Data-Driven Journalism: Predicting the Success of News Articles

Sana Ishtiaq

Department of Computing and Software, McMaster University

***Abstract*—** In the digital era, the rapid expansion of online content has transformed how news is distributed and consumed, presenting both opportunities and challenges for content publishers. The ability to forecast the popularity of online news articles is a crucial element in the dynamic landscape of internet media. It can play a vital role in shaping strategies for content distribution, enhancing reader engagement and revenue generation. This study is pivotal as it delves into analyzing the headlines of news articles to foresee their success. It methodically tackles two main objectives: categorizing articles into 'popular' or 'non-popular' segments and quantifying their popularity in terms of audience engagement and interaction. A suite of machine learning models has been applied, including Logistic Regression, Decision Trees, Light Gradient Boosting Machine (LGBM), and Linear Regression, to create the most optimal model. A notable feature of this study is the utilization of Facebook engagement metrics for elevating the predictive capabilities of these models. We delineate the preprocessing methods employed, including stemming, TF-IDF vectorization, title refinement, and imputation of missing values. Our findings demonstrate a standout performance by Logistic Regression, producing a 96.1% accuracy in classification tasks, considerably outperforming the Decision Tree and LGBM models. Furthermore, the LGBM model excels in regression tasks, predicting popularity scores with superior accuracy. These findings emphasize the effectiveness of tailored machine learning approaches in the strategic management of digital content within the competitive landscape of online news.

1. Introduction

Traditional news media is battling with significant hurdles in the rapidly changing landscape due to the high expenses and the slow pace of printing and distribution. Conversely, the exceedingly sluggish pace of the editorial process has prompted content writers and news reporters to turn to online platforms for publishing their articles and content. [1]

In the past decade, the world has witnessed an extraordinary surge in digital growth [16] and content production. The internet has turned into a vast depository, hosting a myriad of content including articles, blogs, podcasts, and videos. This unprecedented growth is motivated by the readily accessible digital platforms that enable anyone with an internet connection to generate and distribute content. Such democratization of content creation has yielded an expansive array of voices and perspectives available to the online audience. [2]

In response to these advancements and challenges, there is a dire need for data-driven insights. Publishers and creators must utilize analytical tools to scrutinize user engagement metrics, discern search trends, and shape their content strategies accordingly. Such data-driven techniques enable creators to make informed decisions on content creation, marketing, and optimization for different digital channels. Predictive analytics further help by underscoring emerging trends, topics and keywords, which allow content creators to create targeted content. [3].

Predicting news article popularity is increasingly difficult because people often lack the time to read complete articles, typically judging the article by the title. Additionally, social media, [4] particularly Facebook, massively influences an article's success. Interactions like comments, shares, and reactions on Facebook are key indicators of the impact of the content. A primary focus of this study is leveraging Facebook engagement data to optimize the prediction of news popularity, showcasing the pivotal role of social media in this context.

Machine learning has become a vital instrument across various fields. [5] However, existing ML techniques sometimes fall short in capturing the complex patterns of user interactions with online news. This study intends to aid journalists and news publishing websites in creating strategies for successful articles by generating headlines that are most impactful. It brings forth a methodology that employs traditional machine learning models to predict news articles' popularity and categorization.

By focusing on article classification and popularity score prediction, we provide a systematic approach to better understand the dynamics of online news. In this study, we examine how the appeal of news articles can

be deduced through a dual-focused method: firstly, by categorizing articles into groups deemed popular or not, and secondly, by assessing their popularity through the level of Facebook engagement, which serves as an indicator of an article's success.

The rest of this paper is structured as below: Section 2 discusses the existing techniques and methods, Section 3 discusses the detailed methodology and proposed solution, Section 4 explains the applications of this study, Section 5 represents the experiments and results and lastly, in Section 6, we discuss the conclusions.

2. Related Work

A multitude of studies have been conducted to understand the factors influencing online article popularity, focusing on user comments, social media shares, subscribers, and other metrics to predict article success. This overview highlights some key research in this field, presenting various machine learning methods that have been employed to conduct them. Our study contrasts with these by uniquely using title of the articles to forecast popularity, thereby offering new insights into the field.

Tsai and Wu [6] introduce a study on predicting internet news popularity using a UCI dataset from Mashable News, emphasizing the shift from traditional to online media. Machine learning algorithms, including Random Forest, LightGBM, XGBoost, and One-Class SVM, are applied, with One-Class SVM achieving 88% accuracy. The study contributes by modifying classification boundaries, addressing imbalanced data, and combining Autoencoder and One-Class SVM for optimal predictions and anomaly detection. Preprocessing involves normalization and principal component analysis, adjusting the classification boundary from 1400 to 50,000. Limitations include a single news platform focus, overlooking post-dissemination popularity growth, and potential feature shortcomings.

Additionally, Szabo and Huberman [7] propose a method to predict long-term popularity of online content by modeling early user access on platforms like Youtube and Digg. They find that measuring access within the first two hours accurately forecasts Digg story popularity 30 days ahead, while Youtube video popularity needs a 10-day observation. The study reveals differing time scales due to content consumption patterns, with Digg stories becoming outdated quickly and Youtube videos retaining popularity. A strong linear correlation is demonstrated between logarithmically transformed early and later popularity. Three prediction models are presented, emphasizing the importance of using relative error measures in community portals.

On the other hand, Jääskeläinen, Taimela and Heiskanen [1] propose using predictive analytics to optimize editorial decision-making in news production, employing a 'constructive approach' that combines strategic management and system dynamics in a case study exploration. They deploy neural networks for language analysis to predict news story success on digital channels, meeting the 'weak market test' criteria. While ongoing work is acknowledged, the study showcases a rare collaboration between research and the media industry, emphasizing potential applicability and impact.

Similarly, Akyol and Şen [8] propose a study using supervised learning to predict news popularity in social media, employing twelve datasets across Economic, Microsoft, Obama, and Palestine categories. They utilize Gradient Boosted Trees, Multi-Layer Perceptron, and Random Forest algorithms, The comprehensive dataset includes news items and popularity on platforms like Facebook and Google+, featuring 147 attributes over a two-day period. While the study provides valuable insights, limitations and detailed dataset characteristics are not explicitly discussed.

Subsequently, Deshpande [9] proposes a study to predict online news popularity using machine learning. The UCI machine learning repository dataset is employed, covering news across various categories. Linear Discriminant Analysis (LDA) is used for dimensionality reduction, followed by the implementation of three learning algorithms: AdaBoost, LPBoost, and Random Forest. The emphasis is on enhancing prediction performance, specifically for the "number of shares" variable. AdaBoost proves to be the best model, achieving. Moreover, Tatar *et al* [10] use people's comments to predict the article success. They employed three models using a dataset based on the four years from 20minutes.fr French online news channel. The main aim of this study is to improve the ranking rather than predicting precise attention levels. Similarly, in another of study of Tatar et al, [11], they present a ranking problem with importance of online advertisement and delivery of appropriate content. they utilized two datasets based on the French and Dutch news websites. Their results show that the significance of accurately prioritizing news articles is more important than accurately predicting levels of attention.

Table 1 shows a summary of the related work discussed in this section in a tabular form.

Table 1 Overview of Related Work

Ref	Technique(s)	Dataset	Limitation	Best Result
[6]	RF, LightGBM, XGBoost, and SVM,	UCI	A single news platform focus, overlooking post-dissemination popularity growth, and potential feature shortcomings.	Accuracy: 88%
[7]	Linear regression, constant scaling model, growth profile model	Digg, YouTube	Different sections within Web 2.0 portals and examining voting history in scenarios with a small number of users,	Relative Square Error: 30
[1]	RNN	Interviews from experts	Ongoing work and some confidentiality constraints due to real business sensitivity	Accuracy: 70%
[8]	Gradient Boosted Trees, MLP, and RF	Economic, Microsoft, Obama, and Palestine	Lack of exploration into the generalizability of the models across diverse news categories and platforms	R^2: 0.992
[9]	AdaBoost, LPBoost, and Random Forest	UCI	Lack of interpretability might hinder the practical insights gained from the models. Top of Form	69% accuracy and a 73% F-measure
[10]	Simple Linear Regression	20minutes	Only regression-based model is used. Traditional supervised machine learning models remain unexplored.	MAP: 95%
[11]	Linear, MART, AdaRank	20minutes, Telegraaf	The effectiveness of the proposed linear log popularity predicting model could possibly be impacted by the distinct features of the French and Dutch news websites utilized in the research.	MAP@100: 0.61

3. Proposed Solution

In this section, a comprehensive framework of proposed solution to predict the level of interaction and popularity of online news is discussed, as shown in Figure 1. This study is comprised of three phases; each phase focuses on an important aspect of the proposed solution. First, the characteristics of the dataset will be discussed, Then, the data preprocessing techniques and steps are discussed. These steps include stemming, TF-IDF vectorization, handling missing values, and title cleaning. Lastly, the main part of the proposed solution is detailed. It involves article engagement using a variety of machine learning methods, such as Decision Trees, Logistic Regression, the Light Gradient Boosting Machine (LGBM), and Linear Regression. We also discuss hyperparameter tuning of our models to provide as a holistic understanding of the methodology.

3.1 Dataset

The dataset utilized int his study is taken from Kaggle. [12]. The main reason to use this dataset is to employ text regression and text classification models to predict the engagement score and identifying the top articles based on their titles. It is based on the news articles from September 3, 2019, to October 4, 2019. This range enable us a thorough understanding of the patterns of online news consumption during that period.

3.1.1. Fields in Dataset

There are various fields in the dataset such as, **source_id** uniquely to identify publishers, adopting the lowercase source name with underscores. The author field indicates the article's provider, with the source_name acting as a substitute when author information is private. The **published_at** column precisely timestamps article publication in UTC format. The **content** column encapsulates unformatted article content, truncated for brevity. Finally, the **top_article** field indicates whether an article is listed as a top article on the publisher's website, distinguishing between values 1 and 0 based on its popularity status. The details of the fields are listed in Table 2.

3.1.2. Facebook Engagement Data

To enhance the dataset's richness, we augment it with Facebook engagement data, encompassing essential metrics such as the number of shares, comments, and reactions. This augmentation provides a nuanced perspective on the audience's interaction with the articles, contributing to a more robust analysis.

3.1.3. Authorship Insights

Exploring the dataset reveals a correlation between author names and publication sources. Noteworthy publications include Reuters, BBC News, and Irish Times, highlighting their prominence in the dataset as shown in Figure 2. Similarly, the top three authors identified are The Associated Press, Reuters, and CBS

News, shedding light on the influential contributors in the online news landscape

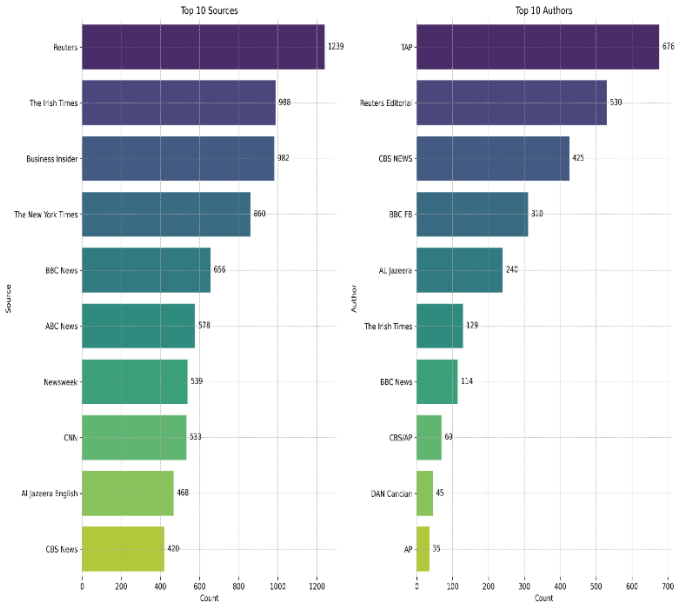

Figure 2. Top Ten Sources and Authors

3.1.4 Consumer Engagement Patterns

After conducting an analysis of consumer engagement throughout the month, Figure 3 has revealed intriguing patterns. Significantly, the considerable consumer engagement observed on the 1st of October suggests a reaction to a noteworthy event or headline that captured attention. Further examination reveals that additional peaks of engagement were observed on the 3rd, 7th, and 12th of September.

Table 2 Fields in the Dataset

Column Name	Description
source_id	The publisher's unique identification is shown as the lowercase source name with spaces replaced by underscores.
source_name	Publisher's name.
author	Author of the article. In situations when publishers do not reveal author information, source_name might be used as a stand-in..
title	Headline of the article.
description	Short article description, often viewable on the publisher's website in pop-ups or suggestion boxes. This field has been consolidated into the "content" column.
url	URL (Uniform Resource Locator) for the article located on the publisher's website.
url_to_image	URL to the main image associated with the article.
published_at	Exact date and time of article publication, presented in UTC (+000) time format.
content	Unformatted content of the article, truncated to 260 characters.
top_article	Shows if the publisher's website lists the article as one of its top articles. If the article is in the popular/top articles group, it has a value of 1, and if not, it has a value of 0.

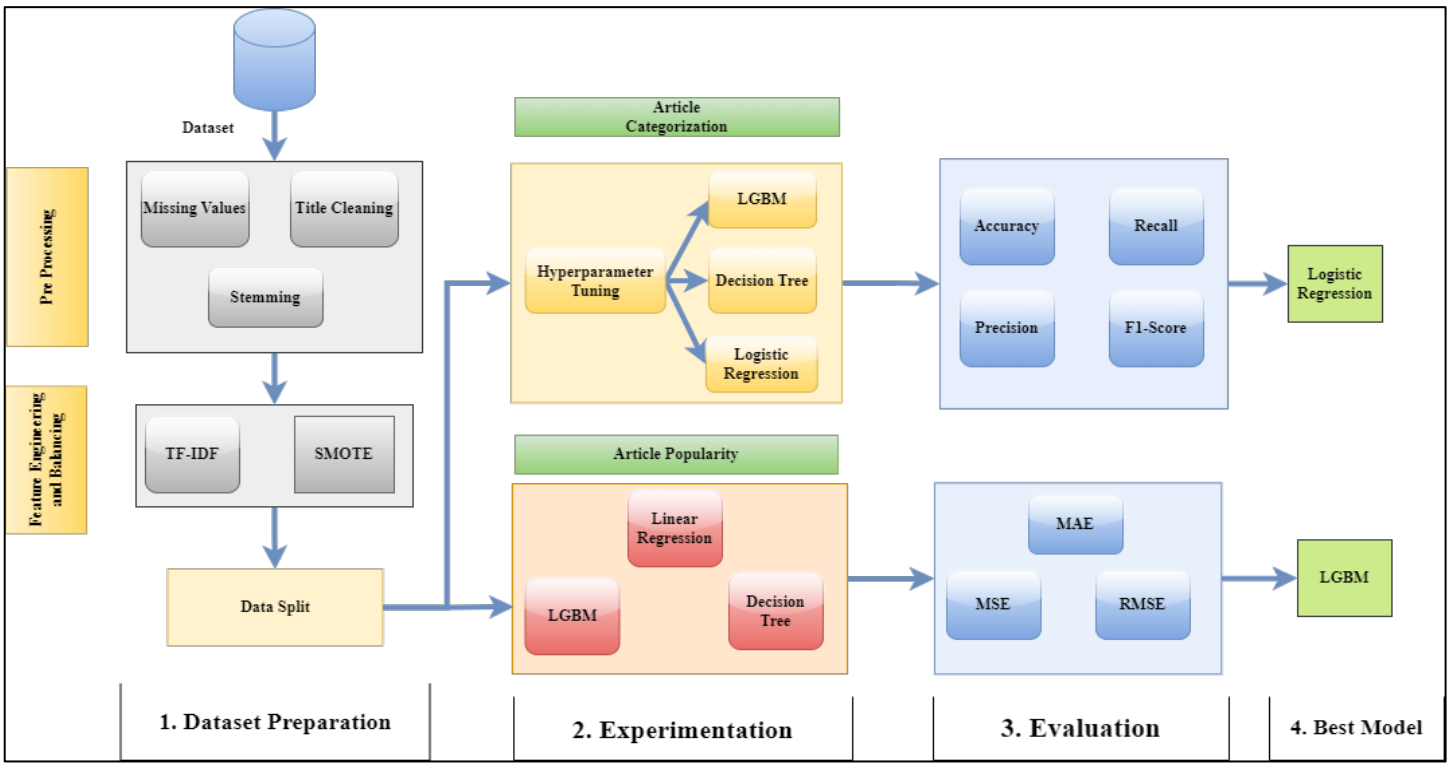

Figure 1. Framework of Proposed Solution

3.2 Preprocessing

Preprocessing methods are vital in the preparation of raw data for analysis. They ensure data integrity and enhance the performance of machine learning models. Various preprocessing techniques were used to clean and improve the quality of the dataset. These preprocessing steps are as follows.

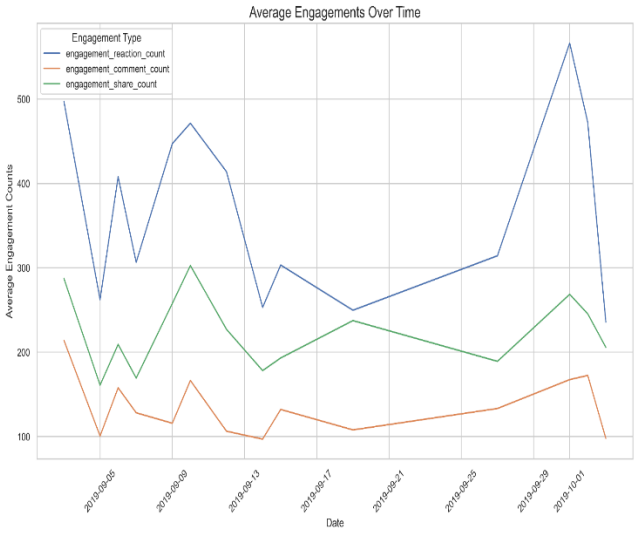

Figure 3. Consumer Engagement Patterns

3.2.1 Handling Missing Values

There were few missing values in the columns of our dataset. These missing values are found in the "content" and "author" columns. The relative strategies have been established to fill these values.

3.2.2 Title Cleaning

The preprocessing techniques are also applied on the textual data. Special characters and phrases such as punctuation marks, hyperlinks and numerical values are excluded from the titles.

3.2.3 Stemming

The words are tracked back to their root by employing stemming. this step is essential to enhance and streamline the data to increase the predictive ability of the models.

3.2.4 TF-IDF Vectorizer

To convert the textual data into a compatible format is very importance in machine learning. We utilize the TF-IDF Vectorizer to achieve this task. In this process, the preprocessed titles are converted into a feature vectors. The TF-IDF Vectorizer considers various factors, such as n-gram range, data type, sublinear term frequency scaling, and inverse document frequency smoothing. This TF-IDF matrix will be used as input for our machine learning models, enabling them to effectively analyze and learn from the intricate patterns present in the textual material.

3.2.5 Oversampling with SMOTE

We employed Synthetic Minority Over-sampling Technique (SMOTE) addressed the class imbalance issue in "top_article" data. Where only 12% of the data is labeled as 1. We replace the missing values fields with '0' after encoding them. We then confirm the binary label distribution, resulting in a column containing only the values '0' and '1'. This initial step is crucial in preparing the dataset for oversampling, ensuring that the artificial samples generated by SMOTE accurately represent both classes and contribute to a more balanced dataset, ultimately improving the performance of our model.

3.2.6 Correlation Analysis

A correlation based heatmap is produced in Figure 4 in order to highlight the interconnections between the engagement metrics and classification of the article. A strong positive correlation is observed among reactions, comments, and shares, indicating that consumers who express their liking through reactions are more inclined to share and comment on the post. Notably, no significant correlation is found between these engagement metrics and the identification of top articles, suggesting that the selection of top articles is primarily influenced by their inherent quality rather than engagement metrics.

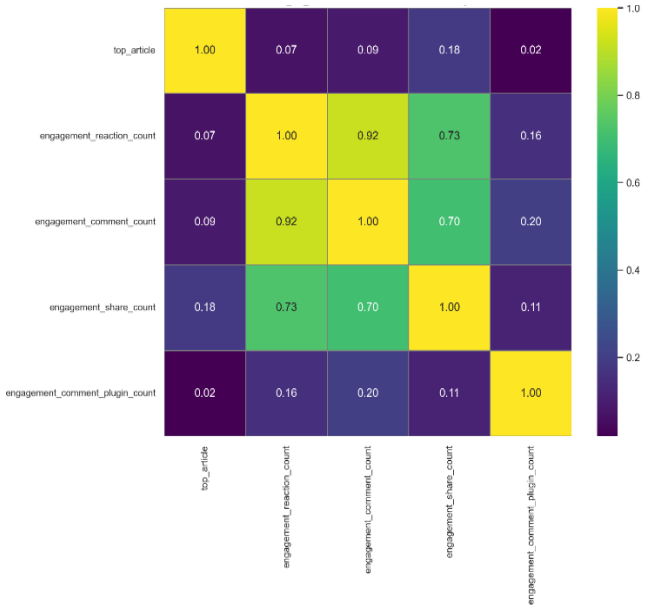

Figure 4. Correlation of engagement metrics with top article

3.2.7 Data Split

To assess the model's efficacy, the dataset was divided into training and testing sets. There were training and testing sets for the feature set (**X**) and the target variable (**Y**). To guarantee a proportionate representation of each

class in the training and testing sets, the split was stratified according to the goal variable (**Y**). The parameter **test_size** was set to 0.2, designating 20% of the data for testing, and a random seed of 40 was applied for reproducibility.

3.3 Proposed Model

Our analysis involves two distinct predictive measures—article category and article popularity score. The classification is characterized as popular or not popular using "top_Article" column as the main variable. In contrast, the regression task aims to predict the popularity score, considering the engagement data as the independent variable and the popularity score as the target variable. Three types of machine learning algorithms both classifiers (Decision Tree, Logistic Regression and LGBM) and regression models(Decision Tree, Linear Regression and LGBM) are employed.

3.3.1 Hyperparameter Tuning

Systematic approach to hyperparameter tuning was applied using RandomizedSearchCV to enhance the performance of our machine learning models. For the LGBMClassifier, the exploration involved key parameters such as learning rate, maximum depth, number of leaves, number of estimators, and maximum bin size. Randomized sampling within specified ranges was conducted to search for optimal configurations, measured against the 'f1_macro' scoring metric. A similar strategy was applied to DecisionTreeClassifier, targeting parameters including criterion, maximum depth, minimum samples split, and minimum samples leaf. In the LogisticRegression model, the tuning process involved adjusting the penalty type, regularization strength (C), solver type, and maximum iteration. To conduct a comprehensive evaluation, the 'f1_macro' score metric was used as a guide for the search.

3.3.2 Classification Models

Three classification models, namely Decision Tree, Logistic Regression, and Light Gradient Boosting Machine (LGBM), are utilized to achieve article classification. The objective variable in these models is the "top_Article" column, where '1' represents a popular article and '0' signifies a non-popular one. The models are trained using engagement data to predict the article category.

- **Decision Tree Classifier:** Selected for decision-making processes because to its interpretability and transparency[13]. The decision to use a decision tree classifier for article classification was based on its interpretability and transparency. These qualities make it suitable for deriving insights from engagement data. The classifier was optimized using specific hyperparameters, including an unrestricted maximum depth, the entropy criterion for impurity measurement, and a requirement for a minimum of two samples to split an internal node and a minimum of one sample to be a leaf node. These hyperparameters were selected using RandomizedSearchCV.
- **Logistic Regression Classifier:** Selected due to its ease of interpretation and simplicity; perfect for binary classification problems such as differentiating between articles that are popular and those that are not [14]. Optimized by RandomizedSearchCV, the classifier's selected hyperparameters are: L2 penalty for regularization, a regularization strength (C) of 9.574, optimizing with the 'liblinear' solver, limiting the number of iterations to 175, and automatically determining the multi-class classification strategy.
- **Light Gradient Boosting Machine(LGBM) Classifer:** Used for huge datasets and complicated tasks due to its scalability and efficiency, it is the best option for engaging-based article classification[15]. The classifier's selected hyperparameters were refined using RandomizedSearchCV. These included a maximum depth of 17, a learning rate of 0.2561, 45 leaves, 97 estimators, and a maximum bin size of 297.

3.3.3 Regression Models:

For predicting the article popularity score, we implement three regression models—Decision Tree Regressor, Linear Regression, and LGBM Regressor. In this scenario, the engagement data serves as the independent variable, and the popularity score becomes the target variable. It is pertinent to mention that we have employed linear regression instead of logistic regression for predicting popularity score because it is more suitable for regression problems as compared to logistic regression. By providing a numerical approximation of the popularity score, these regression models provide highlight the quantitative side of article engagement.

- Decision Tree Regressor: Used because of its transparency and capacity to handle both categorical and numerical features[13], which makes it easier to interpret when forecasting the popularity score. With its default settings, the Decision Tree Regressor enhances the ensemble by forecasting engagement scores.
- **Linear Regression:** Selected due to its ease of use and compatibility with numerical results[14]. With its default settings, the Linear Regression Regressor increases the performance of our model by offering continuous predictions for engagement scores.

- **Light Gradient Boosting Machine(LGBM) Regressor:** Selected for its efficiency and scalability, offering robust performance in predicting numerical values[15].. The LGBM regressor, configured with a learning rate of 0.01, maximum depth of 20, number of leaves set at 50, and 150 estimators, complements the ensemble by providing predictions for engagement scores.

3.4 Evaluation Metrics

We have employed two types of evaluation metrics. These metrics are used to calculate the performance of the classifiers and prediction ability of the repressor's of popularity scores. The details of these metrics are as follows:

3.4.1 Performance Metrics for Classification:

To assess the performance of our classification models, we employ standard metrics including:

- **Accuracy (Acc):** The ratio of correctly classified samples to total samples:

$$\textbf{Accuracy (Acc)}: \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad \text{(i)}$$

- **Precision (Prec):** Precision is named the division of the examples which are actually positive among all the examples that we predicted positive:

$$\textbf{Precision (Prec)}: \frac{(TP)}{(TP+FP)} \quad \text{(ii)}$$

- **Recall:** ratio between the number of Positive samples correctly classified as Positive to the total number of Positive samples

$$\textbf{Recall}: \frac{(TP)}{(TP+FN)} \quad \text{(iii)}$$

- **F1 score:** F1 score is defined as the harmonic mean between precision and sensitivity:

$$\textbf{F1 Score}: 2 \times \frac{Precision \times Recall}{(Precision+recall)} \quad \text{(iv)}$$

3.4.2 Metrics for Popularity Score Prediction:

For assessing the effectiveness of our regression models in predicting popularity scores, we employ:

- **Mean Absolute Error (MAE):** The average absolute differences between predicted and actual values, providing a straightforward measure of prediction accuracy.

$$MAE = \sum_{i=1}^{n} \left| \frac{Y_{actual} - Y_{predicted}}{n} \right| \quad \text{(v)}$$

- **Mean Squared Error (MSE):** The average of the squared differences between predicted and actual values, amplifying the impact of larger errors.

$$MSE = \sum_{i=1}^{n} \left(\frac{Y_{actual} - Y_{predicted}}{n} \right)^2 \quad \text{(vi)}$$

- **Root Mean Squared Error (RMSE):** The square root of MSE, offering a metric in the same unit as the target variable, providing a more interpretable measure of prediction error.

$$RMSE = \sqrt{MSE} \quad \text{(vii)}$$

4. Applications

The proposed solution could be helpful across many domains because to its capability to predict the engagement and popularity of online news articles. Some notable application areas include:

4.1 Content Optimization in Media Outlets

This will have important significance on the way content is optimized within media when it comes to online news articles and their ability to predict engagement and popularity. The predictive model enables editors to prioritize on those topics and formats that relate well with the target audience. With this data-based approach, the media houses can design a content production strategy that will attract readers easily.

4.2 Digital Marketing and Advertisement

The engagement prediction model is an important tool in the digital marketing and advertising landscape. The model's insights can also be used by marketers in refining content strategies so that the content which will enthrall the targeted audience is selected. It involves a process of re-orientation, which leads more precise direct marketing campaigns and high returns.

4.3 User Engagement Analytics

Engagement prediction model is based on its ability to integrate technology into social networks and online user interface for better customer experience. Online platforms have now an opportunity to improve the user experience by providing more personalized content recommendation for consumers, based on their own views regarding which news are "hot".

4.4 Decision Support for News Editors

News editors can rely on the model's predictive abilities concerning article success rate as decision support. However, by being ahead of the curve, editors have the confidence to place an article where they anticipate it would best interest and attract readers. They are well positioned to formulate relevant headings, which can be used in promoting various content strategies.

4.5 Social Media Management

The engagement prediction model improves the capability of social media managers to postnews stories. It's a way for companies to understand which pieces are more likely to elicit response to their social media marketing strategy, thus improving performance and effectiveness of content delivery that is often shared across multiple online sources.

4.6 Personalized Content Recommendations

The proposed solution is also significant in the field of creating personalized content for delivering. Platforms also can provide individualized recommendations by incorporating engagement prediction models in content recommendation systems. This entails using projected popularity of news stories to provide information tailored to the interests of each user, consequently improving the quality of service, and boosting the number of active users in online media.

4.7 Adaptive Social Media Strategies

These insights could greatly aid social media strategies, hence. The engagement prediction model can be used by social media management teams to adapt and optimise their content sharing strategies. Predicting which posts are likely to attract more engagement on social media is critical to ensure that strategically selected articles reach their intended readers, optimizing the campaign's effectiveness.

5. Experiments and Results

In the experiments and results section, we rigorously examine the effectiveness of our proposed machine learning models in predicting online news engagement. The study includes two important experiments: the first focuses on article classification, evaluating the models' ability to reliably identify articles as popular or not, and the second on predicting article popularity scores.

5.1 Experiment 1: Article Categorization

The observed variations in model performance, particularly in the context of overfitting, provide significant perspectives into the applicability of each technique for the given dataset as shown in Figure 5. Overfitting is evident in the Decision Tree (DT), as demonstrated by a significant gap between the training and cross-validation lines. The difference indicates that the model became too complicated and has over fitted to the training data, limiting its generalizability to new, unidentified data. In contrast, while the LGBM lines are currently separated, they have the potential to converge given a larger dataset for training. The closeness of the Logistic Regression lines indicates reasonably good training, demonstrating the model's capacity to generalize effectively to testing data. These findings highlight the importance of dataset sufficiency, emphasizing the goal of reducing the gap between training and cross-validation lines for best model performance.

The first experiment involved testing a three machine learning models, namely Decision Tree, LGBM (Light Gradient Boosting Machine), and Logistic Regression, using critical performance metrics such as accuracy, precision, recall, and F1-score. The Decision Tree model demonstrated good precision (88.4%), recall (86%), F1-score (87.1%) and accuracy (87%). Thus, this indicates that the model can separate the documents into specific topics shown in Figure 6.

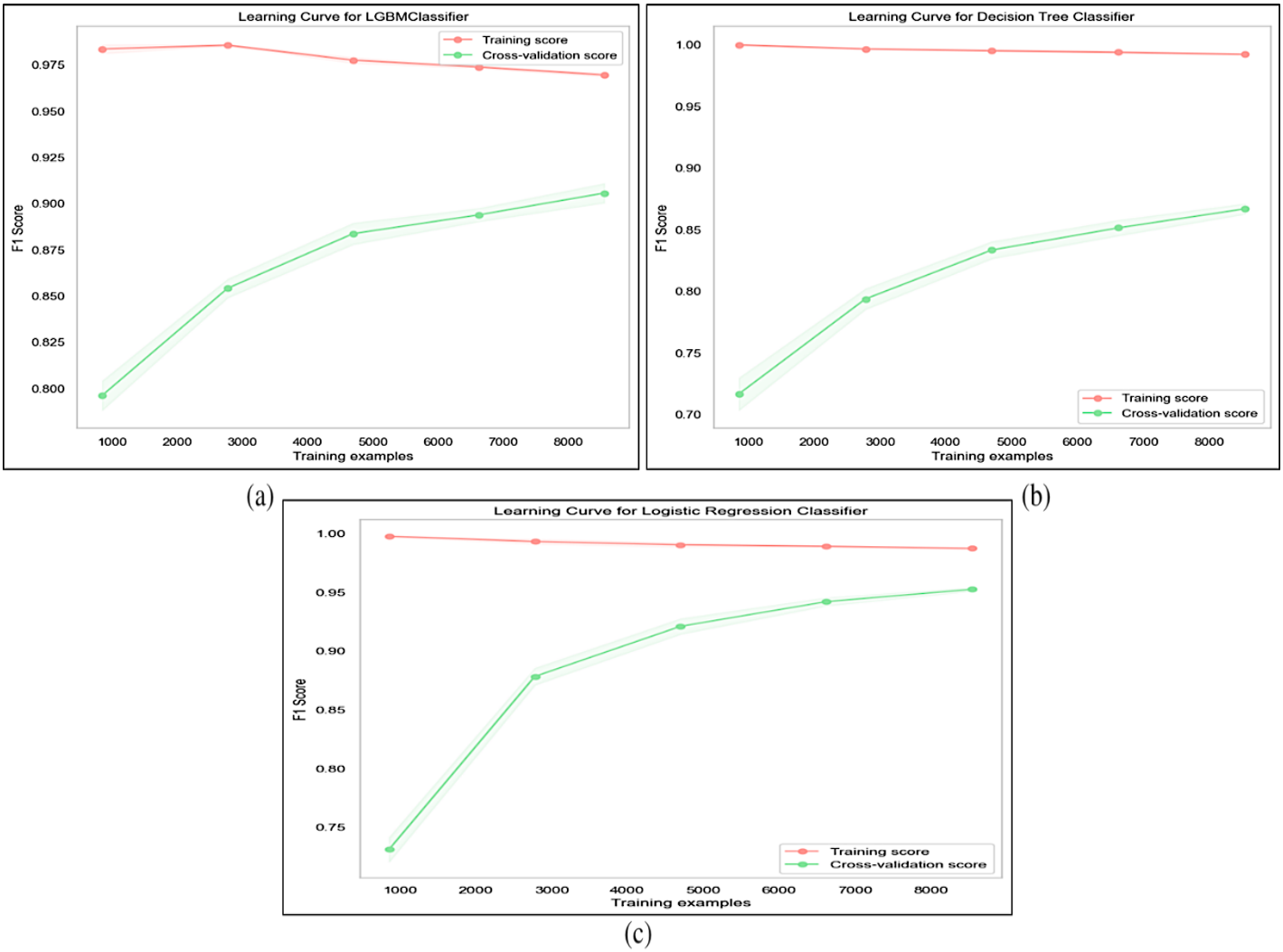

Figure 5. Learning curves of the classifiers: (a): LGBM, (b) Decision Tree, (c) Logistic Regression

The LGBM model, known for its efficacy and scalability, outperformed the Decision Tree with an accuracy of 91.3%. It consistently delivers excellent results, with precision, recall, and F1-score metrics of 95.4%, 88%, and 91.6%, respectively. Since the LGBM model employs a gradient boosting framework with an ensemble of decision trees, its robust performance indicates that it can accurately categorize news items.

The logistic regression model was better than the decision tree and LGBM for article classification. The model exhibited very high levels of accuracy with a score of 96.1% and impressive precision, recall, and F1-score of 93.4%, 98.6% and 96% respectively. The fact that logistic regression managed to identify such a remarkable outcome shows its ability to distinguish popular and less popular news stories.

The outcome shows an overall progress towards performance of all three models considered. While Decision Tree is a good benchmark model, LGBM has higher precision and repeatability than Decision Tree. Logistic regression is the best article classification model because of its high predictability and precision, thus leading to its application in real time settings.

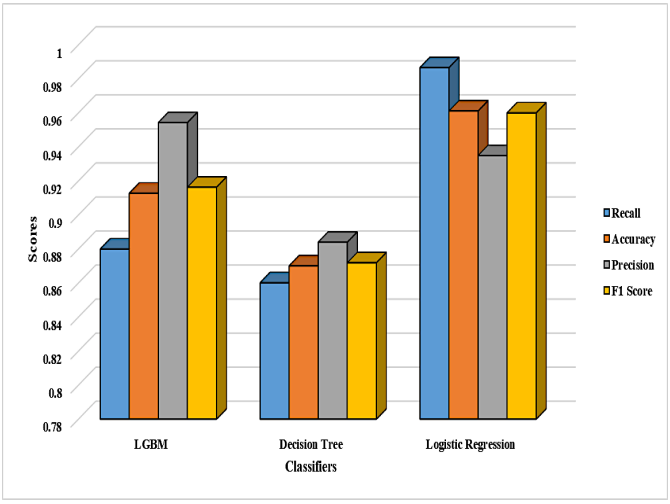

Figure 6. Performance Comparison of Classifiers

We therefore used the logistic regression mode to establish the significance of each attribute whereby, as expected, logistic regression did perform better than other models. Several factors determine the model's coefficients and increase the chances that an article will be popular. For instance, there are various words like "Trump" or "Trudeau". However, articles with negative features have less likelihood of getting popular. Examples of these qualities include words such as

"journal" and "fashion". Some selected attributes for Table 3 as follow:

Table 3. Important Features

Feature Name	Coefficient significance value
Trump	6.3
Kill	4.5
Brexit	3.2
Saudi	2.9
PM Johnson	2.8
Trudeau	2.7

5.2 Experiment 2: Article Popularity Score

The second experiment utilized crucial evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to assess three different machine learning models: Decision Tree, Light Gradient Boosting Machine (LGBM), and Logistic Regression. It is pertinent to mention that the mean in the popularity column was used as a threshold for the popularity score.

The Decision Tree showed good classification results. Nevertheless, it did not predict popularity scores correctly and committed numerous mistakes in this regard. With a large MAE of 2.83, MSE of 13.26, and a huge RSM of 3.64, the model's precision was faulty for that specific activity.

However, the LGBM model performed better compared to others when it came to the prediction of the article popularity scores. This showed higher accuracy and precision with a lower MAE of 2.25, MSE of 6.94 and RMSE of 2.63. LGBM model's gradient boosting design helped in improving the actual prediction through capturing the most minute fluctuation in popularity score. Therefore, this is where the model proves its worth in improving predictions while relating with online news.

Likewise, it has been noted that the Logistic Regression performs well within classification of articles as well predicting their popularity scores. In fact, it had competitive indexes comprised by, among others, Mean Absolute Error = 2.35; Mean Squared Error = 8.66; and Root Mean Squared Error = 2.94, which corresponded to it ability for popularity

Further analysis revealed that in a number of sub-spaces, i.e., various areas of the prediction tasks, model accuracy differed significantly. Clearly, figure 4 illustrated LGBM had higher popularity score compared to Decision Tree or Logistic Regression. The flexibility and appropriateness of LBGMT towards predicting different aspects of online news engagement is brought into light in this paper.

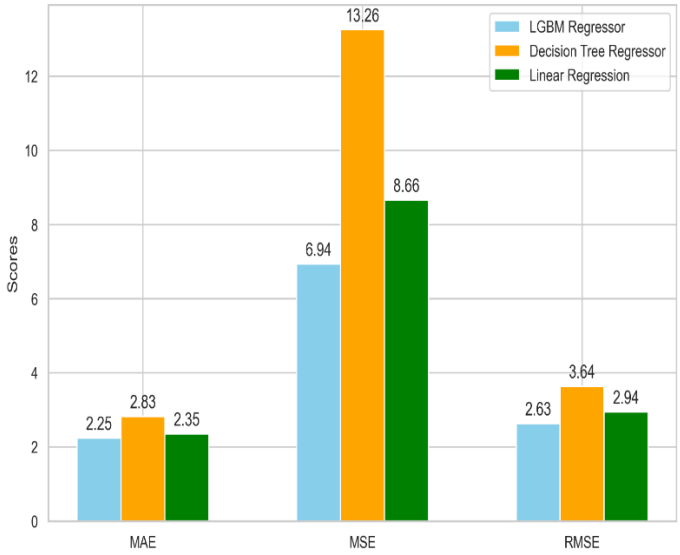

Figure 7. Performance Comparison of Regressors

6. Conclusion

In this study, we employed machine learning methods to investigate the prediction of the engagement and popularity levels of online news articles. The research implemented various models and focused on two primary experiments: predicting article popularity scores as well as article categorization applying Logistic Regression, Decision Tree, LGBM, and linear regression. Logistics Regression was found to be efficient among the tests carried out, being able to distinguish between popular or non-popular news on an average accuracy level of 96.1%. However, LGBM performed well in Article Popularity Score Prediction, with significantly lower error rates.

To enhance the models' resilience and applicability across a wider range of articles, this study will expand its scope in the future by incorporating additional diverse data sources. Machine learning models can better respond to the changing dynamics of online news consumption by recognizing a broader range of content. Furthermore, exploring sophisticated text analysis approaches, such as using pre-trained models like BERT, has potential for understanding complex textual nuances, enhancing the precision of engagement prediction even further. Furthermore, to keep up with the ever-changing nature of online material consumption, future study might focus on evaluating success in real-time. Creating frameworks for continuous learning from audience data allows for the dynamic improvement of content strategies, enabling a more responsive and adaptable approach to content generation and dissemination.

References

[1] A. Jääskeläinen, E. Taimela, and T. Heiskanen, "Predicting the success of news: Using an ML-based language model in predicting the performance of news articles before publishing," *ACM Int. Conf. Proceeding Ser.*, pp. 27–36, 2020, doi: 10.1145/3377290.3377299.

[2] M. Gayberi and S. G. Oguducu, "Popularity prediction of posts in social networks based on user, post and image features," *11th Int. Conf. Manag. Digit. Ecosyst. MEDES 2019*, pp. 9–15, 2019, doi: 10.1145/3297662.3365812.

[3] D. Liao, J. Xu, G. Li, W. Huang, W. Liu, and J. Li, "Popularity prediction on online articles with deep fusion of temporal process and content features," *33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019*, pp. 200–207, 2019, doi: 10.1609/aaai.v33i01.3301200.

[4] K. O. Yakunin, S. B. Murzakhmetov, R. R. Musabayev, and R. I. Mukhamediyev, "News Popularity Prediction Using Topic Modelling," *SIST 2021 - 2021 IEEE Int. Conf. Smart Inf. Syst. Technol.*, no. April, 2021, doi: 10.1109/SIST50301.2021.9465884.

[5] C. Xiao, C. Liu, Y. Ma, Z. Li, and X. Luo, "Time sensitivity-based popularity prediction for online promotion on Twitter," *Inf. Sci. (Ny).*, vol. 525, no. July, pp. 82–92, 2020, doi: 10.1016/j.ins.2020.03.056.

[6] M. J. Tsai and Y. Q. Wu, "Predicting online news popularity based on machine learning," *Comput. Electr. Eng.*, vol. 102, no. February, p. 108198, 2022, doi: 10.1016/j.compeleceng.2022.108198.

[7] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, 2010, doi: 10.1145/1787234.1787254.

[8] K. Akyol and B. Şen, "Modeling and predicting of news popularity in social media sources," *Comput. Mater. Contin.*, vol. 61, no. 1, pp. 69–80, 2019, doi: 10.32604/cmc.2019.08143.

[9] D. Deshpande, "Prediction Evaluation of Online News Popularity Using Machine Intelligence," *2017 Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2017*, pp. 1–6, 2017, doi: 10.1109/ICCUBEA.2017.8463790.

[10] A. Tatar, P. Antoniadis, M. D. De Amorim, and S. Fdida, "Ranking news articles based on popularity prediction," *Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012*, no. February 2007, pp. 106–110, 2012, doi: 10.1109/ASONAM.2012.28.

[11] A. Tatar, P. Antoniadis, M. D. de Amorim, and S. Fdida, "From popularity prediction to ranking online news," *Soc. Netw. Anal. Min.*, vol. 4, no. 1, pp. 1–12, 2014, doi: 10.1007/s13278-014-0174-8.

[12] SLETER, "Internet news data with readers engagement," *Kaggle*, 2020. https://www.kaggle.com/datasets/szymonjanowski/internet-articles-data-with-users-engagement/data (accessed Dec. 01, 2023).

[13] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *J. Chemom. A J. Chemom. Soc.*, vol. 18, no. 6, pp. 275–285, 2004.

[14] X. Song, X. Liu, F. Liu, and C. Wang, "Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis," *Int. J. Med. Inform.*, vol. 151, p. 104484, 2021, doi: 10.1016/j.ijmedinf.2021.104484.

[15] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, and W. Zeng, "Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data," *Agric. Water Manag.*, vol. 225, no. August, p. 105758, 2019, doi: 10.1016/j.agwat.2019.105758.

[16] International Telecommunication Union. (2021). Measuring digital development: Facts and figures 2021. https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf