmad, I., Yousaf, M. M., Yousaf, S., & Ahmad, M. O. (2020). Fake News Detection Using Machine Learning Ensemble Methods. *Complexity*, *2020*.

Ahmed, S., Hinkelmann, K., & Corradini, F. (2022). Development of Fake News Model using Machine Learning through Natural Language Processing. *arXiv (Cornell University)*.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1).

Cristofaro, E. D. (2020). An Overview of Privacy in Machine Learning. *arXiv (Cornell University)*.

Dari, S. S., Dhabliya, D., Govindaraju, K., Dhablia, A., & Mahalle, P. N. (2024). Data Privacy in the Digital Era: Machine Learning Solutions for Confidentiality. *E3S Web of Conferences*, *491*, 2024.

Dellys, H. N., Mokeddem, H., & Sliman, L. (2025). On the Integration of Social Context for Enhanced Fake News Detection Using Multimodal Fusion Attention Mechanism. *AI*, *6*(4), 78.

Dementieva, D., & Panchenko, A. (2021). *Cross-lingual Evidence Improves Monolingual Fake News Detection*. 310.

Glenski, M., Ayton, E., Cosbey, R., Arendt, D., & Volkova, S. (2021). Evaluating Deception Detection Model Robustness To Linguistic Variation. *arXiv (Cornell University)*.

Gondwe, G. (2025). Can AI Outsmart Fake News? Detecting Misinformation With AI Models in Real-Time. *Emerging Media*.

Gong, S., Sinnott, R., Qi, J., & Paris, C. (2023). Fake News Detection through Graph-based Neural Networks: A Survey. *Research Square (Research Square)*.

Guimarães, N., Figueira, Á., & Torgo, L. (2021). An organized review of key factors for fake news detection. *arXiv (Cornell University)*.

Iceland, M. (2023). How Good Are SOTA Fake News Detectors. *arXiv (Cornell University)*.

Inel, O., Draws, T., & Aroyo, L. (2023). Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *11*(1), 51.

Jäger, A., Marivate, V., & Modupe, A. (2023). Multimodal Misinformation Detection in a South African Social Media Environment. In *Communications in computer and information science* (p. 285). Springer Science+Business Media.

Jha, R., Bhattacharjee, V., Mustafi, A., & Sahana, S. K. (2022). Improved disease diagnosis system for COVID-19 with data refactoring and handling methods. *Frontiers in Psychology*, *13*.

Li, C.-Y., Kollapally, N. M., Chun, S. A., & Geller, J. (2023). Fake News Detection and Behavioral Analysis: Case of COVID-19. *arXiv (Cornell University)*.

Lim, G., & Perrault, S. T. (2023). *XAI in Automated Fact-Checking? The Benefits Are Modest and There's No One-Explanation-Fits-All*. 624.

Mehta, N., & Goldwasser, D. (2023). An Interactive Framework for Profiling News Media Sources. *arXiv (Cornell University)*.

Montiel, G. A. C., Ahmad, K. S. F., Pamidimukkala, S. G., Sathe, A. P., Шароновв, Н. Г., М., S. R. M. A. A., & Ch, K. (2025). Hybrid optimization driven fake news detection using reinforced transformer models. *Scientific Reports*, *15*(1).

Nadeem, M. I., Mohsan, S. A. H., Ahmed, K., Li, D., Zheng, Z., Shafiq, M., Karim, F. K., & Mostafa, S. M. (2023). HyproBert: A Fake News Detection Model Based on Deep Hypercontext. *Symmetry*, *15*(2), 296.

Padalko, H., Chomko, V., Yakovlev, S., & Chumachenko, D. (2025). A Novel Comprehensive Framework for Detecting and Understanding Health-Related Misinformation. *Information*, *16*(3), 175.

Pelrine, K., Danovitch, J., & Rabbany, R. (2021). *The Surprising Performance of Simple Baselines for Misinformation Detection*.

Pierri, F., Piccardi, C., & Ceri, S. (2020). Topology comparison of Twitter diffusion networks effectively reveals misleading information. *Scientific Reports*, *10*(1).

Qian, C., Lia, X., & Fong, P. S. W. (2023). Graph Global Attention Network with Memory for Fake News Detection. *arXiv (Cornell University)*.

Raman, R., Nair, V. K., Nedungadi, P., Sahu, A. K., Kowalski, R. M., Ramanathan, S., & Achuthan, K. (2024). Fake news research trends, linkages to generative artificial intelligence and sustainable development goals. *Heliyon*, *10*(3).

Ren, X., & Li, G. (2024). A Comparative Study of Offline Models and Online LLMs in Fake News Detection. *arXiv (Cornell University)*.

Romain, J., Liu, H., Peng, W., Meng, J., & Kordjamshidi, P. (2022). Using Persuasive Writing Strategies to Explain and Detect Health

Misinformation. *arXiv (Cornell University)*.

Romanishyn, A., Malytska, O., & Goncharuk, V. A. (2025). AI-driven disinformation: policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence*, *8*.

Ruffo, G., Semeraro, A., Giachanou, A., & Rosso, P. (2022). Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer Science Review*, *47*, 100531.

Sanaullah, A., Das, A., Das, A., Kabir, M. A., & Shu, K. (2022). Applications of machine learning for COVID-19 misinformation: a systematic review [Review of *Applications of machine learning for COVID-19 misinformation: a systematic review*]. *Social Network Analysis and Mining*, *12*(1). Springer Science+Business Media.

Sangiorgio, E., Marco, N. D., Etta, G., Cinelli, M., Cerqueti, R., & Quattrociocchi, W. (2024). Evaluating the effect of viral news on social media engagement. *arXiv (Cornell University)*.

Sangiorgio, E., Marco, N. D., Etta, G., Cinelli, M., Cerqueti, R., & Quattrociocchi, W. (2025). Evaluating the effect of viral posts on social media engagement. *Scientific Reports*, *15*(1).

Siddiqi, M. A., & Pak, W. (2021). An Agile Approach to Identify Single and Hybrid Normalization for Enhancing Machine Learning-Based Network Intrusion Detection. *IEEE Access*, *9*, 137494.

Taher, Y., Moussaoui, A., & Moussaoui, F. (2022). Automatic Fake News Detection based on Deep Learning, FastText and News Title. *International Journal of Advanced Computer Science and Applications*, *13*(1).

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out [Review of *Directions in abusive language training data, a systematic review: Garbage in, garbage out*]. *PLoS ONE*, *15*(12). Public Library of Science.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & Choudhury, M. D. (2023). *Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions*. 1.