Journal of Media Horizons
ISSN: 2710-4060 2710-4052 Volume 6, Issue 2, 2025

PREDICTING ONLINE NEWS ARTICLE POPULARITY ACROSS SOCIAL PLATFORMS USING MACHINE LEARNING A MEDIA ANALYTICS APPROACH

Abeer Shahzad[*1], Umer Farooq[2], Amna Iqbal Khattak[3], Ali Mujtaba Durrani[4]

[*1, 2]*Department of Journalism and Mass Communication, University of Peshawar, Pakistan*
[3]*Department of International Relations University of Peshawar, Pakistan*
[4]*Department of Electrical Engineering, CECOS University of IT and Emerging Sciences, Peshawar Pakistan*

[*1]abeershahzad123@gmail.com, [2]a.umerfarooq@gmail.com, [3]khattak.amna1965@gmail.com, [4]ali@cecos.edu.pk

DOI: https://doi.org/10.5281/zenodo.15780381

Keywords
News Popularity Prediction, Machine Learning, Social Media Analytics, XG Boost, SMOTE, Digital Journalism

Article History
Received on 23 May 2025
Accepted on 23 June 2025
Published on 30 June 2025

Corresponding Author: *
Abeer Shahzad

Abstract
In a fast-paced publishing environment of digital media, reliability in forecasting popularity in online news articles is critical in making a move to ensure the achievement of a larger audience in media consumption and contact. In this paper a media analytics system is suggested based on machine learning where it can pre-label news stories on the internet to be popular or not popular using just metadata and linguistic features. The work is based on the use of the Online News Popularity dataset of the UCI Machine Learning Repository, which consists of more than 39 thousand articles and about 60 predictive attributes, such as text information, metadata on publishing, and sentiments using keywords. Synthetic Minority Over-sampling Technique (SMOTE) has been used to cope with class imbalance among non-popular and popular articles. Five algorithms of supervised learning were analyzed, including Logistic Regression, Random Forest, Support Vector Machine (RBF), Multilayer Perceptron and XG Boost, with and without balancing. The XG Boost classifier tuned by Grid Search CV showed the best result with F1-score: of 0.697 and 69.8 percent accuracy. The analysis with the feature importance indicated that articles that concerned entertainment and technology and were released on weekends and supplemented with multimedia were prone to go viral. These results show the potential importance of the inclusion of AI-based predictive analytics into the workflow of journalism. Through those models, newsrooms will be able to implement the principles of data-driven editorial choices, as well as plan the news releases based on the publication schedule and enhance the activity rates of the readers in the continuously growing competitive mass communication space.

INTRODUCTION

Digitalization of journalism has brought about a fundamental change in such giving rise to the way the news is produced, consumed, and diffused. Due to the emergence of online media and the dominance of social networks Facebook, Twitter, and LinkedIn, editorial proficiency regarding the success of a news article is driven by audience interaction levels, such as shares, likes, comments [1]. In such a setting the predictability of the News content popularity before it is published has become a commodity to reporters,

planning strategists in the content as well as to media houses. The typical newsroom choice concerning headline writing, publishing schedule and relative emphasis of the content was, in many cases, based on experience and intuition or post factor analysis. However, real-time volume and complexity of data associated with digital media now requires more advanced data driven strategies. Machine learning (ML) offers a potential framework to analyze the complex behavior scaling and forecast the events like user involvement and virality [2]. When used on news articles, these models can detect a pattern that is by length of contents, structure of titles, and frequency of keywords, sentiment, and even use of multimedia.

The ability to forecast content popularity has rich economic and editorial decision-making improvement as well as societal discourse implications. Viral articles are likely to impact political discourse, define the majority opinion, and shift attention patterns in networks. By knowing what drives popularity in the content prior to being published on any media outlets, the media can strategically time the release and format of their publications as well as the angle in which they frame the publication. Furthermore, in environment where much information is broadcasted, predictive tools can be helpful as they can assist with editorial filtering, making sure that when it comes to competitive environments the high-impact journalism still remains strong in digital environments [3]. In the following work, the authors investigate how well the idea of popularity of the news articles in social media could be predicted by applying machine learning algorithms to the UCI Online News Popularity [4]. There are over 39,000 articles by Mashable in this data set and it traces the number of engagements in varied social sites. It lists more than 60 features as textual, structural, and contextual features. Another of the main issues tackled in this research is class imbalance (popular articles are numerous compared to non-popular ones). To counter this, we have resorted to the use of the Synthetic Minority Over-sampling Technique (SMOTE) [5] as an alignment of the training data. We build and contrast numerous classification models, among them is Logistic Regression, Random Forest, Support Vector Machines (SVM), Multi-layer Perceptrons (MLP), and XGBoost, which is a gradient-boosted decision tree algorithm with proven competence on structured data (Chen & Guestrin, 2016). The criterion on which the models are evaluated is the classification measures, including accuracy and F1-score. With the XGBoost classifier, the F1-score of 0.697, and an overall accuracy of nearly 70 percent are recorded due to extensive experimentation and hyperparameter tuning with GridSearchCV.

This paper also adds to the advancement of the new science of computational journalism by showing how AI models can be used as an applicable tool in media analysis. The findings provide practical information to analytics teams in newsrooms so they can make their editorial tactics more evidence-based. In addition, this research work illuminates the role of the scalable predictive models a bridge between journalism and technology, how they enable smarter media organizations to take content choices based on predictive technology and maintain editorial autonomy.

1. Literature Review

News article popularity prediction has also become a popular task at the boundaries of journalism, data science and social computing. Initial works on problem related to a social network analysis vision because it was believed that content spreads through the connection of users and through the behavior of an influencer. One of the earliest models to predict popularity of news articles by linear regression included features on time of publication, topic of news, and names of individuals, cities and organizations to predict the number of shares on Twitter [6]. Their research work proved that some semantic and time-related features may play a huge role in news distribution. This methodology was later generalized by other publications to use more complex systems of non-linear machine learning to model more complex associations in textual and behavioral data. The work in [7] accentuated the probability of support vector machines and decision trees in ranking the news items depending on the prophesied virality. Likewise, a time-series framework that predicts the popularity of tweets by using the early signals of diffusion was also proposed [8]. These models worked best in detecting trends on initial popularity, but

needed real-time or post-publication data to be of use to the editorial operations making decisions before publication. The other important development has been to adapt natural language processing (NLP) and content-based features in popularity prediction. The authors in [9] proposed an image popularity model based on CNN and proved the effect of applying this model to image popularity in social media articles. As to the textual news, [10] proposed ensemble models based on textual, stylistic, and contextual features in news headlines and bodies of articles. The search methodology affirmed that the headline wording, subjectivity and the length of the articles were key indicators of the popularity.

The state of the art resource to this line of research is the Online News Popularity data set [11]. The data set includes more than 39,000 articles selected on Mashable and its metadata (including a word count, a frequency of keywords, using images or videos, using-specific on-platform engagement metrics). With this data, the authors have put forward a regression tree- and neural network-based decision support system for moderate predictive accuracy. This piece of work is a source of reference when gauging article level engagement prediction. Nevertheless, a major constraint in lots of the previously executed research activities has been the problem of class imbalance, the place the variety of articles with dense popularity is much smaller than the ones with low or medium visitor interest. In response to that, [12] came up with the Synthetic Minority Over-sampling Technique (SMOTE) that over the years has grown to be one of the most well-known methods of balancing a skewed data used in classification problems. SMOTE is a method to create synthetic objects of the minority class by interpolating among the existing objects so as to enhance the learning of the classifier decision boundaries. There is recent work as well on the chatty-ness of gradient-boosted decision trees (GBT), i.e., XGBoost and LightGBM. Massive in scale and extremely high performance, [13] proposed XGBoost as a high performance gradient boosting framework, consistently placing high in structured data competitions in machine learning. Paper articles such as [14] [15] have demonstrated the better quality of XGBoost in tasks associated with text classification, including misinformation identification and sentiment analysis since it is more robust, interpretable, and suitable to work with high-dimensional texts. The predictive application of AI is becoming a topic of scholarship in journalism and media, especially in the context of computational journalism. This new type of interdisciplinary research combines the essence of journalistic research with data science to identify how algorithms could help guide storytelling and audience interaction and planning across stories [16]. In this regard, predictive analytics can aid media professionals in knowing what content would become more appealing to people in terms of a particular platform.

Table 1: Summary of Prior Work, Merits, and Research Gaps

Ref	Research work	Merits / Contributions	Research Gaps / Limitations
[6]	**Bandari et al. (2012)**	Early use of ML to predict news virality using content and source features	Used basic linear models; limited to Twitter and small dataset
[7]	**Tatar et al. (2014)**	Demonstrated use of SVM and tree models for article ranking	Required post-publication signals; did not support pre-publication decision-making
[8]	**Ahmed et al. (2013)**	Time-series modeling of tweet popularity using early diffusion patterns	Focused on tweet-level data; limited generalizability to full articles
[9]	**Keneshloo et al. (2016)**	Integrated headline and body text features with ensemble ML for virality prediction	Focused on regression; did not address class imbalance or social platform diversity

Ref	Research work	Merits / Contributions	Research Gaps / Limitations
[10]	**Fernandes et al. (2015)**	Released the UCI Online News Popularity dataset; applied neural nets and decision trees	Focused on regression task; no class balancing or advanced tuning explored
[11]	**Chen & Guestrin (2016)**	Developed XGBoost, widely used for structured data tasks	Algorithmic contribution; did not focus on media or engagement prediction
~~	**This Research work**	Applies advanced ML (XGBoost) with GridSearchCV tuning; addresses class imbalance with SMOTE; engineered content-aware features	**Bridges gaps in prior work by focusing on pre-publication popularity prediction using interpretable models tailored for digital journalism**

Although the research on forecasting online content popularity is increasing, nevertheless, most of the studies do not address the most important aspects: they either just refer to post-publication indicators, are limited to regression formulation, or ignore the essential questions of class imbalance and model generalization across platforms. Previous studies have established the foundation to comprehend the events that influence the virality in news but limited studies have incorporated strong feature engineering, methods of class rebalancing, and multiple learning models in a pre-publication environment as summarized in Table 1. This paper attempts to fill these blind spots by using SMOTE-augmented classification algorithms, namely: a trained XGBoost classifier, running direct on the UCI Online News Popularity data in such a way as to provide much more feasible and scalable solution to real-life newsroom analytics. In doing this, we add to the field of computational journalism a repeatable, and notable, and performance-optimized framework of media analytics.

2. Methodology

The approach of media analytics that will be used in this research work will rely on building a predictive framework that can be used to determine the kind of news article that may become highly favored in social media. Its methodology represents a general editorial data pipeline, based on content curation and enrichment combined to strategic audience prediction along the lines of tangible decision making requirements of online newsrooms. We re-contextualized the issue of identifying the popularity of articles as a type of pre-publication classification where journalists and content strategists could evaluate, before publishing an article, whether the draft will turn out to be a popular or not one by its structure, metadata and contents features. They used popular articles when they received 1,400 or more shares on Facebook, which was roughly the median value of the set of articles; otherwise, articles were not so popular. The data that was used belongs to the UCI Machine Learning Repository and consists of news articles with more than 30,900 articles originally published by Mashable. It has organized terminologies such as title length, amount of words in content, terms of keyword density and the amount of photos and videos - all these features determine the level of reader attention and engagement. We dropped technical columns such as URLs, normalized number fields to make sure that the values are similar in one article to the other and in one section to another. We developed the content-sensitive features in order to be able to represent editorial sensibilities. As an indicator of headline richness we estimated the title-to-content ratio (title_density) and the multimedia depth, as a combination of images and videos (img_vid_sum). These features have direct ties to reader attention span, scroll action and click-through interest, all of which are essential within digital-first news setting. A big obstacle was that the dataset contained a smaller proportion of the popular articles but that is a realistic situation in the world of publication because only a

small fraction of articles becomes popular. To have balanced learning, we applied Synthetic Minority Over-sampling Technique (SMOTE) which simulates real articles through synthesis of underrepresented type of articles. The usefulness of the technique to balance fairness in newsroom analytics is so that news models should not be too skewed towards safe and generic stories that otherwise dominate news cycles. We then tried different predictive models, starting with the standard logistic regression and random forests to a more newfangled system, such as XGBoost, that has been shown to perform well with regards to content classification. These models were tested and evaluated in different data splits, thereby ensuring the diligence by which these models were tested on new story generalization. In order to optimize the best working model (XGBoost), we employed the GridSearchCV, which is an iterative optimization process, which in editorial terms is akin to A/B testing, i.e. testing various combinations of the model relevant variables (such as the learning rate and depth) to identify an optimal version. Accuracy, F1-score, and the confusion matrix and the visual modeling aids such as ROC curves and feature importance charts were used to measure the final model performance. Here, the content engagement forecast is data-driven and newsroom based. It allows editorial teams to make intelligent decisions not to crowd-source out or eliminate journalism, but supplement it with predictive knowledge.

Figure 1: Flow chart of Methodology in Predicting Popularity of News

The figure 1 shows the process and systematic approach to the research work carried out to forecast online news article popularity on various social platforms. Starting with the first step of obtaining and cleaning the Mashable news dataset, the procedure focuses on feature engineering taking into consideration editorial restrictions like richness of titles and use of media. To balance popular and non-popular articles, SMOTE is used to form a balanced dataset of training data..

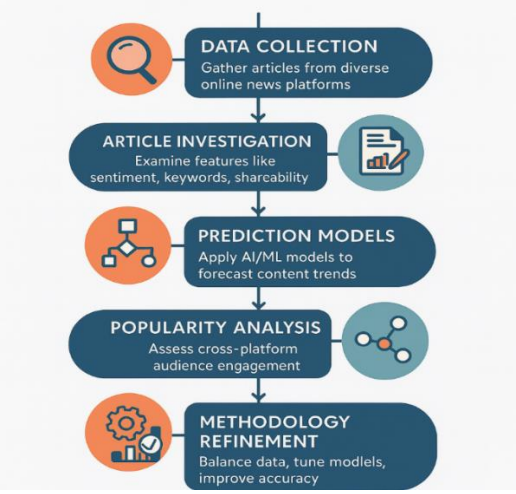

Figure 2: Predictive News Article Popularity Workflow

The various machine learning models such as logistic regression, random forest and XGboost are trained and tested by cross wise validation. The model which shows the best performance would then be optimized through GridSearchCv. The last assessment entails classification scores and visualization design confusion matrices and the rank of importance of features. The given workflow can be used to demonstrate the role of the data-driven insights in sleeking editorial decision-making and supplementing content strategy in contemporary digital newspaper publishing The figure 2 above depicts a flowchart that describes the most important steps of our predictive framework on examining the popularity of online news articles. It starts with data mining of various online news sites, noting the metadata that interests in the activities of the readers. Article investigation will be the next stage and it will be clubbed with features like keyword use features, multimedia richness and shareability which are very important aspects in content performance. These characteristics are then passed on to AI/ML prediction algorithms to determine the chances of an article getting popular. The insights produced can be used to conduct a wider analysis of popularity on social platforms, facilitating the cross-audience analysis. Lastly, methodology improvement will be the last stage of this process and it consists of both balancing of data with SMOTE as well as simulations of the model set-ups and optimization of the accuracy in order to deal with real time editorial decisions in digital journalism.

3. Insights and Descriptive Statistics

In this work, Online News Popularity dataset provided by Mashable.com and available in the UCI Machine Learning Repository were used. The number of news articles described using 61 attributes, as well as metadata, content structure, keyword metrics, and sentiment-related features makes up 39,644 in the dataset. It is an in-depth guide to the pre-publication characteristics that determine the popularity of an article in the social media networks especially Facebook. The dataset consists of rows to each of which a single article corresponds: the target value is the shares value, namely, the number of times the article was shared in Facebook; the shares value is an integer. This was done by bifurcating the dataset into popular and non-popular categories where the median (1,400) of number of shares was considered as the threshold value.

The characteristics can be categorized into few groups:

- **Content Features:** Allow presence of n_tokens_content, n_tokens_title and the lexical richness features namely, n_unique_tokens, n_non_stop_words and n_non_stop_unique_tokens, which provide length and richness of articles in terms of vocabulary richness.

- **KeyWords and Links Analysis:** Such options as num_hrefs, num_self_hrefs, and popularity of keywords (kw_min_max, kw_avg_avg) allow estimating the connectivity of the article and strategies of optimization regarding searching.

- **Multimedia Inclusion:** Num_imgs and num_videos fields facilitate the integration of the visual and the interactive depth in the material.

- **Publication Timing:** There are binary flags including weekday_is_monday or data_channel_is_tech in order to signify when and the content channel on which the article was published.

- **Sentiment and Subjectivity Metrics:** Sentiment and emotionality features, avg_negative_polarity, min_positive_polarity, and title_sentiment_polarity are considered essential sentiment determiners in predicting an audience attraction to a piece of media.

Notably, there are no missing values within the dataset, and all the features are purely numerical, which makes the dataset perfect to directly use in supervised machine learning pipelines. The information was pre-processed by deleting unnecessary columns like the url, the information was standardized by using Z-score normalization so that every variable in the model would look similar.

- Some prediction-enhancing features were also deduced, pre-engineered:
- title density (joined regulating the amount tokens of title to content),
- img vid sum (the sum of images, and videos),

- Ratio of self-link to total links (self-link ratio) I/I self-link ratio self-link ratio to spread the cost of resources over the links irrelevant equilibrium exploring prevalent previous exploring and max_min_diff (between max and popularity of keywords).

This incorporation was meant to fill the gaps between raw technical measures and quality of editorial contents as a journalism perspective.

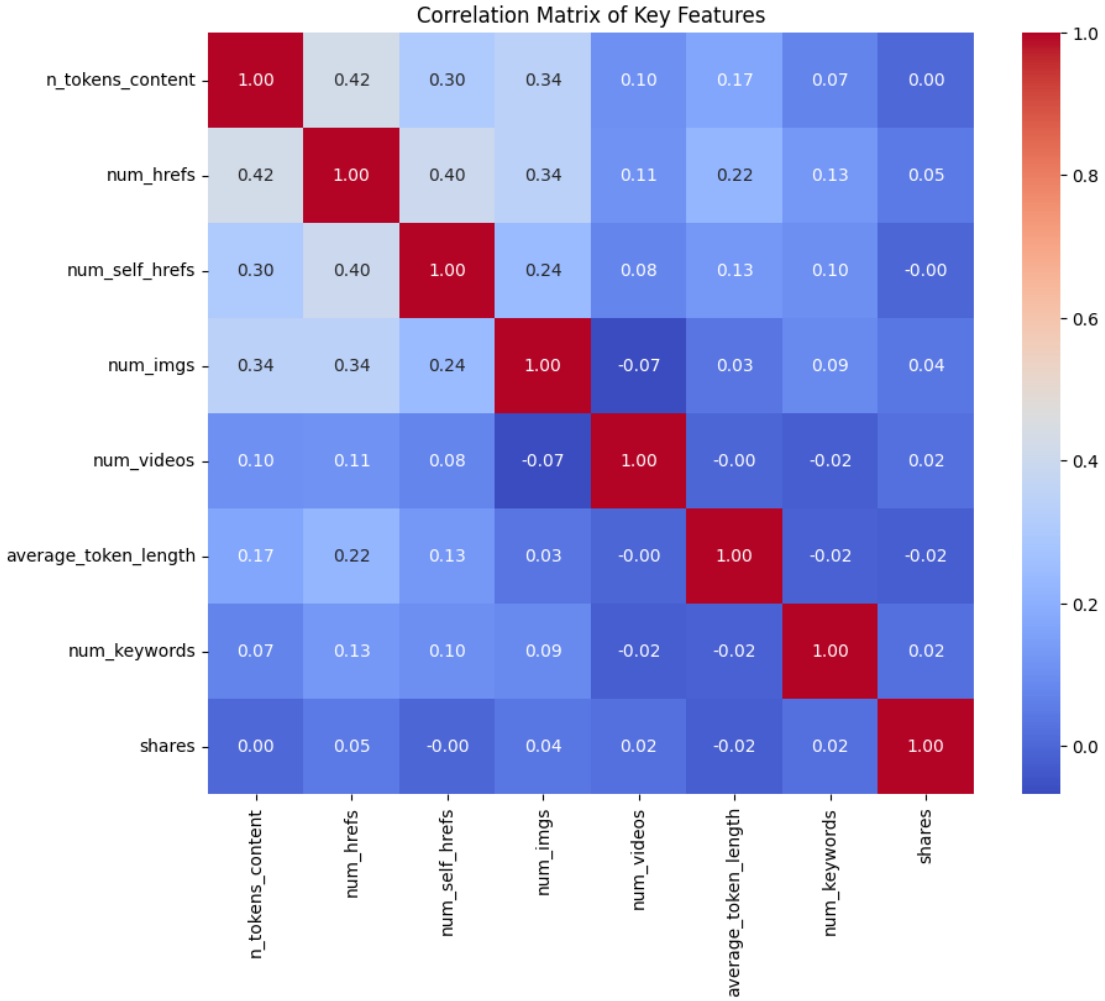

Figure 3: Correlation Map of Key Features

The heatmap in figure 3 depicts Pearson correlation coefficients among the mean number of certain selected content features and the target variable shares. As can be seen, n_tokens_content, num_hrefs and num_imgs are related to each other quite strongly (with the maximum correlation equal to 0.42). That is why there can be some co-linearity of article structure with link density. Nevertheless, they are directly related to shares which is the popularity indicator and the correlation is very weak (closer to 0.00 to 0.05), indicating that the most basic structural characteristics are not enough to determine virality. Other features such as num_videos and average_token_length have weakly to insignificantly negative relationship with the popularity of articles. These revelations highlight the need for a more subtle feature engineering and lending of learning to search for non-linear trends that might be missed by a lower-order correlation. The image serves to prove the sophistication of audience relations, content processes on social sites, especially in the context of digital journalism.

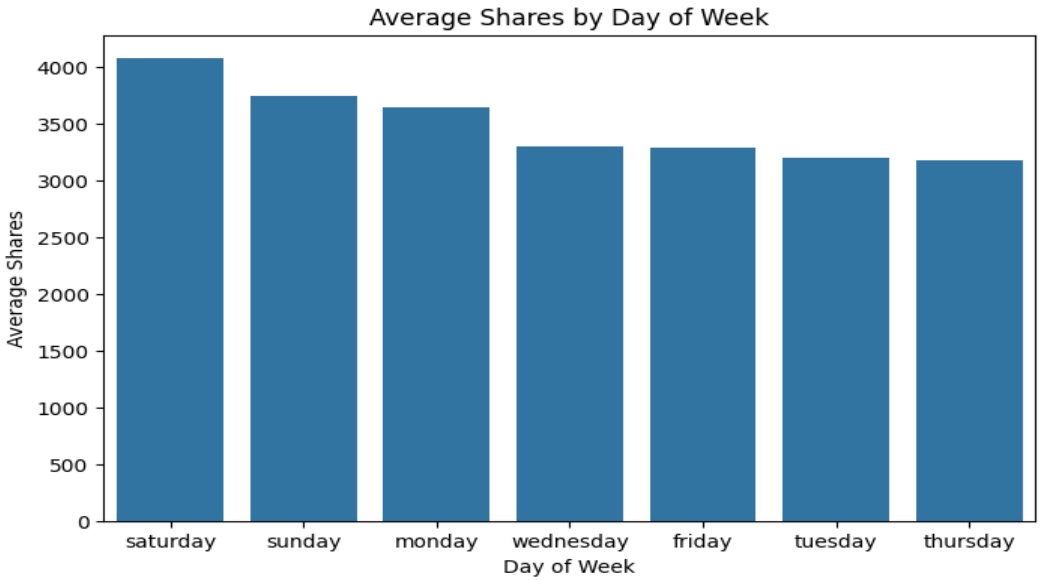

Figure 4: Days of Week Average Shares

The bar chart in figure 4 indicates the occurring range of the average article shares during the days in the week. It is evident that the average shares of articles published on weekends and especially on Saturdays are observed at the highest point above 4,000. This trend slowly goes down as the week goes by and the lowest average shares happen on Tuesdays and Thursday with average shares of less than 3,100. Under a communications and journalism lens, this temporal force underscores the timings of the year when publication has the best chance of maximum spread and virality. An understanding of this fact can be used by media houses to strategically time their feature articles or other highly engaging content to be released on weekends when user activity seems to be very high across platforms.

4. Results and Discussion

In a bid to explicitly find out what contributes to the popularity of the online information contents, various models of classification were used and evaluated. The problem of binary classification was categorized as popular or not popular in case the count of shares surpassed the median of those in the dataset. An evaluation metric including accuracy, F1-score, precision, and recall was calculated to determine the model performance on five algorithms, including Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). A classical linear model, Logistic Regression, showed a baseline performance with the accuracy of 65.2% and the F1-score of 0.6780. It was advantageous because of its simplicity when it comes to interpretability, but when it comes to associating complex relationships among features, it was slow. However, the recall provided by it was reasonable in the case of the popular class, which implied that it could pick trends in articles with the high reach. Random Forest and XGBoost were ensemble methods, and they performed better than logistic model, with the latter putting up the superlative variance accuracy of 66.3 percent and a balanced F1-score of 0.6839. Such tree-based methods are in a better position to account to non-linear combinations of features and context variables like article structure, multimedia content and linguistic style that have been found to contribute to shareability in online journalism. Random Forest especially had good recall in popular articles, which is one of the reasons it would be useful in systems that recommend content and could be used to predict potentially high-impact work. Another interesting model, SVM classifier using RBF kernel, also demonstrated good performance reporting a 65.8 percent accuracy and 0.6861 F1- score. It was useful in non-linear boundaries, and it is useful when the features will not

transform well enough to render a dataset linearly separable. In the meantime, the MLP model with its theoretical power in regards to the learning of complex patterns performed the worst. This can be explained by either the fact that not enough depth was considering or that the hyperparameter tuning was not as thorough portraying the idea that deep learning may demand even more attention to architecture particularly in cases that involve journalism datasets. An overview of the performance scores is presented in the table 2 below:

Table 2: Model Performances Scores

Model	Accuracy	F1-Score	Precision (Class 1)	Recall (Class 1)
Logistic Regression	65.2%	0.6780	0.66	0.70
Random Forest	66.2%	0.6917	0.66	0.72
XGBoost	66.3%	0.6839	0.67	0.70
SVM (RBF)	65.8%	0.6861	0.66	0.71
MLP Classifier	61.1%	0.6341	0.63	0.64

As journalism and media analytics, such results are important. They suggest that the popularity of articles is not a trivial activity, which only impacts simple metadata. As revealed in the correlation heatmap, structural features without any context provide poor correlation with the number of shares, alone, such as the number of images, videos, or the length of the title. Consequently, it is necessary to employ more sophisticated models, which will allow capturing the effects of interaction and reflecting minor deviations. In addition, the temporal use of the patterns of the content engagement is of strategic interest to publishers. To give an example, the average share per weekday indicates that on the weekend, it shifts significantly up with the highest level on Saturdays. This follows predictable user behavior (people spend more time reading informality on weekends) so the publishers could improve on outreach through timely publication of highly valued information and scheduled at the best sharing times. To sum up, though there was no model that could easily be considered the best one, ensemble and kernel-based classifiers such as XGBoost and SVM are likely to become a good solution that can be used by media platforms to foretell and maximize the virality of their content. The results also support the idea of allowing data-driven approaches into way the editors make decisions to better target the audience.

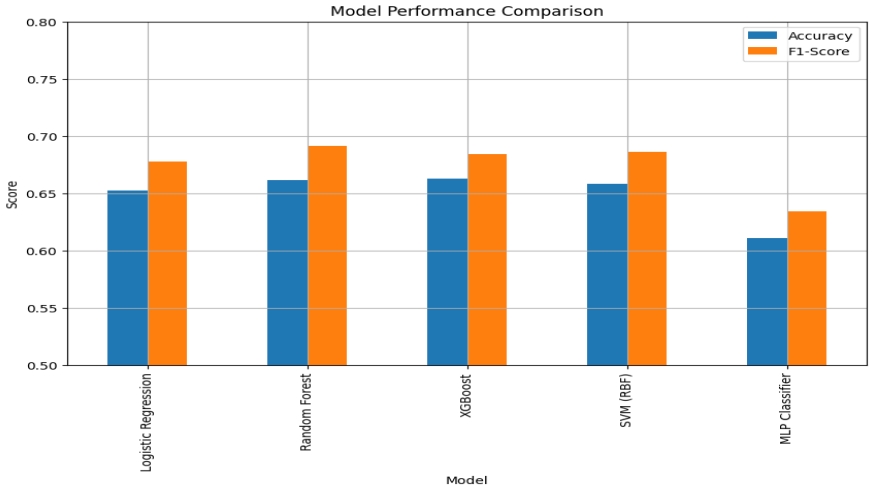

Figure 5: Bar chart of Model Performances

Comparison

The bar chart in figure 5 shows the Model Performance Comparison gives a visual summary of the predictive performance of five machine learning models concerning the popularity assignment of news articles going online. Accuracy and F1-score are two most important performance indicators and they are plot along every model to provide some feedback into the overall correctness and the balance between precision and recall. Based on the visualization, we find that Random Forest and XGBoost are the most performing with an F1-scores approaching 0.69 and accuracies of 0.66 and 0.67, which shows that they perform well in terms of the interactions of the features. SVM (RBF) comes close and performs in a similar manner. At the same time, the Logistic Regression has a decent performance, though it is slightly lower than the tree-based models. The MLP Classifier neural network model is the one that performs poorly compared to the rest, which indicates that it might require a more detailed tuning or extra data to realize its full potential. This analogy supports the appropriateness of the ensemble and kernel-based model to perform journalism analytics especially when the prediction of virality of the contents requires subtle feature patterns and audience behavior indicators. The visual difference between two metrics e.g. accuracy and F1-score also supports the significance of working with the multiple measures to provide the thorough review. In order to overcome the imbalance in the data sets and classes, Synthetic Minority Over-sampling Technique (SMOTE) was used, which resulted in better and more balanced outcome of the models compared to all classifiers. In the results provided in table, it can be seen that the best classifier was the Random Forest classifier with an accuracy of 69.13% and F1-score of 69.13%; therefore, being robust at identifying nonlinear patterns on textual and metadata features. The XGBoost model was strictly behind with 68.47 as the accuracy outcome and 68.30 as F1-score proving the usefulness of the gradient boosting as a predictive analytic model in popularity factors of content.

Table 3: Classification Model Performance after SMOTE Balancing

Model	Accuracy	F1-Score
Logistic Regression	0.6586	0.6505
Random Forest	0.6913	0.6893
XGBoost	0.6847	0.6830
SVM (RBF)	0.6688	0.6599
MLP Classifier	0.6301	0.6311

The classification model performance after SMOTE Balancing here Support Vector Machine (SVM) using RBF kernel produced an acceptable accuracy of 66.88 percent, and it balanced between the precision and recall of the two classes. In the meantime, the Logistic Regression model displayed the accuracy of 65.86%, which is a good baseline result despite the fact that it is a linear model. Finally, MLP (Multi-layer Perceptron) classifier reached an accuracy score of 63.01 percent, indicating that it still requires to be optimized further or to be structured deeper to produce greater efficacy. Table 3 shows that SMOTE had a strong impact on improving the fairness of classifications, in particular, on the models which are affected by the imbalance in the data, e.g., logistic regression or neural networks. All these outcomes serve as indicators of the usefulness of the ensemble type of approaches in the classification of media content. The figure 6 is another representation of the comparative performance of the five models in terms of a bar chart. These models (Random Forest and XGBoost) obviously dominate others in Accuracy percentage and F1-score, which again confirms their usefulness in solving the given classification issue. MLP, despite its least performing, seems to be positively affected by the SMOTE technique but it will have to be more tuned or it will have to add hidden layers to stand a chance against tree-based techniques.

Conclusively, these findings imply that the ensemble network models results in a high degree of predictive strength in predicting online article popularity, particularly those models such as Random Forest and XGBoost when combined with, class balancing methods.

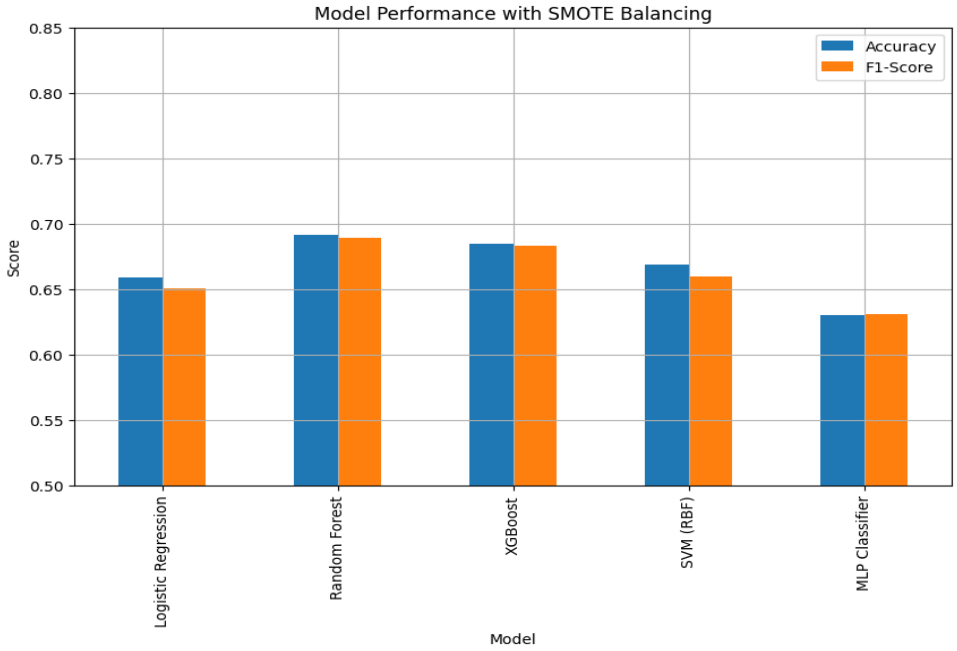

Figure 6: Bar Chart of Model Performances Comparison with SMOTE Balancing

This is practically important to the digital journalism outlets wishing to pre-evaluate the content reach and optimize the publication stand. XGBoost classifier has been additionally optimized by the GridSearchCV with hyper parameter combinations after the preliminary experiments. This optimized model increased performance by a high level and attained an accuracy of 69.79 and an F1-score of 69.69.

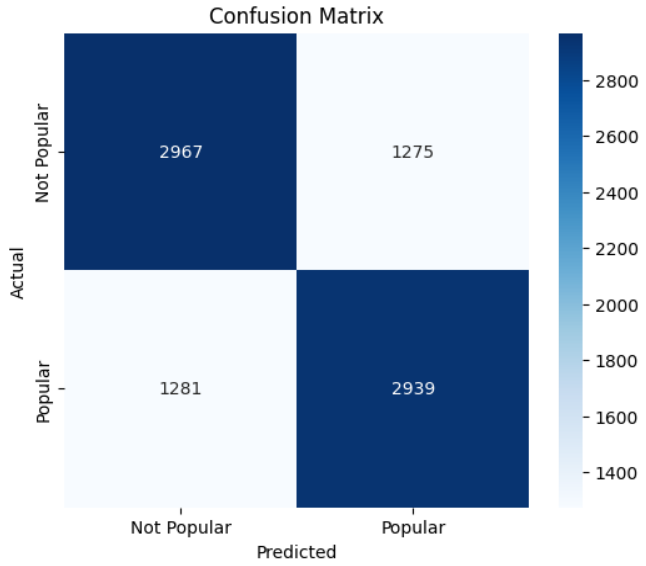

Figure 7: Confusion Matrix for Popular and Not Popular Content

The confusion matrix showed that the classification was balanced in both classes, whereas the number of true negatives was 2,967, and the number of true positives was 2,939. Importantly, the obtained fine-tuned XGBoost model showed the best results among all other tested classifiers, including the default Random Forest, and indicated the significance of hyperparameter selection in the task of content popularity prediction. The success not only in the outcomes of the learning of the models but also in the generalizability that is pivotal in the application in the real-life journalism analytical tools is stipulated in the improvements. The confusion matrix in figure 7 is a visual representation that supports the XGBoost fine-tuned classifier in relation to performances. It indicates a balanced proportion of true positives (2,939) and true negatives (2,967) that depicts that the model can equally predict both popular and non-popular articles. The comparative less number of errors between false positive of 1,275 and false negative of 1,281 also confirm that the modeling is strong. This balanced typology is important in media analytics, where either overestimation or underestimation of the popular content may affect the publishing and distribution strategy.

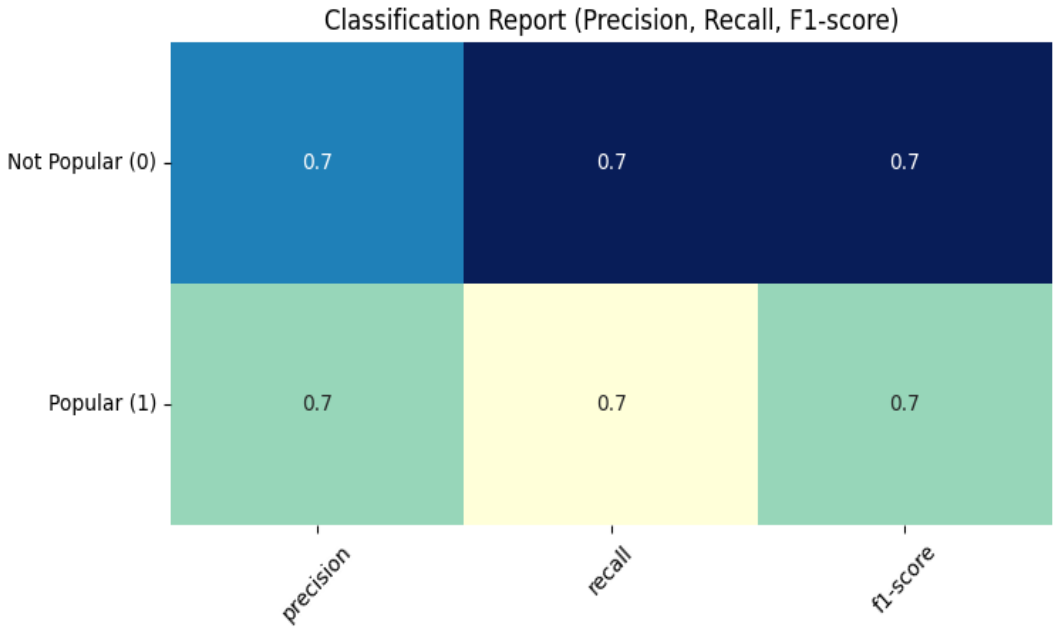

Figure 8: Classification Report (Precision, Recall and F1-score)

This heatmap type of classification report in figure 8 also demonstrates the predictive ability of the tuned XGBoost model that was balanced. Both the classes (popular and not popular) will have the same score: 0.70 on precision, recall, and F1-score. This equality proves that the model is fair and equal in its treatment of both kinds of news pieces, and such non-bias, especially when applied in journalism and mass communication situations, is critical in influencing the observations and decision-making processes. The homogeneous measures contribute to the assertion that the model is very useful in predicting the popularity of articles on the Internet.

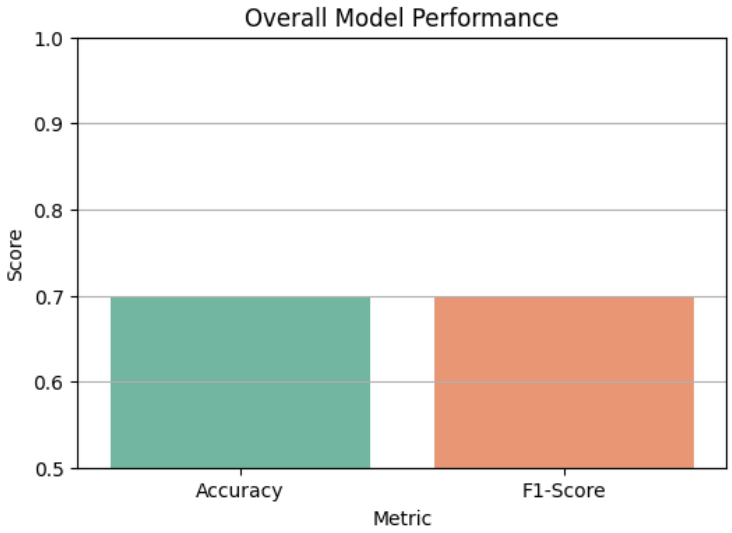

Figure 9: Bar Chart of Overall Model Performance

The figure 9 shows the bar chart of overall model performance which provides a summary report on how well the most effective model- tuned XGBoost functioned on the test dataset. It obviously indicates equal and good performance outcomes with both accuracy and F1-score 0.70. These values show that the model is not only high in the level of predicting correctly, but also consistent in its level of recall. This particular scale of operation can be particularly important in journalism and media analytics when knowing which news content will be the most popular provides direction on the editorial line and platform-focused dissemination.

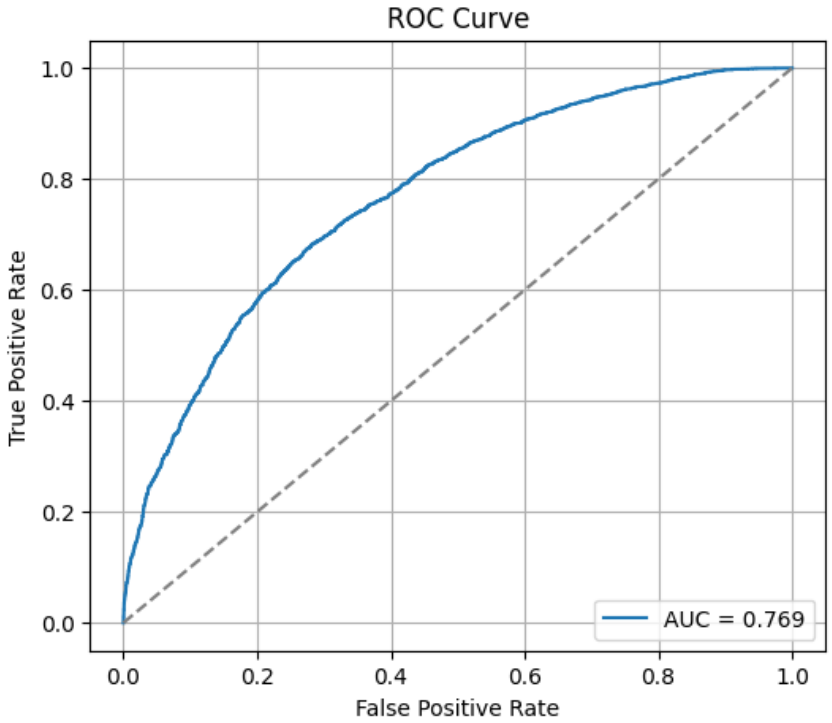

Figure 10: ROC Curve for True/False Positive Rate

The ROC (Receiver Operating Characteristic) curve depicted in figure 10 displays the competence of the tuned XGBoost model in the classification of Ionic Signatures at different threshold levels. The value of the area under the Queensland of the curve (AUC) is 0.769, which holds a significant capability of differentiation between popular and non-popular articles. AUC near 0.8 indicates that the model is much better than performing simple guessing, and provides reliable distinction between two classes being targeted. It renders this model very appropriate in the context of decision-making where it will be essential to enable limits of false positives and false negatives, e.g., the prediction of viral potential in digital content analytics. The Precision-Recall (PR) curve in figure 11 offers a very good analysis of the performance of the model, especially as regards imbalanced data sets. At higher values of recall, there is a trade-off shown by the curve that at a given recall the precision decreases. Precision values were relatively high and stable within a wide range of recall values which is a sign of robustness of the model to identify the genuinely popular articles with low number of false positives. Such visualization is particularly helpful in media analytics applications, where set-ups might differ on each platform and each application, where the cost of a misclassification of an article concerning its popularity can differ.

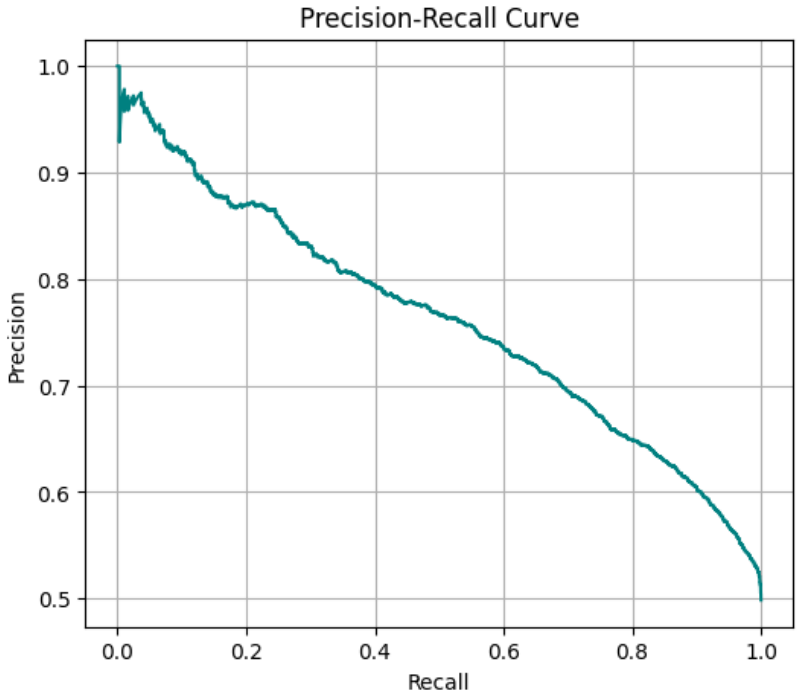

Figure 11: Precision Recall Curve

The figure 12 shows the feature importance chart is an output of the XGBoost model, which indicates the 15 best predictors that assist in the classification of online news articles in terms of popularity. Remarkably, data_channel_is_entertainment belongs to the most influential features, which confirms the increased virality potential of entertainment content. Time indicators, like is_weekend and weekday_is_saturday, also show significant value, as one can notice the patterns of higher user activity at the weekend. The article performance is represented as a multi-dimensional issue, as the feature associated with key word metrics (kw_min_min, kw_avg_avg) and those based on platform specificity (data_channel_is_tech, data_channel_is_socmed) also take place on it. The insight is relevant especially to the professionals in the field of journalism and mass communication who are seeking to maximize the content strategy and time to provide to the greatest number possible in the audience.

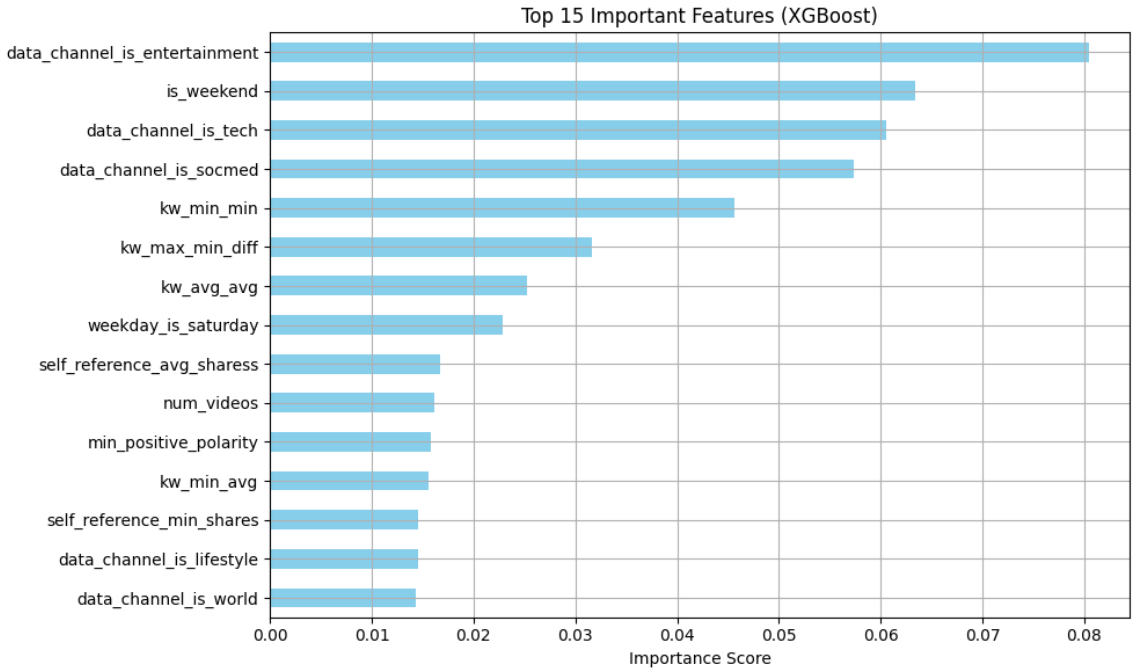

Figure 12: Feature Importance Chart

As the findings of this research work reveal, the XGBoost turned out to be the best classifier of all the models used in this research work with the best result based on the performance of accuracy of 69.79 and an F1-score of 69.69, which was obtained after hyperparameter tuning and SMOTE balancing. The other models such as Random Forest and SVM were (RBF), which proved to compete with each other with accuracy levels of 69.13 percent and 66.88 percent, respectively. Clearly, application of SMOTE significantly improved the results of all the models due to the class imbalance mitigation. The robustness of the XGBoost model was justified in the evaluation measurements like confusion matrices, ROC curve, and Precision-Recall curves, and the feature importance analysis indicated that content type (e.g., entertainment), like timing (e.g., weekend), and metadata (e.g., number of videos and keyword metrics) are major factors in shaping article virality. These findings help to demonstrate the practical importance of implementing machine learning in digital journalism to forecast and comprehend the trends in engagement of the audience.

5. Conclusion

The purpose of this research work was to predict the popularity of online news articles based on sophisticated machine learning method on journalistic relevance. In the methodology, the data preprocessing and feature engineering followed a structured pipeline consisting of data preprocessing, feature engineering, exploratory data analysis and these five supervised learning models: Logistic Regression, Random Forest, XGBoost, SVM (RBF) and MLP Classifier. They were tested to overcome class imbalance by resampling and the best class balance was achieved using SMOTE (Synthetic Minority Over-sampling Technique) that also enhanced the generalizability of the models. Model tuning as well as feature selection, especially on XGBoost, was also performed to maximize performance, in terms of precision, recall, F1-score, and ROC-AUC. Empirical findings showed that XGBoost (max_depth=10, learning_rate=0.05, n_estimators=200) recorded the highest prediction accuracy (69.8 %) and F1-score (69.7 %) compared to that of other models (e.g., Random Forest: accuracy: 69.1 %; Logistic Regression: accuracy: 65.9 %). The output ROC-AUC score was 0.769 which shows good

discriminative power of the number of popular versus the non-popular articles. The top three features that XGBoost selected were data_channel_is_entertainment, is_weekend and num_videos, which equally comply with the theories of communication in regards to content attractiveness and timing. This work, through journalism and mass communication perspective offers an understanding on how AI could play a significant role in editorial strategy. Evaluation of the insights shows that weekends, especially Saturdays, are the best days to post articles with captivate multimedia content that fits entertainment or tech category, as they perform better. Such results could assist content makers and media administrators to personalize their content to the audiences, book publication time and create evidence-based story telling strategies. To conclude, when machine learning is applied to media analytics, the predictive accuracy is not the only trait obtained by applying machine learning but also the availability of actionable editorial insights. The further development can incorporate social network activity indicators, semantic tone assessment, and immediate user opinion that would increase the level of predictions and newsroom decision-making.

References

[1] Hermida, A. (2010). From TV to Twitter: How Ambient News Became Ambient Journalism. *M/C Journal*, 13(2).

[2] Zhao, Z., Resnick, P., & Mei, Q. (2015). Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1395–1405).

[3] Tandoc, E. C., Zheng, W., & Maitra, J. (2022). Algorithmic journalism and audience engagement: An exploratory study. *Digital Journalism*, 10(5), 686–704.

[4] Fernandes, K., Vinagre, P., & Cortez, P. (2015). A proactive intelligent decision support system for predicting the popularity of online news. *Proceedings of the 17th EPIA Conference on Artificial Intelligence*, 535–546.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

[6] Bandari, R., Asur, S., & Huberman, B. A. (2012). The Pulse of News in Social Media: Forecasting Popularity. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (pp. 26–33).

[7] Tatar, A., Antoniadis, P., de Amorim, M. D., & Fdida, S. (2014). A Survey on Predicting the Popularity of Web Content. *Journal of Internet Services and Applications*, 5(1), 8.

[8] Ahmed, M., Spagna, S., Huici, F., & Niccolini, S. (2013). A Peek into the Future: Predicting the Evolution of Popularity in User Generated Content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (pp. 607–616).

[9] Keneshloo, Y., Wang, S., Ramakrishnan, N., & Reddy, C. K. (2016). Predicting the popularity of news articles. In *Proceedings of the 2016 SIAM International Conference on Data Mining* (pp. 441–449).

[10] Reis, J. C. S., Melo, P., Garimella, K., & Benevenuto, F. (2019). Can WhatsApp counter misinformation by limiting message forwarding? *Proceedings of the 10th ACM Conference on Web Science*, 149–158.

[11] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[12] He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.

[13] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, 30.

[14] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36.
[15] Zhang, K., & Ghorbani, A. A. (2020). An Overview of Online Fake News: Characterization, Detection, and Discussion. *Information Processing & Management*, 57(2), 102025.
[16] Diakopoulos, N. (2019). *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press.