Online News Popularity Prediction

Aditya Jain - 2021511
Shrey - 2021562
Vasu Kapoor - 2021573
Vinayak Sharma - 2021574

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

- UCI Online news popularity dataset is a valuable resource in the field of journalism.

- Online news articles that have a high popularity score are more likely to be shared and commented on by the readers.

- In an era dominated by the internet and social media, understanding what makes the online news article popular and shareable is very important for editors, media networks, and content creators.

Motivation

- **Factors** affecting the popularity of Online News Article
  - Topic
  - Headline of the article.
  - Images
  - Text Quality

- The task that this dataset aims to address is the prediction of news articles popularity by looking at the number of shares it receives on social media.

Literature Review

- "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News" by Fernandez, Vinagre and Cortez.
- NLP algorithms and IDSS.
- Proposed a Binary Classification problem.
- Search methods on Random Forests, AdaBoost and KNN for best performance

Load a YouTube video on the big screen, and there's a chance the quality could appear somewhere between an 8-bit video game and a fuzzy LEGO art project. There's not a whole lot YouTube can do about that; it's the one truly democratic, worldwide video network. Quality of uploads is bound to be all over the map.

No Verified Videos, Some Nudging

So how to overcome our wariness of using YouTube on the TV? During our conversations, I suggested the service start verifying accounts, Twitter-style -- you get a tick next to your name if you consistently post videos that look great on a 42-inch screen, say. (Because as we know from experience, simply saying a video is HD when you upload it doesn't make it so.) You could also use the verification process to clamp down on one of YouTube's most terrible scourges: the vertical video.

Meanwhile, the Ascend D2 is Huawei's chance at taking on the superphone market. The Ascend D2 has a 5-inch full HD display with a resolution of 1920x1080 pixels. That gives it the same pixel density as the HTC's Droid DNA on Verizon.

Like it's bigger sibling, the Ascend D2 runs on Android 4.1 and has a quad-core 1.5GHz processor. In addition to a 13-megapixel backside illuminated camera and a 3000mAh battery, the Ascend D2 was built to be water resistent. Huawei demonstrated its ability to work under wet conditions by throwing water on the device during the press conference.

Credit:

NLP ALGORITHMS

# n_tokens_t...	# n_tokens_...	# n_unique_t...	# n_non_sto...	# n_non_sto...	# num_hrefs
12.0	219.0	0.663594466988	0.999999992308	0.815384609112	4.0
9.0	255.0	0.604743080614	0.999999993289	0.79194630341	3.0
9.0	211.0	0.575129530699	0.999999991597	0.66386554064	3.0

Literature Review

- "Online News Popularity Prediction" by Feras Namous, Ali Rodan.
- Classification problem.
- Extracted 20 best features of the dataset using fischer score.
- 11 ML algorithms.
- Random Forests - 66%, MLP - 65%

Dataset description

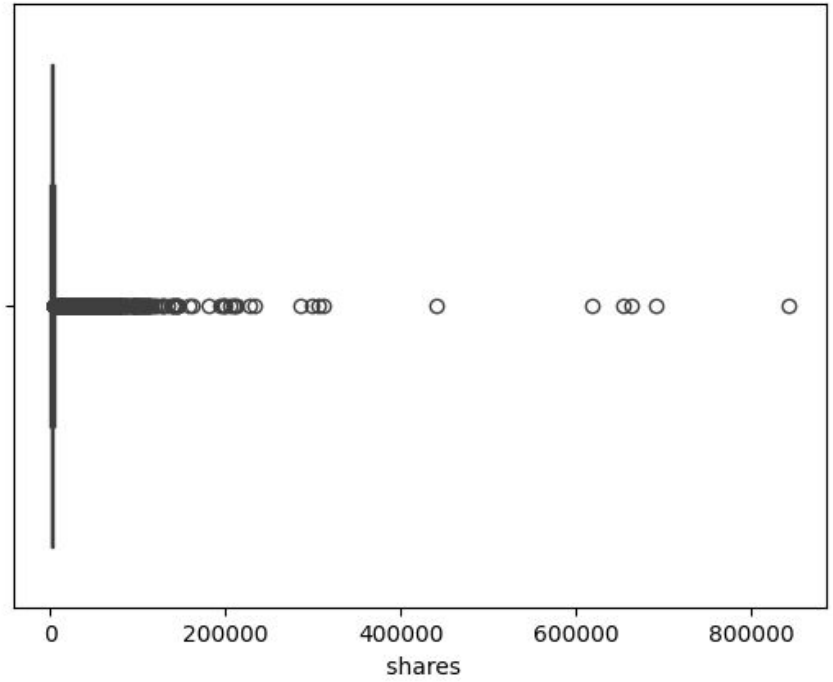
- The UCI Online News Popularity dataset underwent prior preprocessing by Fernandez et al.

- Multiple features were derived from the articles using NLP processes, encompassing aspects such as title subjectivity, Latent Dirichlet Allocation (LDA), the frequency of positive and negative words, release date, article type, data channel, image count, and more.

Dataset description

- The dataset comprises a total of 58 predictive attributes, alongside two non-predictive attributes, and one target attribute, which is denoted as "shares."

- It is important to note that the target variable is highly skewed.
  So, we turned the dataset into binary classification of popular or unpopular.



Dataset description

Collinearity:

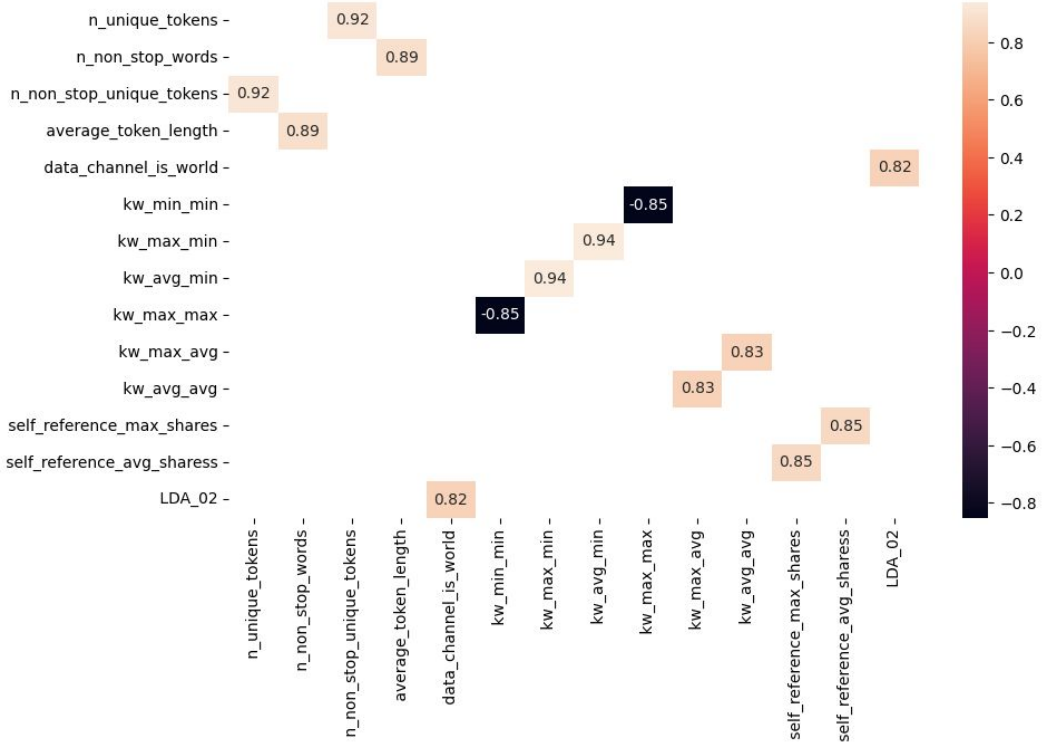
Used Correlation Heatmap Analysis Conducted.

Key Findings:

- Strong Correlations Identified:
  - Indicates Redundancy in Certain Feature Pairs.
  - Correlation Coefficient Threshold: 0.8.

Action Taken:

- Feature Removal Strategy:
  - Removed one Feature from Highly Correlated Pairs.
  - Enhances Efficiency and Prevents Redundancy.



Dataset description

Outlier detection:

Importance of Outlier Detection:

- Quality Assurance: Essential for Data Integrity.
- Algorithm Used:
  1. Local Outlier Factor (LOF) Algorithm
  2. Isolation Forest Algorithm

Challenges Faced:

- Issue Identified: Outliers Not Detected as Expected.

Dataset description

Dimensionality Reduction

- Mutual Information Analysis: Identified Variables with Limited Contribution.

Approach:

- Method Used: Principal Component Analysis

Significance:

- Reducing the features while preserving maximum information and Interpretability

Challenges Faced:

- Issue Identified: Accuracy is getting reduced

FEATURES	MUTUAL INFORMATION
self_reference_min_shares	*0.043105*
self_reference_max_shares	*0.040332*
LDA_02	*0.036253*
kw_max_avg	*0.033929*
kw_min_avg	*0.030945*
LDA_03	*0.029853*
LDA_00	*0.028202*
self_reference_avg_sharess	*0.027294*
kw_avg_avg	*0.027271*
is_weekend	*0.021205*

Methodology

- Various classification techniques are used for predictions over the dataset.
- Classification Techniques Used:
    - Logistic Regression
    - Decision Tree Classifier
    - Random Forest Classifier
    - Bagging
    - AdaBoost
    - Gradient Boosting Classifier
    - MLPClassifier
    - SVM
    - Bernoulli Naive Bayes
    - Stacking Classifier

Methodology

- **Logistic regression:** Uses Sigmoid Function to modal the probability that a particular input belongs to a class.
- **Decision Tree Classifier:** Partitions the dataset based on the input values.
- **Random Forest Classifier:** Combines 100, 300, and 500 decision trees to create robust classification model.
- **Bagging:** Training multiple instances of learning algorithm (we used logistic regression) on different subset of training data. It is used to reduce overfitting.
- **AdaBoost:** Assigns weights to misclassified instances, then it gives more emphasis to the difficult-to-classify examples. We used logistic regression as our learning algorithm.

Methodology

- **Gradient Boosting Classifier:** Combines weak learning models (generally, decision trees). We used it since it captures the complex patterns in the data.
- **MLPClassifier**(Artificial Neural Network): It is a feedforward neural network.
- **Bernoulli Naive Bayes:** Variant of Naive Bayes Algorithm, used for binary classification problems. Assumes the target variable as Bernoulli distribution.
- **Stacking Classifier:** Combines the prediction of multiple base models to improve overall predictive performance.

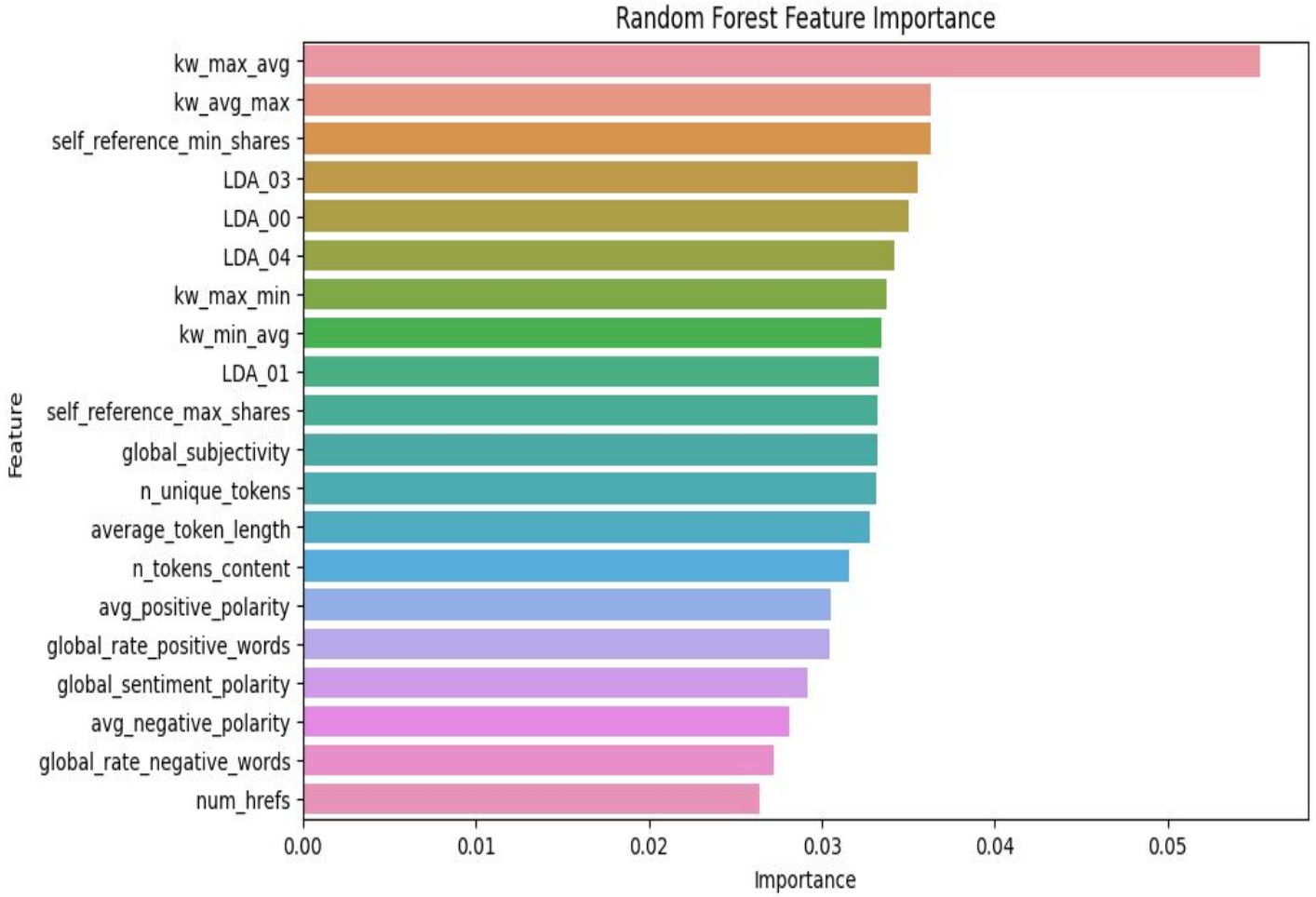
Results and Analysis

MODELS	TRAINING	TESTING
	Accuracy	Accuracy
Logistic	65.2	65.3
MLP	62.3	65.3
Bagging (LogReg)	65.5	65.13
Random Forest(500)	100	68.1
AdaBoost (LogReg)	64	63
Gradient Boosting	71	67.1
SVM	66.3	66
Naive Bayes	63.17	63.13
Stacking Classifier	90	67

Random Forests with 500 estimators, **Gradient Boosting Classifier** with 300 and **Stacking Classifier** with both random forests and gradient boosting classifier combined estimators gave us the best results.

We also managed to increment the performance of certain models from the models used by authors in our literature review. (Logistic Regression, Random Forests).

Results and Analysis



Feature Importance as per **Random Forests** were also analysed to find the basis of which feature affects the popularity of the new article. Max shares of avg keyword is the most important as per the graph.

Conclusions

Consequently, our continuous target variable has undergone a discretization process. This approach served two primary objectives:

- It facilitates the evaluation of different models based on accuracy metrics. By correctly classifying extreme values into their respective labels, all models can be compared based on the classification of non-extreme data points, enabling meaningful improvements.

- It provides more informative feedback, especially for news writers, as it predicts a label (e.g., "popular" or "unpopular") associated with each data point rather than an exact value that has less significance.

Conclusions

Recommendations for stakeholders are highlighted as follows:

- Publishers: The statistical summarization of articles, proves adequate for predicting popularity pre-publication. Entities prioritizing reputation and trust can confidently rely on our methodology for enhanced security and reliability assurance.

- Writers: Authors are advised to incorporate references to already-popular articles, enhance subjectivity in titles, and integrate more visual elements such as photos and videos. Additionally, it is recommended to refrain from using negative words, while avoiding multi-topic discussions whenever possible.

Conclusions

Furthermore, as cited earlier, we take pride in highlighting our notable performance in the literature, particularly within non-deep learning architectures. Our commitment extends towards future explorations in this domain, along with analysis of prevailing and emergent high-impact trends on top of popularity prediction.

We thank the entire machine learning course faculty for giving us the opportunity and guidance to perform on this project.

Timeline

Week 1-2: Exploratory Data Analysis

Week 3: Feature engineering

Week 4: Pre-processing for Regression

Week 5: Model Generation for Regression

Week 6: Threshold generation for discretization and Preprocessing for Classification

Week 7: Model Generation & Hyperparameter Tuning for Classification

Week 8-9: Model Evaluation & Ensembling Methods

Week 10: Comparisons and Conclusions

Team member's contribution

Individual Contributions:

- ***Aditya Jain:*** Data Analysis, Model Generation, Report
- ***Shrey:*** Data Analysis, Pre-processing, Report
- ***Vasu Kapoor:*** Data Analysis, Model Generation, Literature review, Report
- ***Vinayak Sharma:*** Pre-processing, Report, Model generation, Data Analysis, Literature review

References

[Models]

[EDA and RFE]