

A Project Report on

**Prediction of Online News Popularity**

Submitted in partial fulfilment for the award of the degree of

**Post Graduate Diploma Management**

In **Business Analytics**

Submitted by

**Vaibhav Tyagi**

R19DM054

Under the Guidance of

**Mr. Ratnakar Pandey**

Chief Data Scientist, ffreedom app

REVA Academy for Corporate Excellence

**REVA University**

Rukmini Knowledge Park, Kattigenahalli,

Yelahanka, Bangalore – 560064

**August, 2023**



# Candidate’s Declaration

I, **Vaibhav Tyagi,** hereby declare that I have completed the project work towards **Post Graduate Diploma Management in Business Analytics** at REVA University on the topic entitled **Prediction of Online News Popularity** under the supervision of **Mr. Ratnakar Pandey, Chief Data Scientist, ffreedom app.** This report embodies the original work done by me in partial fulfilment of the requirements for the award of the degree for the academic year **2023.**

Place: Bengaluru Name of the Student: Vaibhav Tyagi

Date: Signature of Student



# Certificate

This is to Certify that the Project work entitled **Prediction of Online News Popularity** carried out by **Vaibhav Tyagi** with **R19DM054,** a bonafide student of REVA University, is submitting the first-year project report in fulfilment for the award of **Post Graduate Diploma Management in Business Analytics** during the academic year **2023**. The Project report has been tested for plagiarism and has passed the plagiarism test with a similarity score of less than 15%. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said Degree.

Signature of the Guide Signature of the Director

Mr. Ratnakar Pandey Dr. Shinu Abhi

Chief Data Scientist, ffreedom app Director, Corporate Training

External Viva

Names of the Examiners

Place: Bengaluru

Date:



# Acknowledgment

At the outset, would like to indebt our Mentor Mr. Ratnakar Pandey for his time, valuable input, and guidance. His experience, support, and structured thought process guided me to be on the right track toward the completion of this project.

Heartily thankful to Dr. J.B. Simha our chief mentor for their time-to-time guidance. Extremely gifted and fortunate to have Dr. Shinu Abhi as our academic counselor and director. Her in-depth knowledge and her passion for lucidly delivering the subjects have helped me a lot. Special thanks to her for her guidance towards the entire coursework. Also, thank all the course faculty of the PGDM - BA program for providing me with a strong foundation in various concepts of analytics and machine learning.

Also acknowledge the support provided by Hon’ble Chancellor, Dr. P Shayma Raju, Hon’ble Vice Chancellor, Dr. M. Dhanamjaya, and Registrar, Dr. N Ramesh in all circumstances.

Last but not the least, I would like to sincerely thank my respective family for giving me the necessary support, space, and time to complete this project. I will certify that the work done by me for conceptualizing and completing this project is original and authentic.

Place: Bengaluru

Date:



# Similarity Index Report

This document attests that the project report titled **"Online News Popularity Prediction"** was compared to other documents using a similarity detection service. The process and outcome are given below. The plagiarism report is attached in the appendix.

Software Used: Turnitin

Date of Report Generation:

Similarity Index in %:

Total word count:

Name of the Guide: Mr. Ratnakar Pandey

Place: Bengaluru Name of the Student: Vaibhav Tyagi

Date: Signature of Student

Verified by: Dincy Dechamma

Signature

Dr. Shinu Abhi,

Director, Corporate Training

# List of Abbreviations

|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Abbreviation** | **Long Form** |
| 1 | LR | Logistic Regression |
| 2 | DT | Decision Tree |
| 3 | AUC | Area Under the Curve |
| 4 | RF | Random Forest |
| 5 | URL | Uniform Resource locator |
| 6 | VADER | Valence Aware Dictionary and Sentiment Reasoner |
| 7 | WOM | Word of Mouth |
| 8 | CRISP-DM | Cross-Industry Standard Process for Data Mining |
| 9. | RFECV | Recursive feature elimination with cross-validation |
| 10. | KNN | K-Nearest Neighbours |

# List of Figures

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Name** | **Page No.** |
| Figure No. 5.2.1 | The cross-validation score versus the number of features selected using RFECV for logistic regression estimator |  |
| Figure No. 5.2.2 | The cross-validation score versus the number of features selected using RFECV with RF estimator. |  |
| Figure No. 5.2.3 | The cross-validation score versus the number of features selected using RFECV with Adaboost estimator. |  |
| Figure No. 5.3.1 | Performance of three classifiers under default parameter setting. |  |
| Figure No. 5.3.2 | Evaluation of three classifiers under default parameter setting. |  |
| Figure No. 5.4.1 | Performance metrics for three supervised learning refined hyperparameters. |  |
| Figure No. 5.4.2 | Evaluation metrics for three supervised learning refined hyperparameters. |  |
| Figure No. 7.1 | Count of popular/unpopular news over different days of a week |  |
| Figure No. 7.2 | Count of popular/unpopular news over different article categories |  |
| Figure No. 7.3.1 | Data projected on the first two principle important Components |  |
| Figure No. 7.3.2 | Data is projected on the first three principle important components. |  |
| Figure No. 8.1 | Correlation between default features |  |
| Figure No. 9.1 | Content length vs Number of shares |  |
| Figure No. 9.2 | Text sentiment popularity vs Number of shares |  |
| Figure No. 9.3 | Comparisons of popular vs unpopular percentages in Mean and Median |  |
| Figure No. 10.1 | Comparison of subjective information using Sentimental analysis in Positive, Negative, and neutral |  |
| Figure No. 10.2 | Correlation between sentiment score positive on title vs social shares |  |
| Figure No. 10.3 | Top 10 Negative features by VADER score |  |
| Figure No. 10.4 | Top 10 Positive features by VADER score |  |
| Figure No. 10.5 | Top 10 Neutral Features by VADER |  |
| Figure No. 10.6 | The word cloud shows that specific keywords may be more informative and technological |  |

# List of Tables

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Name** | **Page No.** |
| Table No. 5.1 | List of predictive attributes of the dataset. |  |
| Table No. 5.2.1 | 29 features were selected using RFECV with logistic regression estimator. |  |
| Table No. 5.2.2 | 18 features were selected using RFECV with Adaboost estimator |  |
| Table No. 5.3.1 | Metrics score of three classifiers under default hyperparameters. |  |
| Table No. 5.4.1 | Metrics score of three classifiers under refined hyperparameters. |  |
| Table No. 9.1 | Regression model error comparison |  |
| Table No. 9.2 | Classifier comparison between four models |  |
| Table No. 11.1 | Random Forest (RF) Model for news classification |  |

# Abstract

Due to its rapid dissemination and easily accessible attributes, the Internet has supplanted traditional newspapers and periodicals as the major platform for broadcasting public news. Thus, estimating Internet news's rising profile is an increasingly important topic.

Due to advancements in the Internet and technology, online news items and blogs have become increasingly popular. Even while print news isn't entirely gone, more and more individuals choose to check the Internet for the day's events because it's quick, easy, and free. Pew Research found in 2018 that slightly more than 50% of Americans acquire their news in some form online. Forbes reports that the figure reached 55% in 2019 and is still rising. Using data on Mashable articles from the previous two years, machine learning models that forecast the popularity of an article given a set of features about that article—such as the number of words in the article, time of week on which article appeared, and the average sentiment polarity of the article content—have been created. Divide the continuous predictor variable "shares" into two categories before running these models: "Low" popularity for articles with multiple shares below the median share total for all articles, and "High" popularity for articles with numerous shares above or equal to that median total.

Three machine learning models predict the level of popularity of a certain article using the Naive Bayes, Random Forest, and K-Nearest Neighbours (KNN) algorithms. Overall, the accuracy gained from the three models is subpar and is not much improved by the various data processing methods. The models are performed with supervised discretization using Python and the Valence Aware Dictionary and Sentiment Reasoner (VADER) module. The best overall performance, however, calculated and achieved by using a Random Forest model with supervised discretization and without "looking" at the test set of (67%).

Keywords: Text Mining, Natural Language Processing, Sentiment Analysis

# Table of Contents

[Candidate’s Declaration 2](#_Toc131593709)

[Certificate 3](#_Toc131593710)

[Acknowledgment 4](#_Toc131593711)

[Similarity Index Report 5](#_Toc131593712)

[List of Abbreviations 6](#_Toc131593713)

[List of Figures 6](#_Toc131593714)

[List of Tables 7](#_Toc131593715)

[Abstract 8](#_Toc131593716)

[Table of Contents 9](#_Toc131593717)

[Chapter 1: Introduction 10](#_Toc131593718)

[Chapter 2: Literature Review 12](#_Toc131593719)

[Chapter 3: Problem Statement 18](#_Toc131593720)

[Chapter 4: Objectives of the Study 19](#_Toc131593721)

[Chapter 5: Project Methodology 20](#_Toc131593722)

[Chapter 6: Business Understanding 27](#_Toc131593723)

[Chapter 7: Data Understanding 28](#_Toc131593724)

[Chapter 8: Data Preparation 31](#_Toc131593725)

[Chapter 9: Modeling 33](#_Toc131593726)

[Chapter 10: Model Evaluation 36](#_Toc131593727)

[Chapter 11: Analysis and Results 39](#_Toc131593728)

[Chapter 12: Conclusions and Future Scope 40](#_Toc131593729)

[Bibliography 42](#_Toc131593730)

# Chapter 1: Introduction

This document attests that the project report titled "Online News Popularity Prediction" was compared to other documents using a similarity detection service. The methods Random Forest, linear regression, and Adaboost are frequently used in categorisation. Predicting how well internet news will be recognised via machine learning techniques this study investigates several models and feature sets, and then applies various machine learning algorithms based on the selected characteristics. Mashable, a well-known international multi-platform media and entertainment firm, is the source of the data utilised in this analysis. Area Under the Curve, Functionality Measure, and Precision are compared to find the most precise method based on the evaluation of the data. Based on the findings, One of the simplest prediction models, Random Forest may achieve an accuracy of around 67%.

The proliferation of smartphones and Internet connectivity has increased the number of news websites covering a wide range of issues, including politics, economics, business, entertainment, technology, fashion, and many more. We think that our study will be useful to online news organisations in gauging the interest in their stories prior to publication.

Academic studies frequently examine how to foresee the proliferation of internet news. User comments augment content analysis on the internet. It turns fame into a contest in which the winners are the days on which the most enticing stories were published. To classify how interesting a piece of internet news is, we employ a ranking support vector machine (SVM).

The work of predicting the popularity of web material still confronts a number of significant difficulties, despite the fact that it has significant effects in many different fields. First, a variety of elements make prediction challenging, such as the popularity of a piece of information can be influenced by its quality or relevance to viewers. Second, it can be challenging to both gather and provide information to the prediction engine about how real-world occurrences relate to the content itself. Finally, predicting intricate social interactions and information cascades at the microscopic level is quite difficult. Fourth, the forecast could also be challenging due to material that is hard to get, such as context from sources outside of the web, regional conditions, and circumstances that have an impact on the population. Last but not least, the prediction could also be challenging depending on network factors like network topology and how different web layer interactions interact.

The era of reading and sharing information and news has gained popularity as it has become part of people’s entertainment nowadays. Hence, predicting the popularity of news before its publication on how news includes different topics about the world and how it will affect an individual in both positive and negative influences plays a crucial role. Samples of 39484 observations from the articles are considered from the Mashable website, which is published between the years 2013 and 2015 on Mashable’s official website (*UCI Machine Learning Repository: Online News Popularity Data Set*, n.d.).

This study challenges the traditional binary job of evaluating popularity as either popular or unpopular and instead employs multiple advanced techniques such as Random Forest, Adaptive Boosting, and SVM. The study also uses real-time rolling windows to operate within. In addition, this study utilizes the well-known news source, Mashable (mashable.com/), which has not been previously examined for popularity forecasting, and compiles a comprehensive dataset covering the last two years, which is a much larger timeframe than what is typically used in literature.

In recent years, the online news sector has grown dramatically, and machine learning algorithms are being used more frequently to analyse huge databases and forecast readership. In this work, a variety of machine learning techniques were employed to predict the recognition of online news based on a collection of variables, including Random Forest, linear regression, and Adaboost. Mashable, a well-known international media and entertainment organisation that offers online news and information, provided the dataset for the study.

The goal of the study was to assist online news organisations in estimating before a story is published how popular it will be. The calibre and applicability of the news to viewers are just two examples of the many variables that might affect how popular internet news is. It might also be difficult to compile and present data regarding how actual events connect to the material itself. Furthermore, it can be challenging to forecast intricate social interactions and information cascades at the microscopic level.39,484 observations from Mashable's website between 2013 and 2015 were used in the study. Different machine learning algorithms were used to analyse the dataset, and the results were assessed using AUC, F-measure, and precision to identify the best accurate algorithm. The simplest model for making predictions was determined to be the Random Forest method, which had a 67% accuracy rate.

The study's overall goal was to offer knowledge on anticipating the popularity of online news and assist online news organisations in making judgements about the news articles they publish. Large datasets and the application of machine learning algorithms can assist improve the quality of online news material while also shedding light on the elements that affect its popularity.

# Chapter 2: Literature Review

Ananto et al. employed Genetic Algorithms along with Machine Learning Methods to forecast the popularity of online content, as noted by Wicaksono and Supianto in 2018. The Random Forest algorithm's parameters include the number of Decision Trees, minimum sample leaf size, leaf weight percentage, and optimizing the hyperparameter K in the KNN technique. The crossover and mutation rates were both set to 0.5, and the study performed ten iterations. The team used Scikit-learn's evolutionary algorithm search CV with a population size of 50 and a generation number of 1000 to implement this approach. Despite the use of evolutionary algorithms and hyperparameter optimization, the KNN method produced the quickest results, with an accuracy of only 58% in 61 seconds. In comparison, the Random Forest algorithm took 659 seconds and achieved an accuracy of 67%.

When analysing the question of popularity, Hensinger et al. frame it as a competitive setting in which the day's most appealing content win out. They propose a linear "appeal" function as a solution, which is essentially a bag-of-words model. The parameters of this linear function are learned from pairs of articles that have different levels of popularity while being published on the same day and page. Using a year's worth of data from six distinct sources, an appeal function was trained with the help of Support in Ranking Using Vector Machines. [(Hensinger et al., 2013)].

**Rizos and his colleagues** employed a framework to predict a variety of news article popularity data, including the number of comments, users, votes, and a contentious Ness score. In three online news post datasets, it is demonstrated how the recommended graph-based characteristics significantly improve the process of identifying contested articles and the prediction of all popularity indices. That is to illustrate the complexity of the two networks of social interaction. In a news-focused Reddit dataset, a method for controversiality prediction that previously employed solely fundamental comment tree traits found a 5% relative improvement (Rizos et al., 2016).

In an article entitled "Social Media Outpaces Print Publications in the U.S. as a News Source," Elisa Shearer presents data from the Pew Research Centre showing that social media platforms are now more popular than print newspapers among Americans. In an article entitled "Social Media Outpaces Print Publications in the U.S. as a News Source," Elisa Shearer presents data from the Pew Research Centre showing that social media platforms are now more popular than print newspapers among Americans. One in five Americans now regularly gets their news from social media, surpassing the number who read print newspapers (16%). The poll also indicated that 27% of respondents aged 18-29 regularly visit news websites, making them the second most popular news source among this age group. For news, 36% of Millennials and Gen Zers choose social media, whereas 27% prefer news websites, 16% prefer TV, 13% prefer radio, and 2% prefer print. Younger Americans are distinctive because they don't rely on just one platform as their elders did, who relied mostly on television. Between the ages of 18 and 29 and between the ages of 30 and 49, about half of the population regularly uses just one news source.

The purpose of the research conducted by Namrata Godbole and titled "Large-Scale Sentiment Analysis for News and Blogs" is to create a system that can give entities referenced in news stories and blogs favourable or negative ratings. There are two stages to this system: (1) detecting emotions, and (2) collecting, analysing, and ranking those emotions. During the first stage, views are matched with the appropriate entities, and during the second, sentiments are aggregated and scored. This research looks at the value of various scoring methods for a massive database of news and weblogs. The research also intends to broaden the scope of spatial analysis of news items to include sentiment maps, which can pinpoint areas of the world where people are more likely to have positive or negative feelings about specific topics.

Igor Suciu, Increased Use of Social Media as a Source of News among Americans. More and more people in the United States are reading news articles on social media.. It's no secret that the vast majority of Americans don't follow the news through reading newspapers or tuning in to the evening broadcast. As the official newspaper increasingly focuses on local news and sports scores, social media platforms like Facebook and Twitter are becoming increasingly attractive alternatives. A recent study by the Pew Research Centre found that 55% of Americans now "often" or "often" acquire their news from social media, up 8% from the previous year. Up from 20% in 2018, about 30% of respondents said they "frequently" accessed the news. "Knowing that Russians used Facebook to influence a sizable number of political views with their conflict with Ukraine; and despite knowing, it had an impact on the presidential election," Safko said.

Ren and Yang's goal was to find the best machine learning model and characteristics to foretell the success of online news. Mashable is a popular news website, and the researchers used data from there to try out 10 different learning algorithms on. Each algorithm's efficiency was evaluated and compared to the others. To improve accuracy and minimise the amount of features, feature selection techniques were used. According to the findings, Random Forest was the best prediction model, with an accuracy of 70% when using the best settings.

A Survey of Social Media Prediction, by Sheng Yu and Subhash Kak (Yu & Kak, n.d.). Brands are mentioned in about 19% of tweets, however the vast majority of these posts (almost 80%) are emotionless advertisements. We made a decision to research the impact of online communities and widespread e-word-of-mouth (eWOM) on product uptake so that we may better predict its future popularity. The number of supporters and advocates is staggering. As of 2009, however, 77.9% of user pairings were one-way. This means that just 22.1% of user couples are in committed relationships. As a result, people's inner circles of real friends are more tighter and more fundamental. Even if the results of this kind of study are still quite inaccurate, social media has provided us a new means to gather, extract, and objectively utilise the wisdom of crowds at little cost and with great efficiency.

Tatar, Alexandru, How to gauge what will go viral on the web: a survey [(Tatar et al., 2014)]. By filtering out irrelevant content, users can focus their attention, the most valuable commodity in the digital realm, on the data that truly matters to them. In a market where companies are allocating up to 30 percent of their advertising budget online. There is no universally accepted definition of "viral" material online. Increased online activity or presence may be a sign of popularity, whether for an individual or a business. The application of content popularity prediction to the planning of networking solutions is not well documented. However, more scalable content distribution strategies that proactively repeat information in response to user demand may be developed by understanding the dynamics of online content popularity.

The article "Predicting Online News Popularity" ((Predicting Online News Popularity (Part 1) | by Benjamin Dornel | Towards Data Science, n.d.)) was written by Benjamin Dornel.. Determine the factors that contribute to a story's success or failure in the news. The first step is to examine online New York Times stories and make predictions about their popularity using factors like word count, headline length, and abstract length. These are wordless interactive articles. Since this information isn't truly 'Missing At Random' (MAR), we opted against using imputation to fill in the gaps. Articles with shorter abstracts and headlines (50-130 words) tend to do better. Popular items are more likely to come from the Politics, Games, and Washington news departments, while those from the Sports, Culture, and Podcasts desks tend to be ignored. In comparison to obituaries and straight news reporting, news analysis and interactive articles tend to be more read.

Using machine learning and a variety of data mining methods, Feras Namous's Online News Popularity Prediction [(Namous et al., 2019)] sought to identify the most effective model and feature set for making such predictions. The popular news website Mashable was the source of the data. Precision was used to assess the results, recall, and F-measure, with the latter two metrics compared to determine the best option. In addition, we compared our results to those of other studies using the same dataset. The top prediction models are revealed to be Random Forest and Neural Network, both of which attain an accuracy of 65% with the best-case settings. With an F-measure of around 65%, the best performance was achieved by using 0.1 as a learning rate, four hidden layers, and Random Forests or Multilayer Perception. However, combining Bagging or AdaBoost with Random Forests into a hybrid model opens up further room for development.

The Future of Newspapers in an Age of Social Media by Koustav Mukherjee [(Mukherjee, n.d.)]. Multiple studies, from a variety of angles, have examined the impact of social media on the journalism business. People claim to have entered a new era in which ideas can be communicated in as few as 140 characters, thanks to the rapid expansion of social media platforms like Orkut, Twitter, Facebook, and Instagram. Journalists' thoughts on how social media has altered their work and the way they do it are explored in this research. The development of technology has been crucial in the production of these aggregations. Because of the increased competition, established media outlets will strive to uphold their reputations by providing consumers with accurate and engaging news coverage.

Rasmus Kleis Nielsen (Alessio Cornia, Kleis Nielsen, & Antonis Kalogeropoulos, n.d.) examines the difficulties and opportunities facing the news industry in today's increasingly mobile, digital, and social media-driven environment. It's a sign that we're entering a new era characterised by more competition for people's attention as the world becomes increasingly mobile, digital, and social media-centric. The majority of people nowadays acquire their news from digital sources, mostly online and on their mobile devices (especially smartphones) through shared links and recommendations from friends and family. There are 31% "Proactive Participants" whose regularly post writings, comment on them, occasionally blog or participate in campaigns, and 48% "Passive Consumers" who might converse with friends and coworkers both online and offline about news, but refrain from participating in more participatory forms of media use such as posting or commenting. This made traditional forms of media such as television and newspapers more challenging to succeed in. Their importance in the newsmaking process remains high, but their influence as news carriers is dwindling.

Data Mining and Prediction from Online News Feeds, by Arpit Garg (Online News Feeds, n.d.). The goal of the Online News Feed Prediction System is to compare and contrast various prediction algorithms by analysing their respective implementation methodologies. Media outlets may utilise this strategy to increase readership and revenue. The goal is to figure out how to get there. With LinkedIn as the target, "Topic" transformed into four columns of zeroes and ones, and the remaining parameters chosen as features, the prediction accuracy was 54.38%. After resegregating "Topic" into bins of 0, 1, 2, 3, and "SentimentTitle" and "SentimentHeadline" into bins of -1, 0, and 1, the testing set's accuracy score was 0.547082797083, or 54.7% for criteria. When switching the criteria to Entropy, the accuracy increases to 54.64 percent. Python is used for the SVM implementation, and a train test split is utilised to split the data (60% to 40%) between the training and testing sets. Facebook is the most well-liked social media platform, followed by LinkedIn, while Google Plus is the least. Headline popularity rankings for Microsoft aside, LinkedIn and Google Plus are quite similar. The popularity rankings are heavily influenced by the social media site through which the news is released.

(Predicting Online News Popularity, n.d.) K.S.NagaHaritha. There has been a rise in interest in finding ways to forecast the reach of online news stories. Logistic Regression (LR) is a machine learning technique used to estimate the probability of an outcome involving a categorical dependent variable. There are just two possible values for the dependent variable in LR: 1 (yes, success, etc.) or 0 (no, failure, etc.). To prioritise challenging circumstances, the AdaBoostclassifier ensemble classifier first updates the weights of several copies of the fitted classifier on the original dataset.. Sklearn () provides a tool called AdaBoostClassifier. There is a customizable limit on the size of the boosted ensemble of estimators. It is well accepted that remodels are low-bias, high-variance models. Training and testing RF takes significantly longer when 500 trees are used in the forest, but this setup yields superior results in terms of ACCURACY F1-SCORE and Area Under the Curve (AUC).

# Chapter 3: Problem Statement

Using machine learning techniques, it is possible to forecast, before publication, whether an online news piece will become popular. The quantity of shares indicates how popular something is. The article is classified as popular if it receives more shares than a set threshold; otherwise, it is classified as unpopular. Thus, the challenge is to properly categorise the goal label (popular/unpopular) of the articles that are to be published using a set of article attributes and the best machine learning model.

*Since this problem can be approached as a binary classification problem, three classification learning algorithms including logistic regression, Random Forest (RF), and Adaboost are implemented and compared. The best model is selected based on the indicators presented in the project methodology chapter.*

# Chapter 4: Objectives of the Study

The major goal of this study is to make:-

Predictions about whether an article will be popular or not are based on numerous characteristics, such as the number of words it contains and the number of times it has been shared.

When working on visualizing the features, it was discovered that there was little correlation between dependent and independent features. However, when each column was compared to shares, some insights were gained, such as the fact that articles with titles that contained neither fewer words nor more words received more shares.

1. Any article will be analyzed by users. In addition to providing a forecast, this would analyze the article's emotion to determine if it is favorable, moderate, or negative.
2. Need to productize this application which will the online bloggers or online writers estimate for the future whether the article is having a good response or not. Also, helps in predicting whether related news going to publish in coming days.

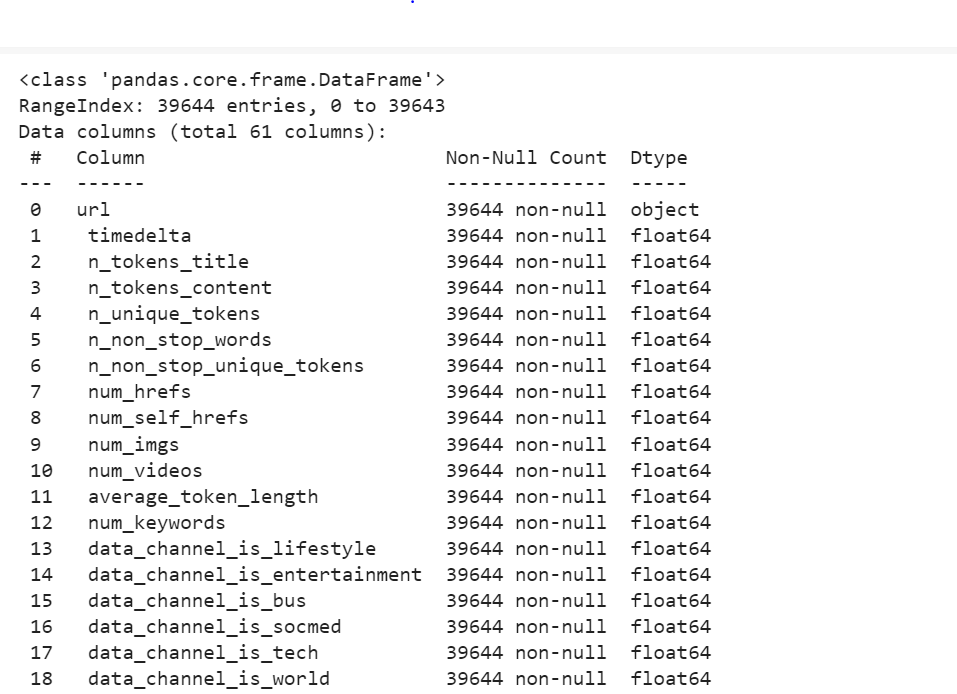
# Chapter 5: Project Methodology

To solve the business objective, the traditional Cross Industry Standard Process for Data Mining “CRISP-DM” methodology is used which involves the following tasks.

**Data Exploration and Visualization**

There are 39,644 articles in the dataset, all of which were pulled from the online news website Mashable between January 2013 and January 2015. The UCI Machine Learning Repository is where it is retrieved from.

TABLE 5.1: LIST OF DATASET



Earlier, the dataset was pre-processed. For instance, a one-hot encoding strategy was used for the categorical characteristics, such as the day of publication and article type, while a log transformation was applied to the skewed features, such as the number of words in the article.

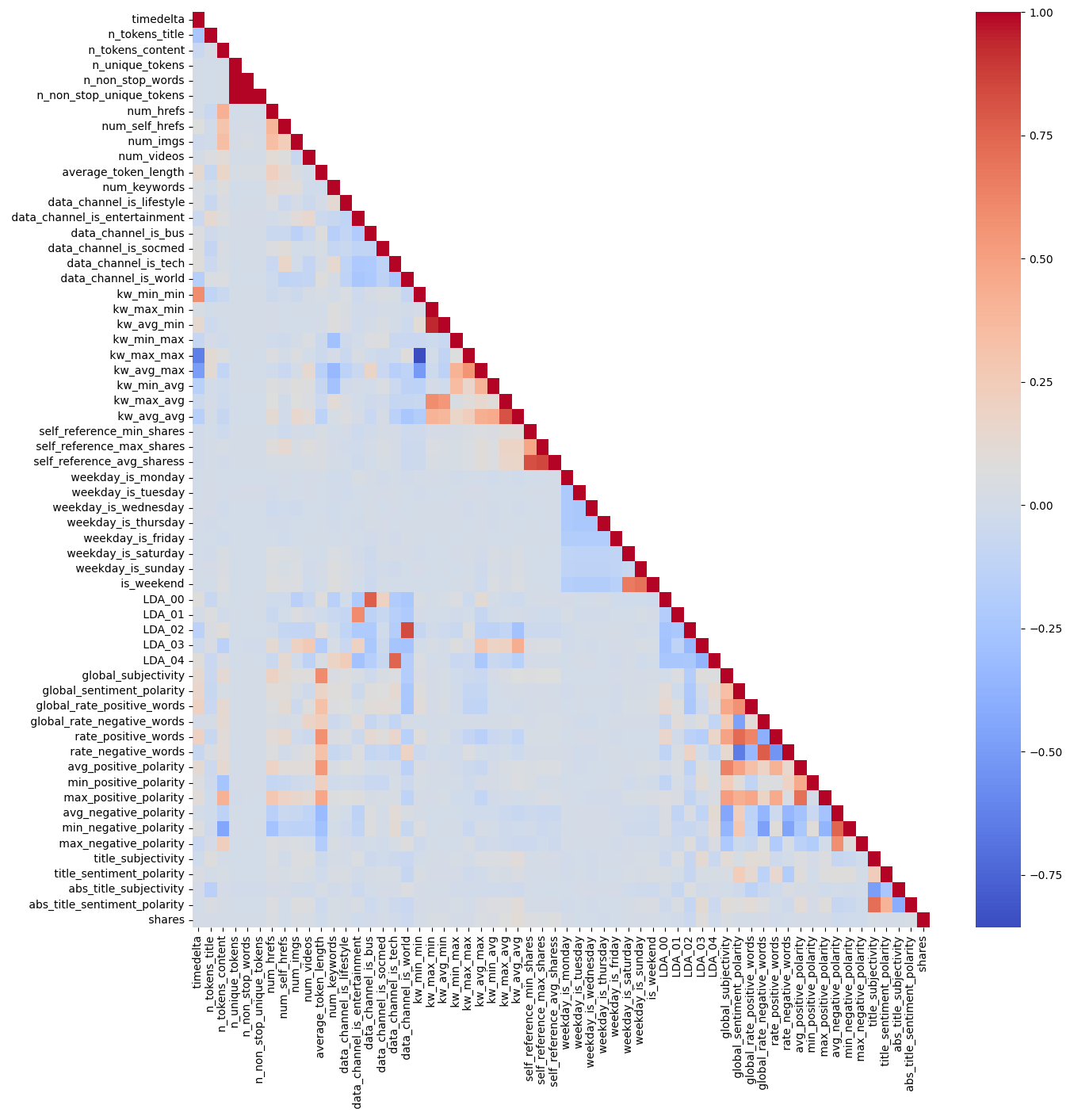


Fig. 5.1.1: Heatmap of a correlation matrix

**Data Pre-processing**

As was discussed in Section 5.1, preliminary work on the data has been completed. A one-hot encoding approach was used to modify categorical information like publication day and article category, while a log transformation was used to skewed features like the number of words in the article. In order to ensure that each feature is given equal weight during supervised learning, the dataset is further pre-processed by normalising the numerical features to the interval [0, 1]. To make the continuous target characteristic into a boolean label, choose its median as the threshold.

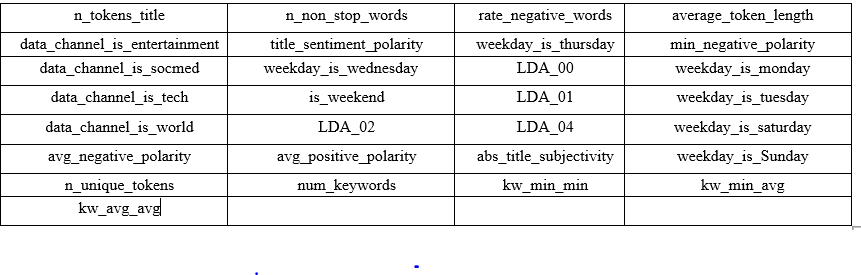
Chart, line chart

Description automatically generated

Fig. 5.2.1: RFECV for logistic regression estimator: cross-validation score versus feature selection

There are 58 features in the dataset, thus it makes sense to do a feature selection in order to lower the data noise and speed up the algorithm's execution time, as illustrated in Fig. 5.2.1.

TABLE 5.2.1: 29 FEATURES SELECTED USING RFECV WITH LOGISTIC REGRESSION ESTIMATOR



Selecting the most important characteristics for a classifier may be automated with the help of a technique called Recursive characteristics Elimination with Cross-Validation (RFECV). A useful function, REFCV(), is available in Sklearn.

Table 5.2.1 demonstrates that a logistic regression estimator is required for RFECV. correlation between the cross-validation score and the proportion of features chosen When more than 29 elements are present, the score decreases. From the original set of 58 features, the RFECV algorithm picks the 29 most relevant ones. Table 5.2.1 details the selected 29 characteristics. These 29 features also account for the characteristics of the article category and the days of the week. The next steps include using an RF estimator to carry out RFECV with the following characteristics:-

n\_tokens\_title, n\_unique\_tokens, num\_hrefs,num\_self\_hrefs, n\_tokens\_content num\_imgs, num\_videos, num\_keywords, avg\_token\_length, LDA\_00, kw\_min\_max, LDA\_01, kw\_min\_min, data\_channel\_is\_entertainment, kw\_max\_min, kw\_min\_avg, data\_channel\_is\_tech, global\_sentiment\_polarity, avg\_positive\_polarity, avg\_negative\_polarity, global\_rate\_negative\_words , title\_subjectivity, kw\_max\_avg, self \_reference\_max\_shares, LDA\_02, LDA\_03, kw\_min\_max, abs\_title\_sentiment\_polarity ,kw\_avg\_avg, self\_reference\_avg\_shares, global\_subjectivity, global\_rate\_positive\_words, rate\_positive words, min\_positive polarity, title\_sentiment\_polarity, abs\_title\_sentiment\_polarity ,m\_length

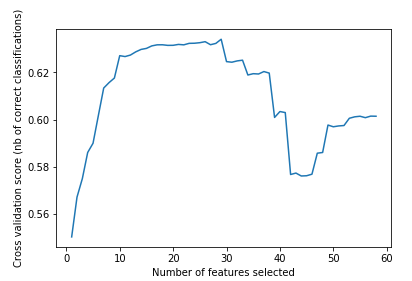


Fig. 5.2.2: The number of features chosen by RFECV with the RF estimator plotted against the cross-validation score.

The relationship between the cross-validation score and the number of chosen characteristics is seen in Fig. 5.2.2. The 56 characteristics chosen for RF by RFECV are nearly all of the features available in the dataset. 'n nonstop words' and 'data channel is lifestyle' are the only two characteristics that are not considered by RFECV while evaluating RF.

Lastly, run RFECV with Adaboost estimator on the following features:-

n\_unique\_tokens, num\_hrefs, num\_self\_hrefs, num\_imgs, num\_videos, num\_keywords, data\_channel\_is\_socmed, data\_channel\_is\_entertainment, data\_channel\_is\_tech, kw\_max\_min, data\_channel\_is world, kw\_avg\_max , kw\_max\_max, kw\_min\_max, kw\_min \_avg, kw\_max\_avg, self\_reference\_avg\_sharess, kw\_avg avg, self\_reference\_min\_shares weekday\_is\_ Friday, LDA \_01, global\_subjectivity, , is\_weekend, LDA\_00 global\_rate\_positive\_word, avg\_positive\_polarity, , rate\_negative\_words in\_positive\_polarit, title\_subjectivity

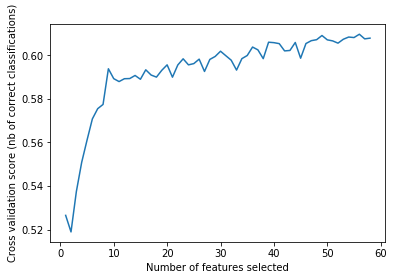


Fig. 5.2.3: RFECV with Adaboost estimator cross-validation score vs feature selection size.

The cross-validation score vs the number of features chosen is plotted in Fig. 5.2.3. The optimal value for "Number of features selected" appears to be 18. Therefore, out of a total of 58 characteristics, the RFECV algorithm only uses the top 18.

TABLE 5.2.2: 18 FEATURES SELECTED USING RFECV WITH ADABOOST ESTIMATOR.

| Feature selected USING RFECV | | | |
| --- | --- | --- | --- |
| n\_non\_stop\_unique\_tokens | num\_self\_hrefs | num\_imgs | num\_videos |
| data\_channel\_is\_entertainment | num\_keywords | kw\_avg\_min | kw\_min\_min |
| data\_channel\_is\_socmed | kw\_min\_max | kw\_max\_max | kw\_max\_avg |
| global\_sentiment\_polarity | kw\_avg\_avg | self\_reference\_min\_shares | self\_reference\_max\_shares |
| global\_subjectivity | kw\_min\_avg |  |  |

As shown in Table 5.2.2, the selected 18 features are listed. As per section 5.3, will going to implement the classification algorithms using these selected features.

**Implementation**

Each method randomly divided the dataset with its own chosen characteristics into a training set and a testing set before deployment. The sklearn functions LogisticRegression(), Random-ForestClassifier() and AdaBoostClassifier(), respectively, implement logistic regression, RF and Adaboost. We’ll start by using the model hyperparameters’ default values. The default hyperparameters for logistic regression are "C = 1.0", "n estimators = 10", "learning rate = 1.0", and "n estimators = 50" for Adaboost. In the next section, parameters need to be refined.

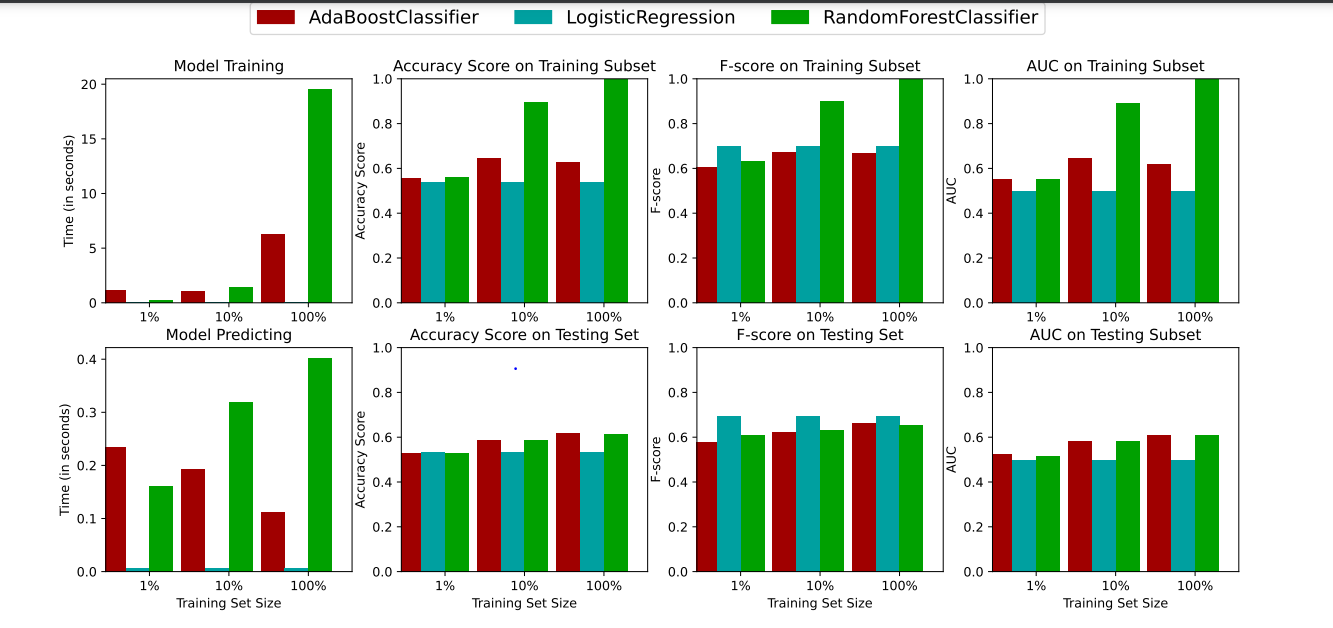


Fig. 5.3.1: Performance of three classifiers under default parameter setting

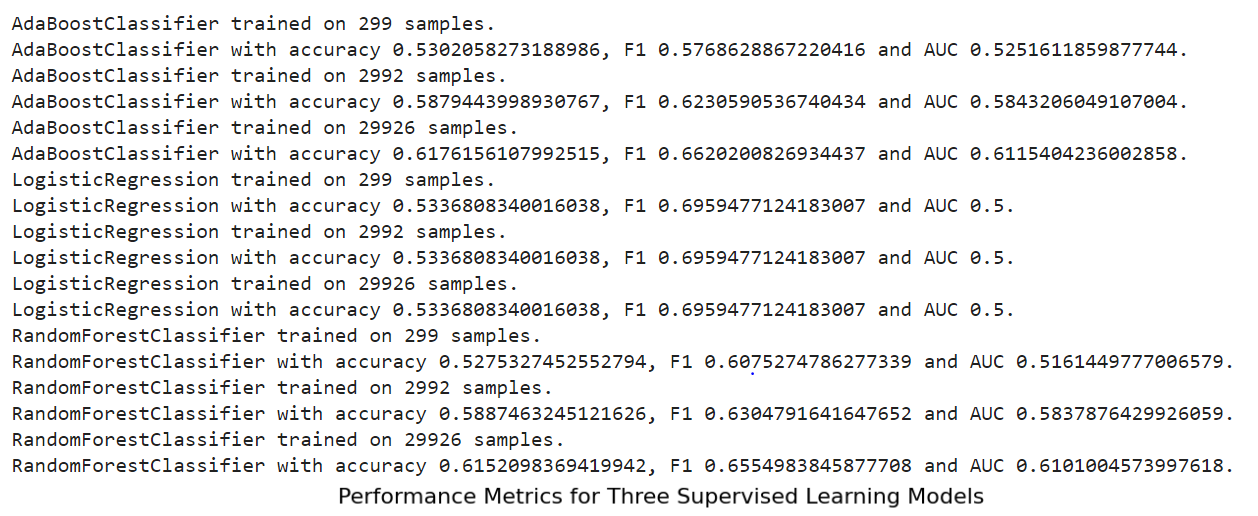


Fig. 5.3.2: Evaluation of three classifiers under default parameter setting.

As shown in Fig. 5.3.1 and 5.3.2, demonstrate the performance and evaluation of the three classification methods using the chosen characteristics for each algorithm.

TABLE 5.3.1: METRICS SCORE OF THREE CLASSIFIERS UNDER DEFAULT HYPERPARAMETERS.

|  |  |  |  |
| --- | --- | --- | --- |
| **Classifier** | **Accuracy** | **F1-score** | **AUC** |
| **Logistic**  **Regression** | 0.533 | 0.6959 | 0.5 |
| **RF** | 0.615 | 0.655 | 0.61 |
| **Adaboost** | 0.6176 | 0.6832 | 0.6513 |

Table 5.3.1 summarises the accuracy, F1-score, and AUC measures. While Adaboost performs best across the board with the default parameter settings, RF outperforms logistic regression in AUC and F1-score, while logistic regression outperforms RF in accuracy. In terms of how quickly models can be trained and evaluated, logistic regression and RF both outperform Adaboost.

**Using some other models**

The provided results show the performance of three different classifiers, Gaussian Naive Bayes, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN), trained on three different sample sizes of 299, 2992, and 29926 samples.

For all classifiers and sample sizes, the accuracy is around 0.53, which means that the models are correctly classifying approximately 53% of the instances. However, an accuracy score alone may not be sufficient to evaluate the performance of a classifier, especially when the classes are imbalanced, and one class dominates the other.

That's where other evaluation metrics like F1 score and AUC come into play. The F1 score takes into account both precision and recall and is a better metric to use when there is an imbalance in the classes. In this case, the F1 score is around 0.66-0.70, which means that the models are not able to distinguish between the positive and negative classes effectively.

Moreover, the AUC score is another metric to evaluate the performance of a classifier, which measures the ability of the classifier to distinguish between the positive and negative classes. An AUC score of 0.5 suggests that the classifier is not able to distinguish between the two classes and is performing no better than random guessing.

Therefore, based on the provided results, it appears that the classifiers are not performing well in classifying the instances. Improving the performance may require using different classifiers, exploring different feature engineering techniques, or increasing the sample size to obtain more representative data.

.

As shown in Table 5.4.1, the three metrics scores are summarized. Compared with the metrics in Table 5.3.1, the metrics of logistic regression and Adaboost are just slightly improved, but the metrics of random forest are significantly improved after tuning by grid search.

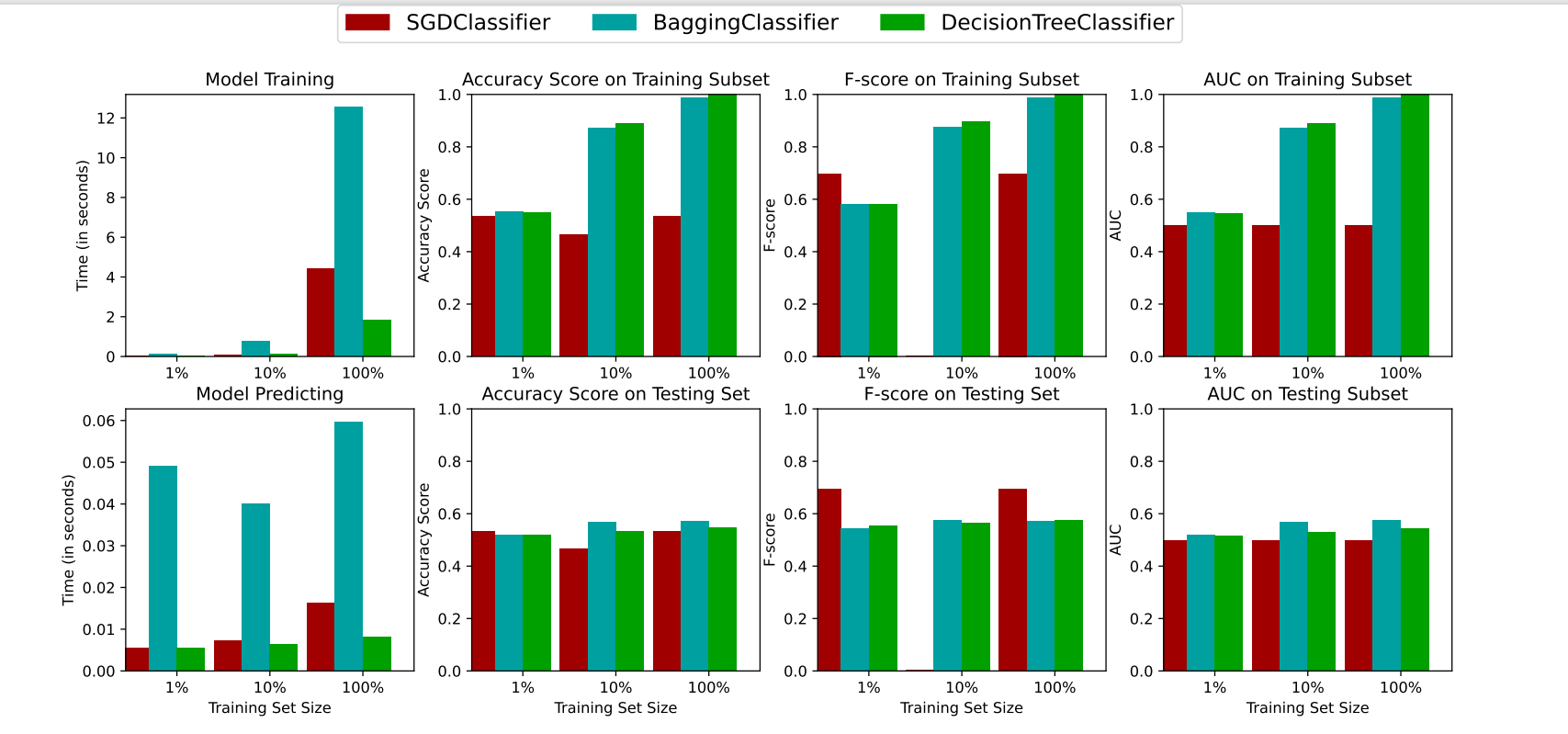


Fig. 5.4.1: Performance metrics for three supervised learning refined hyperparameters.

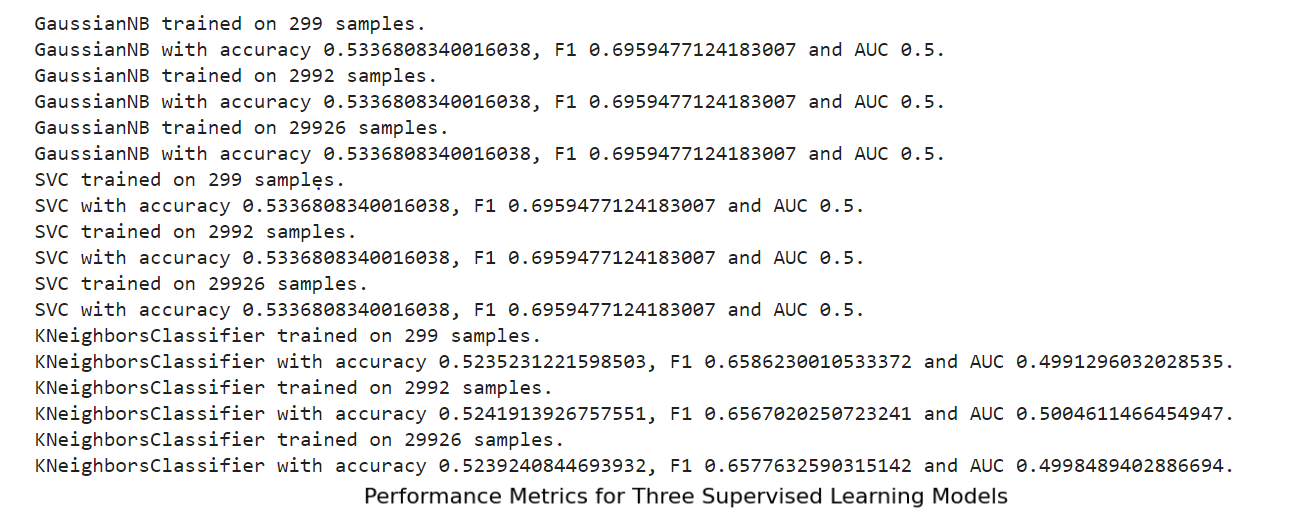


Fig. 5.4.2: Evaluation metrics for three supervised learning refined hyperparameters.

# Chapter 6: Business Understanding

Journalists, advertisers, content providers, etc. value news articles published online as a useful resource. Online news portals have taken over as the main news source with the advent of cell phones and the Internet. The key to reaching a large audience is not just how many people read the content, but also how many people share it on social networking sites like Facebook, LinkedIn, Twitter, etc. The number of shares is a well-known indicator of how popular the news is, which affects how much money is made through ads. As noted in the research, the main difficulty in forecasting the popularity of a blog post is that it heavily depends on human conduct.

The rise of the internet and social media has significantly impacted the way people consume news and information. With the increasing accessibility of cell phones and the internet, online news portals have become the primary source of news for many people. As a result, journalists, advertisers, content providers, and other stakeholders in the media industry place a high value on news articles published online.

One crucial aspect of online news articles is the number of shares they receive on social media platforms like Facebook, LinkedIn, Twitter, and others. The number of shares is an essential indicator of how popular an article is, which ultimately impacts the revenue generated through advertising. The more shares an article gets, the higher the chances of it reaching a larger audience, thereby increasing the chances of generating more revenue through ads.

However, predicting the popularity of an article is not an easy task as it heavily depends on human behavior. The behavior of individuals on social media, including their preferences and interests, significantly influences the popularity of news articles. Therefore, accurately forecasting the popularity of a blog post requires understanding the complex interactions between human behavior and online content.

predicting the popularity of a news article or blog post is challenging because it is influenced by various factors, including the quality and relevance of the content, the timeliness of the topic, the writing style, the format of the article, and the target audience. Additionally, human behavior plays a critical role in determining whether an article becomes popular or not. Factors such as emotions, personal preferences, opinions, social networks, and cultural influences affect whether people share and engage with the content on social media platforms.

As a result, developing an effective tool to predict the popularity of an article online requires analyzing multiple characteristics and factors, including text analytics, sentiment analysis, machine learning algorithms, and social media metrics. This tool can help journalists, advertisers, content providers, and other professionals to understand the potential impact of their content, optimize their marketing strategies, and increase their revenue.

The media industry is facing the challenge of accurately predicting the popularity of news articles and blog posts in the online space. This is crucial because the number of shares an article receives on social media platforms directly impacts the revenue generated through advertising. Therefore, developing an effective tool to forecast the popularity of an article online can help businesses optimize their marketing strategies, increase their reach, and generate more revenue. The tool must take into account various factors that affect the popularity of an article, including text analytics, sentiment analysis, machine learning algorithms, and social media metrics. By analyzing these factors, businesses can gain valuable insights into their target audience's preferences and interests, optimize their content marketing strategies, and create more engaging and relevant content that drives more shares and engagement. Furthermore, understanding the complex interactions between human behavior and online content is essential for accurately forecasting the popularity of an article. Therefore, businesses must continuously refine and update their forecasting tools to keep up with the latest trends and changes in human behavior on social media platforms. Overall, developing an effective tool to predict the popularity of an article online can help businesses in the media industry stay ahead of the competition, increase their revenue, and provide more value to their target audience.

# Chapter 7: Data Understanding

The well-known characteristics, such as the number of keywords and the blog post’s genre, were simple to comprehend. The main challenge was the inability to initially comprehend the analytics information provided in the data set, such as LDA, polarity, and subjectivity, which necessitated additional knowledge of text mining techniques, which was briefly inferred from "Large scale sentiment analysis for news and blogs." When the data were thoroughly examined, it became clear that the news might either be unimportant and shared by no more than one person, or it may be viral and shared by more than half a million individuals. So, while determining the following steps, the diversity of the data was also an important consideration.

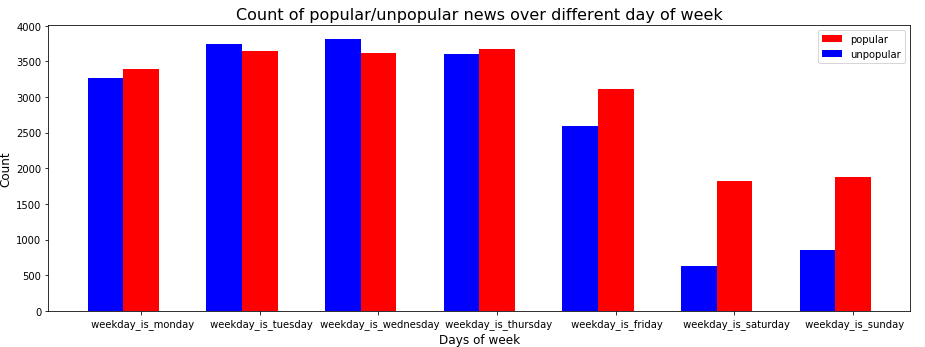


Fig. 7.1: Count of popular/unpopular news over different days of a week

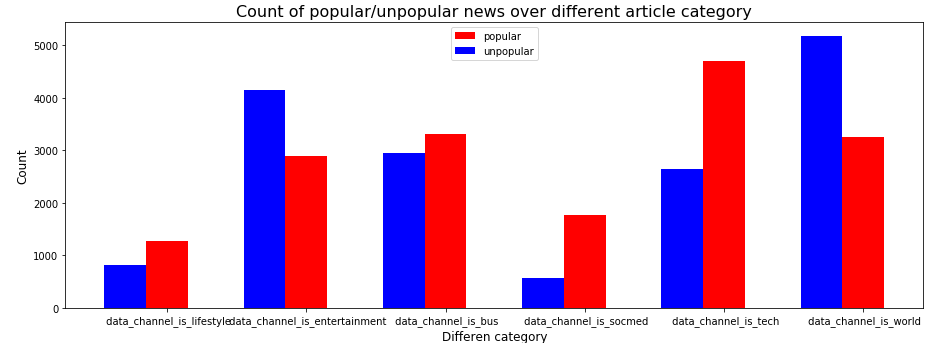


Fig. 7.2: Count of popular/unpopular news over different article categories

Several features, such as the day of the week and the article category, seem to be important. Figure 7.1 displays a graph showing the number of days of the week that particular news stories were the most/least popular. The potential readership of content published over the weekend is higher. This makes sense, as it's probable that weekend readers will spend more time than usual perusing the news online. In Fig. 7.2, we see a scatter plot showing the number of positive and negative news stories across all article types. While the proportion of popular news to unpopular news is much larger in the category of technology ("data channel is tech") and the opposite is true in the categories of the world ("data channel is the world"), social media ("data channel is socmed (social media)"), and entertainment ("data channel is entertainment"). This might indicate that Mashable's audience is more interested in technology and social media than in global and entertainment news.

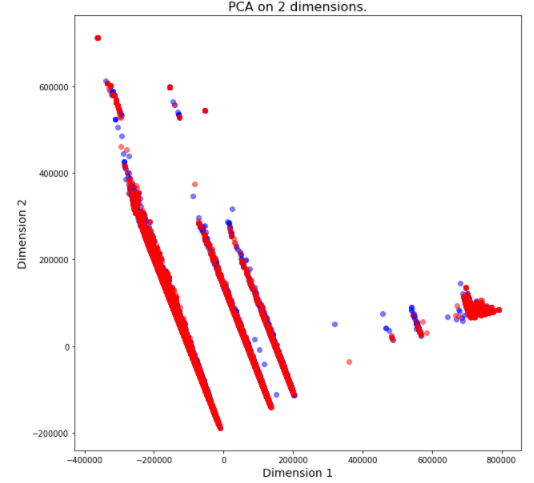


Fig. 7.3.1: Data projected on the first two principle Components.

Chart, radar chart

Description automatically generated

Fig. 7.3.2: Data projected on the first three Principle components (Red: popular; Blue: unpopular)

To better understand the information, PCA will be used. Figure 7.3.1 illustrates how to map a data point into the first two and third components. PCA space does not allow for linear separation of the dataset. Red indicates high popularity and blue indicates low popularity for the data projected on the first two principal components, as illustrated in Fig. 7.3.1 and 7.3.2.

The objective of the data analysis is to predict the popularity of news articles based on various characteristics and factors. The dataset contains news articles published on Mashable, an online news portal, and includes features such as the number of shares, the article category, the day of the week when the article was published, and text analytics metrics such as LDA, polarity, and subjectivity.

The data analysis revealed that predicting the popularity of an article is challenging and depends on various factors such as the quality and relevance of the content, the writing style, the format of the article, and the target audience. Additionally, human behavior plays a critical role in determining whether an article becomes popular or not. The behavior of individuals on social media, including their preferences and interests, significantly influences the popularity of news articles.

The data analysis also revealed that certain features such as the day of the week when the article was published and the article category are relevant in predicting the popularity of an article. Articles published over the weekends have a higher potential to be popular, as people tend to spend more time browsing news online during weekends. The category of technology and social media has a higher proportion of popular news, while the category of world and entertainment has a higher proportion of unpopular news.

PCA was performed to visualize the data and revealed that the dataset is not linearly separable in PCA space. The data projection on the first two and three principal components showed that popular and unpopular news articles are scattered across the space and do not have a distinct separation.

In addition to the features mentioned above, the analysis also includes the use of text mining techniques, such as LDA, polarity, and subjectivity, to gain a better understanding of the data. The LDA technique was used to identify the topics discussed in the news articles, while the polarity and subjectivity analysis helped to determine the sentiment and tone of the articles.

One of the main challenges in analyzing the data was the diversity of the news articles, which ranged from unimportant to viral. This required careful consideration of the various features and factors that could influence the popularity of the articles.

Two significant features that were found to have an impact on the popularity of the news articles were the day of the week and the article category. As shown in Fig. 7.1, articles published over the weekends had a higher potential to be popular, likely due to people spending more time browsing news online on their days off. Additionally, in the category of technology and social media, the proportion of popular news was much larger than the unpopular ones, while in the categories of world and entertainment, the proportion of unpopular news was higher than popular ones. This suggests that Mashable readers prefer news related to technology and social media over news related to the world and entertainment.

To visualize the data, Principal Component Analysis (PCA) was performed, as shown in Fig. 7.3.1 and 7.3.2. The data points were projected onto the first two and three principal components, respectively, and colored based on their popularity. It was found that the dataset was not linearly separable in PCA space, indicating that other techniques might be needed to improve the accuracy of popularity prediction.

Overall, the analysis highlights the importance of considering various features and factors when predicting the popularity of news articles. By leveraging text mining techniques and other data analysis tools, journalists, advertisers, content providers, and other professionals can gain insights into the potential impact of their content and optimize their marketing strategies to increase revenue.

# Chapter 8: Data Preparation

Data Cleaning Dataset has been thoroughly checked using various summarizing techniques for identifying 1. Inconsistent data 2. Inappropriate data 3. Missing data. It’s observed that all the data types i.e., the variable types are consistent. Numeric fields are filled with only numeric values and there are no irrelevant negative values. While exploring the data summary of the input variables few irrelevant values in the n\_tokens\_content attribute are found. The value for n\_tokens\_content is 0 for a few records. With the business understanding, the number of words in the news content can never be 0. Upon checking the news post on the website with the URL provided in the dataset, it’s observed that the rows with value 0 for n\_tokens\_content are misleading and irrelevant data. So, it is excluded and considered missing data.

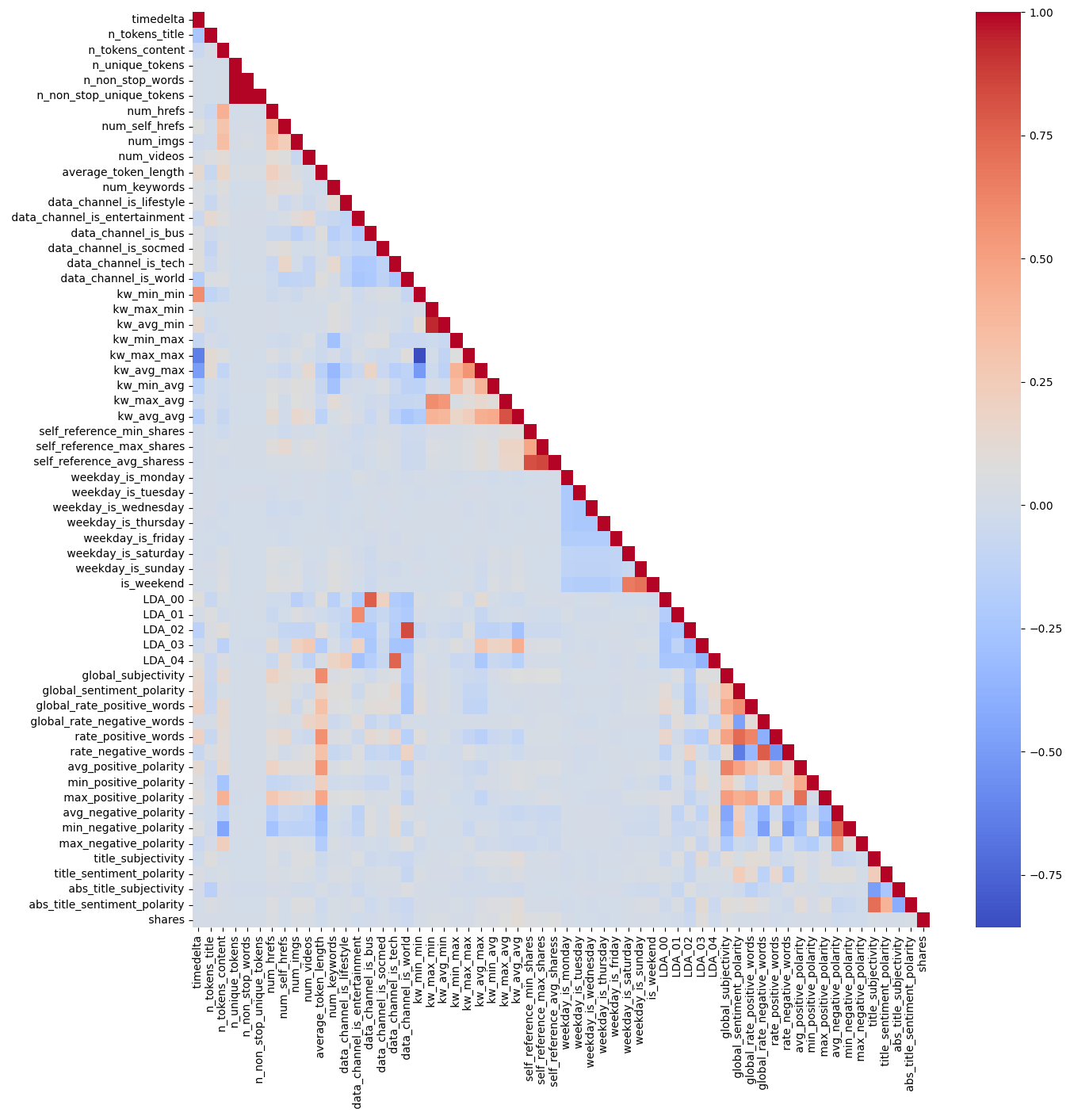


Fig. 8.1: Correlation between default features

As shown in Fig. 8.1, Applied a cluster map to examine the correlation among variables. This seems that some of them have a high correlation. However, correlated information between features is limited. Distribution of weekdays showed that the numbers of shares are highest during the middle of the week and decreased during the weekends; Tuesdays, Wednesdays, and Thursdays are higher than the rest of the days. After some analysis, it is observed that a weekend article has less opportunity to be shared as an article.

# Chapter 9: Modeling

In the previous part, we identified the dependent variables after cleaning the data by eliminating outliers and extraneous information. The correlation matrix is constructed to reveal the associations between the dependent variables and the dependent variable of interest. I looked at the heatmap of feature correlation and found that the majority of the characteristics are independent of one another. In addition, there are over 1400 possible values for the target variable, making accurate prediction difficult. To determine which of several popularity tiers (most popular, most popular, least popular, viral, etc.) a given post falls into, the target variable is binned. A regression model was applied to the problem space and data before the target variable was binned into categories.

Then, explore the relationship between the target (number of shares) and some selected features:

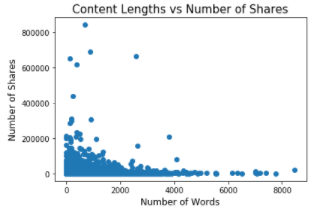


Fig. 9.1: Content length vs Number of shares.

As shown in Fig. 9.1, the Comparison is between several words in the content vs of several shares. It is showing that if the content is big, then there is less chance of sharing whereas if the content size is a short number of users will read and share the news.

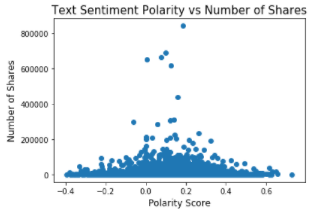


Fig. 9.2: Text sentiment popularity vs Number of shares

As shown in Fig. 9.2, If the content is having sentiments, that will attract the users and will allow users to share also. Initially, regression models were used to predict the number of shares of a piece of news. All features except for news URL and timedelta were included in the model. StandardScaler() was applied to numerical features.

TABLE 9.1: REGRESSION MODEL ERROR COMPARISON

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **R2** | **Mean Absolute Error** | **Median Absolute Error** |
| **LassoCV** | 0.021390 | 3039.425129 | 1675.728808 |
| **RidgeCV** | 0.019512 | 3053.470410 | 1662.516891 |
| **RandomForestRegressor** | -0.070401 | 3456.761696 | 1584.380000 |

Regression models produced substantial median absolute errors, as shown in Table 9.1, suggesting that classification models might be better suited to this issue. The model includes all features with the exception of the news URL and the timedelta. The numerical characteristics were scaled using StandardScaler(). Each categorisation algorithm was given its own pipeline.

Choosing the threshold -- Mean (3395) or Median (1400)

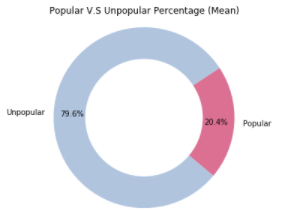
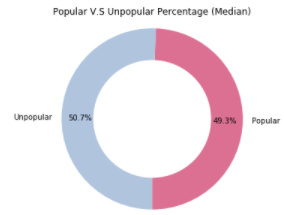


Fig. 9.3: Comparisons of popular vs unpopular percentages in Mean and Median

As shown in Fig. 9.3, If the mean number of shares (3395) is used as the threshold, the popular and unpopular classes would be imbalanced. Additionally, the large range of several shares would make the mean not appropriate to divide the news into two classes. Therefore, decided to choose the median number of shares (1400) as the threshold. LogisticRegressionCV, KNeighborsClassifier, Naive Bayes (GaussianNB), and RandomForestClassifier are the classifiers utilised in this study.

TABLE 9.2: CLASSIFIER COMPARISON BETWEEN FOUR MODELS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **accuracy** | **precision** | **recall** | **f1 score** |
| **LogisticRegressionCV** | 0.641569 | 0.644345 | 0.620777 | 0.641405 |
| **KNeighborsClassifier** | 0.595788 | 0.606519 | 0.529337 | 0.593988 |
| **GaussianNB** | 0.533611 | 0.672938 | 0.118110 | 0.437456 |
| **RandomForestClassifier** | 0.663640 | 0.666842 | 0.644653 | 0.663509 |

As shown in Table 9.2, Accuracy is 67% with a 0.67 F1 Score which shows better results for the Random forest classifier among all the

**Accuracy:** (9.1)

**Precision:**  (9.2)

**Recall:** (9.3)

**F1 score:** (9.4)

- F1-score is used when False Negatives and False Positives are crucial

- F1-score is a better metric when there are imbalanced classes

The logistic regression model predicts the two-category results with a balanced accuracy of 64%. The accuracy curve for different cut-offs is as below, so to maximize the accuracy the threshold limit is chosen as 0.5.

News articles include results from text mining and sentiment analysis and tried to create a Naive Bayes Classifier for this data set. The observed accuracy and f-score are not that satisfactory. So, concluded that the Naive Bayes Model does not fit the data set as expected. The failure can be attributed to the naive assumption about class conditional independence. It is possible that the features in the data set are not independent

The random forest model is built with the same set of input variables, testing set, and training set which was used for logistic regression to understand the accuracy of the models. The random forest model is initially set with 500 trees and 7 splits. Later, confirming the right model, the number of trees, and the number of splits will be tuned to find the optimized results.

Using the F1 score, Random Forest Classifier is the best model.

# Chapter 10: Model Evaluation

The article's headline may benefit from some sentimental examination. The capacity to use unstructured data for meaningful insights is crucial for managing a successful article using emotional data.

To determine whether an author is being subjective or objective, or even positive or negative, as seen in Fig. 10.1, sentimental analysis employs Natural Language Processing algorithms.

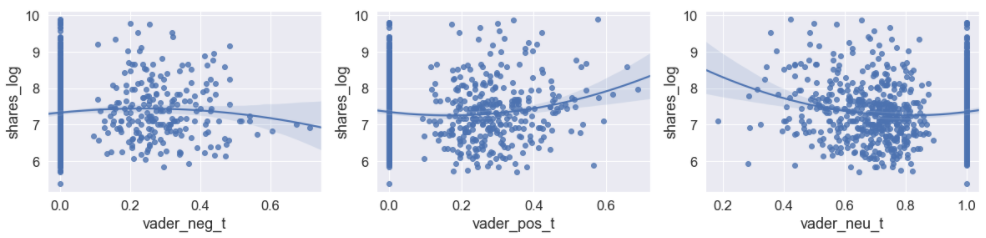


Fig. 10.1: Comparison of subjective information using Sentimental analysis in Positive, Negative, and neutral

As shown in Fig. 10.1, shows the comparison of subjective information using Sentimental analysis in Positive, Negative, and neutral which highlights, that data having positive and neutral sentiments will be viewed more and shared.

Surprisingly, there is a positive correlation between social shares of news stories and a score positive in the title: Between 20-40% shown in Fig. 10.2 of those who score positive in the title are shared. (Note - the logarithmic scale for social shares).

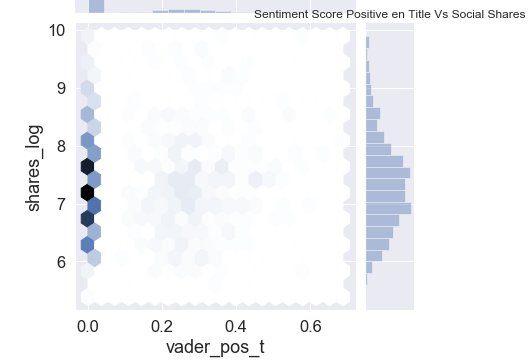


Fig. 10.2: Correlation between sentiment score positive on title vs social shares

As shown in Fig. 10.2, Here’s a correlation between positive sentiments and the number of shares.

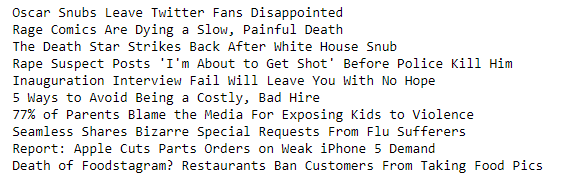


Fig. 10.3: Top 10 Negative features by VADER score

As shown in Fig. 10.3, Using the VADER library, here’s a ranking of the 10 titles with negative scores by the number of social shares of those articles. Abusing language and violent words like kill reflect negative phrases.

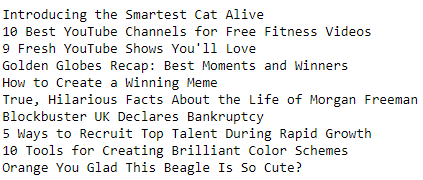


Fig. 10.4: Top 10 Positive features by VADER score

As shown in Fig. 10.4, Here’s a ranking of the 10 titles with negative scores by the number of social shares of those articles. There are 4 records that include numbers. Also, one of them has a video. It seems that strong visuals and including numbers in the title create the biggest impact to share.

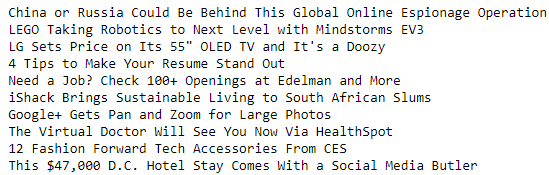


Fig. 10.5: Top 10 Neutral features by VADER

As shown in Fig. 10.5, Here’s a ranking of the 10 titles with neutral scores by the number of social shares of those articles. This section generally offers products, job offers and news related to that is reflected.



Fig. 10.6: The word cloud shows that specific keywords may be more informative and technological

The word cloud as shown in Fig. 10.6 shows that specific words should be more informative, descriptive, and technological. This suggests that the influencing factors that result in the title of an article being popular and experiencing the highest volume of media sharing are ones that: have the placement of a single number in the title, contain strong visual imagery such as videos, and convoy positive messaging.

# Chapter 11: Analysis and Results

The best performance is also given by the RF model as shown in Table 11.1, which achieves a 0.67 of 10 accuracy score, and 0.67 for the F1 score. The metrics of my RF model are an accuracy of 0.663 and an F1-score of 0.663. By comparison, the accuracy and F1 score are all better than the benchmark model. So obtained model achieves comparable performance to the benchmark model and it is significant enough to solve the popular news classification problem.

TABLE 11.1: RF MODEL FOR NEWS CLASSIFICATION

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **accuracy** | **precision** | **recall** | **f1 score** |
| **LogisticRegressionCV** | 0.641569 | 0.644345 | 0.620777 | 0.641405 |
| **KNeighborsClassifier** | 0.595788 | 0.606519 | 0.529337 | 0.593988 |
| **GaussianNB** | 0.533611 | 0.672938 | 0.118110 | 0.437456 |
| **RandomForestClassifier** | 0.663640 | 0.666842 | 0.644653 | 0.663509 |

**Sentimental Analysis**

1. VADER library used to get better sentiment scores.
2. Sentiment analysis with VADER 2. In addition to being effective on texts from other domains, our lexicon- and rule-based sentiment analysis tool is finely suited to the emotions conveyed in social media.
3. Parsed the text With SentimentIntensityAnalyzer.

# Chapter 12: Conclusions and Future Scope

For this project, initially tried to predict the number of shares of a piece of news with regression models. Fit three algorithms, Lasso, Ridge, and Random Forest Regressor, and found that the Median Absolute Error (MedAE) was around 1600, which is very high considering that some news only has a few hundred shares.

Treated this problem as classification and used the median number of shares (1400) as a threshold to divide the news into two classes, popular and unpopular. Fit four algorithms, Logistic Regression, K-Nearest Neighbours, Gaussian Naive Bayes, and Random Forest Classifier, and found that Random Forest Classifier has the best performance using the F1 score as the North Star Metric.

Now tuned the hyperparameters of the random forest classifier based on cross-validation scores since grid/random search was too slow for this dataset. In final model had an F1 score of 0.66 and an accuracy of 66% for the test data.

The business problem was to predict the reach/popularity of the news article. After multiple levels of cleaning and pre-processing the data are stabilized for model building. Since the linear model couldn’t produce better results because of the variance in the data, various numbers of bins are used, and classification algorithms are applied. Upon analyzing various models, the suited random forest model is fine-tuned. Though 2 categories classification provides better and more relevant results, it assumes popularity as a definitive output rather than a ranking methodology. In the next steps, instead of categorizing the news article, a ranking mechanism can be built using the bag of words and other text mining, and clustering methodologies. The ranking methodology can be improved over the period using reinforcement learning by adding the words from popular articles to the bag of words.

**Future Enhancements**

To further improve the performance, the below 3 steps could be used:

(1) Increase the size of the dataset since RF has a strong learning capability and a rich dataset might improve its prediction performance.

(2) Try more advanced cross-validation methods although it might increase the training time.

(3) Engineer and add more relevant features to the original dataset. Users will use all the words in an article as additional features, and then try the classifier such as Naive Bayes to see if it can achieve better performance.

# Bibliography

[1] (UCI Machine Learning Repository: Online News Popularity Data Set, n.d.)

[2] Wicaksono, A. S., and Supianto, A. A. (2018). Hyper Parameter Optimization using Genetic Algorithm on Machine Learning Methods for Online News Popularity Prediction. *IJACSA) International Journal of Advanced Computer Science and Applications*, *9*(12). [www.ijacsa.thesai.org](http://www.ijacsa.thesai.org)

[3] Hensinger, E., Flaounas, I., and Cristianini, N. (2012). Modeling and predicting news popularity. *Pattern Analysis and Applications 2012 16:4*, *16*(4), 623–635. https://doi.org/10.1007/S10044-012-0314-6

[4] Rizos, G., Papadopoulos, S., and Kompatsiaris, Y. (2016). Predicting News Popularity by Mining Online Discussions. *WWW 2016 Companion - Proceedings of the 25th International Conference on World Wide Web*, 737–742. <https://doi.org/10.1145/2872518.2890096>

[5] Godbole, N., Srinivasaiah, M., and Skiena, S. (n.d.). *Large-Scale Sentiment Analysis for News and Blogs*. Retrieved August 19, 2022, from http://www.textmap.com/sentiment.

[6] Kleis Nielsen Alessio Cornia Antonis Kalogeropoulos, R. (n.d.). *Challenges and opportunities for news media and journalism in an increasingly digital, mobile, and social media environment*.

[7] *More Americans Are Getting Their News From Social Media*. (n.d.). Retrieved August 19, 2022, from https://www.forbes.com/sites/petersuciu/2019/10/11/more-americans-are-getting-their-news-from-social-media/?sh=5794a8c43e17

[8] Mukherjee, K. (n.d.). *Advancement of Social Media and Future of Newspaper Industry*. Retrieved August 21, 2022, from www.caluniv.ac.in

[9] Namous, F., Rodan, A., and Javed, Y. (2019). Online News Popularity Prediction. *ITT 2018 - Information Technology Trends: Emerging Technologies for Artificial Intelligence*, 180–184. <https://doi.org/10.1109/CTIT.2018.8649529>

[10] *Online News Feed Data Mining and Prediction*. (n.d.). Retrieved August 21, 2022, from https://1library.net/document/ozloje6z-online-news-feed-data-mining-and-prediction.html

[11] *Predicting Online News Popularity*. (n.d.). Retrieved August 21, 2022, from https://1library.net/document/q5mkpvgy-predicting-online-news-popularity.html

[12] *Predicting Online News Popularity (Part 1) | by Benjamin Dornel | Towards Data Science*. (n.d.). Retrieved August 21, 2022, from https://towardsdatascience.com/predicting-online-news-popularity-part-1-aae9a4f7f1a4

[13] Ren, H., & Yang, Q. (n.d.). *Predicting and Evaluating the Popularity of Online News*.