MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Project Report

on

Popularity analysis of Articles

A project report submitted in partial fulfilment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

in

IT-IOT

Submitted by:

Aditya Joshi

0901IO201005

Harsh Rane

0901IO201031

Faculty Mentor:

Prof. Abhilash Sonkar

Assistant Professor, Department of Information Technology

DEPARTMENT OF INFORMATION TECHNOLOGY
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005

JAN-MAY 2023

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

CERTIFICATE

This is certified that **Aditya Joshi(0901IO201005), Harsh Rane(0901IO201031)** have submitted the project report titled **Popularity analysis of Articles** under the mentorship of **Prof. Abhilash Sonkar**, in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in **Internet of Things** from Madhav Institute of Technology and Science, Gwalior.

Prof. Abhilash Sonkar
Assistant Professor
Information Technology

Dr. Akhilesh Tiwari
Professor and Head,
Department of IT

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Internet of Things at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of **Prof Abhilash Sonkar**, **Assistant Professor**, Department of Information Technology.

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Date:
Place: Gwalior

Aditya Joshi
0901IO201005
III Year
IT-IOT

Harsh Rane
0901IO201031
III Year
IT-IOT

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Information Technology,** for allowing me to explore this project. I humbly thank **Dr. Akhilesh Tiwari**, Professor and Head, Department of Information Technology, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of Prof. **Abhilash Sonkar**, Assistant Professor, Department of Information Technology, for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Aditya Joshi
0901IO201005
III Year
IT-IOT

Harsh Rane
0901IO201031
III Year
IT-IOT

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

As the world goes through an internet boom, every day, more people from developing nations are turning to the internet for News. As a result, the prediction of online news popularity is becoming a trendy research topic. In this project, we aimed at the descriptive and predictive analysis of the popularity of News articles on the internet. The goal is to analyse trends and patterns among the popular News articles and predict the likelihood of the news article to be popular before their publication. The primary dataset is from Mashable.com. This is then extended to articles from Medium.com, along with a similar prediction pipeline.

Keywords: Online news popularity; Predictive analysis; Likelihood.

CHAPTER 1: INTRODUCTION

With the Internet expanding at an impressive pace, there has been a growing preference for online news, which is currently the fastest means of information spread across the world. With news spanning multiple genres and various websites and all websites wanting to cover the most "happening" news and go viral. For online news websites and other content providers or advertisers, it is crucial to predict the popularity of the news articles before its publication. Thus, it is logical and meaningful to use machine learning techniques to predict the popularity of online news articles. Here, the popularity of the article is in terms of the number of shares. This analysis will also help news and content writers on how an article should be written to make it popular.

1.1 Objectives and scope

The objective of newspaper article popularity analysis is to understand which articles are gaining the most attention and traction among readers. By analyzing the popularity of articles, publishers and editors can gain insights into the preferences and interests of their audience, and make data-driven decisions about the type of content they should produce and promote.

1.2 System Requirements

The project ran on a system having the following system requirements:

- 2.4GHz Intel i5-9300h 9th Gen processor
- 8GB DDR4 RAM
- 1TB 7200rpm hard drive
- NVIDIA GeForce GTX 1650 4GB Graphics

CHAPTER 2: LITERATURE REVIEW

[1] The study proposes a machine learning-based approach for predicting the popularity of online news articles using features such as article category, number of images and videos, and social media activity. The proposed model achieved an accuracy of 85.5% in predicting the popularity of news articles.

[2] This study focuses on predicting the popularity of online news articles using content metadata, such as the number of words, images, and videos. The authors use a decision tree-based model and achieve an accuracy of 81.5% in predicting the popularity of news articles.

[3] This study proposes a method for predicting the popularity of online articles based on user comments. The authors use features such as comment count, sentiment analysis of comments, and user engagement to predict the popularity of articles. The proposed model achieved an accuracy of 79% in predicting the popularity of articles.

[4] The study proposes a model for predicting the popularity of news articles using a combination of content-based features and social network analysis. The authors use features such as article length, sentiment analysis, and number of social media shares to predict the popularity of articles. The proposed model achieved an accuracy of 81.9% in predicting the popularity of articles.

[5] This study focuses on predicting the message propagation in Twitter. The authors use a machine learning-based approach and features such as message content, user influence, and social network structure to predict the propagation of messages. The proposed model achieved an accuracy of 79.9% in predicting the propagation of messages.

CHAPTER 3: ANALYSIS PIPELINE

We start with some exploratory Data Analysis. We have different plots, mainly between no. of Claps, Publication, Sentiment, Day of the week and number of Tokens. For the classification task, we implemented three classification algorithms, which are, Logistic Regression, SVC with Gaussian kernel and Random Forest. We tuneour models with different value of hyper-parameter to find the model with the highest accuracy.

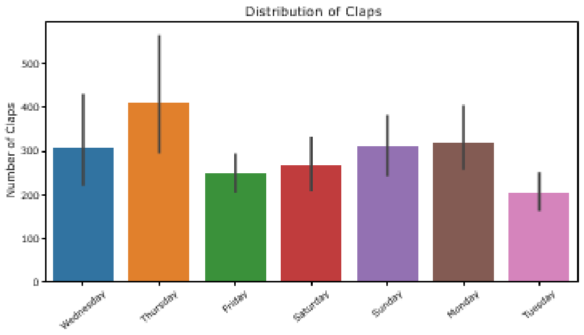


Fig 3.1

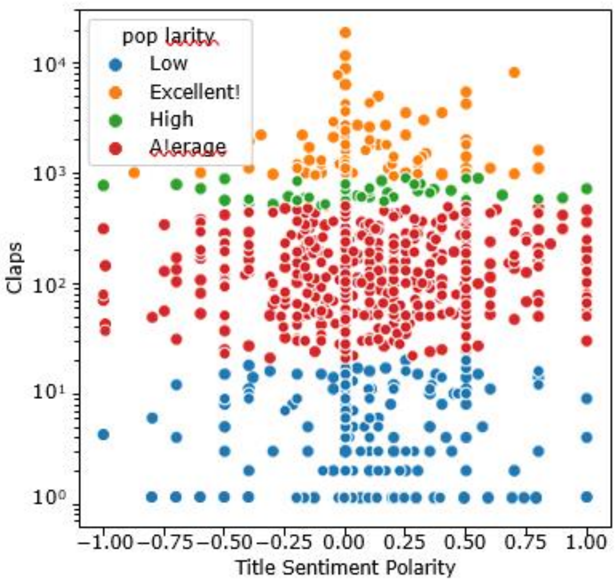


Fig 3.2

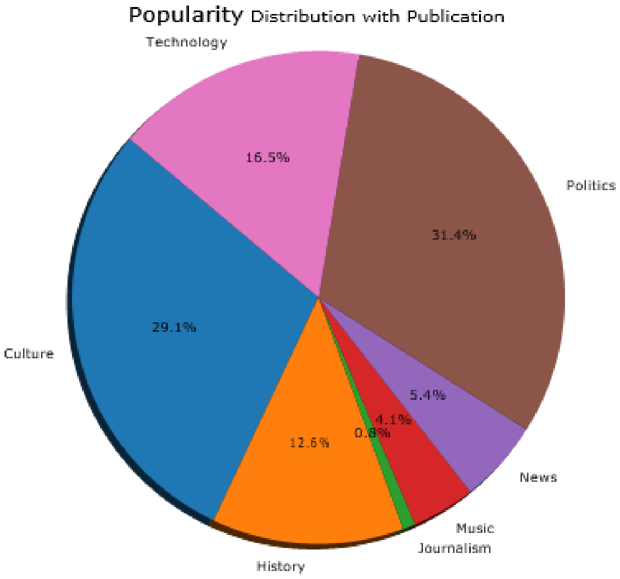

Popularity Distribution with Publication
Technology
16.5%
Politics
31.4%
Culture
29.1%
5.4%
4.1%
0.8%
12.6%
News
Music
Journalism
History

Fig 3.3

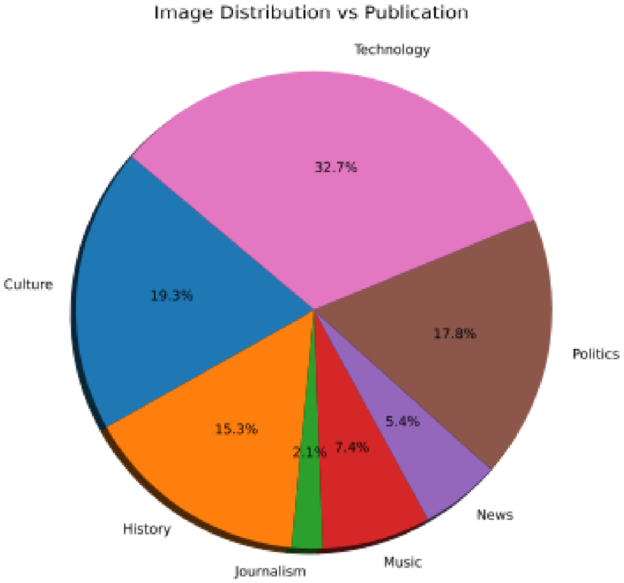
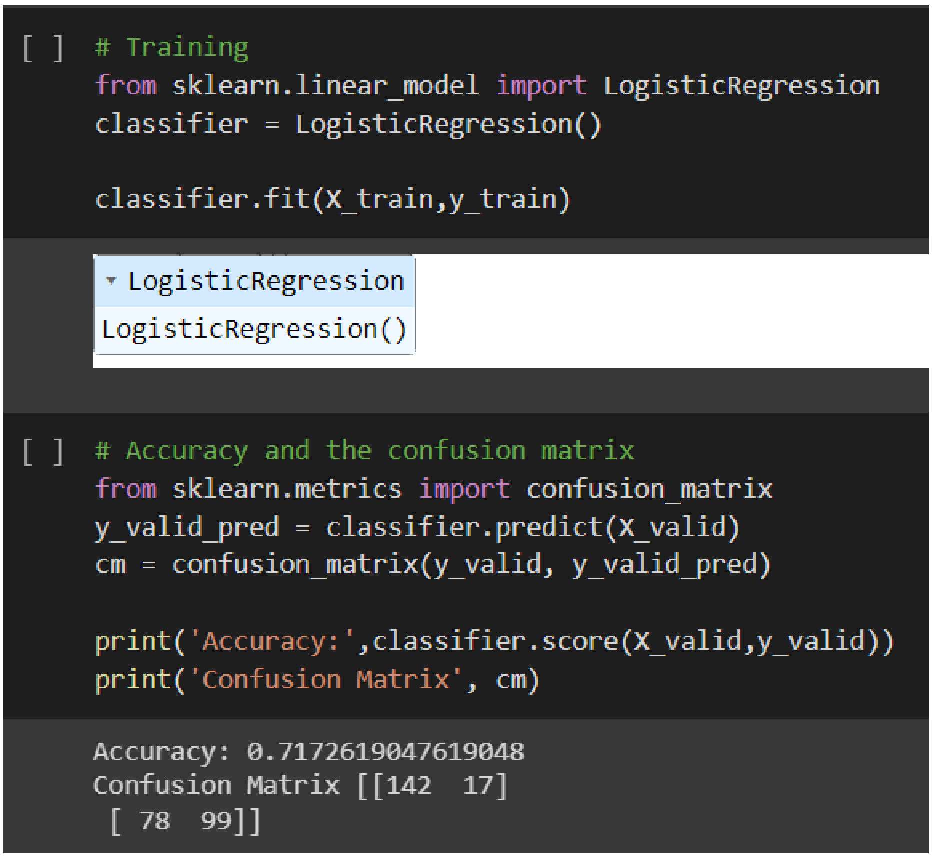


Image Distribution vs Publication
Technology
32.7%
Culture
19.3%
17.8%
Politics
5.4%
15.3%
2.1%
7.4%
News
History
Music
Journalism

CHAPTER 4: RESULTS

4.1 Result

Figure 4.1 shows that for the logistic regression model, the accuracy is best with having regularisation parameter equal to 1, i.e. by setting regularisation parameter as zero. The accuracy achieved using logistic regression is 71.5 % . Figure 4.2 shows that for the SVC model with gaussian kernel, the accuracy increases with the regularization parameter and then become constant after 10. The accuracy achieved using the SVC(with the gaussian kernel) is 70 % . Figure 4.3 shows that for the Random Forest model, the accuracy keeps increasing with the increase in the number of trees. The accuracy achieved with Random Forest, using 300 trees is 78 %. Overall, Random Forest gives the best accuracy, therefore it is the best algorithm to model this classification problem.



```
[ ] # Training
    from sklearn.linear_model import LogisticRegression
    classifier = LogisticRegression()

    classifier.fit(X_train,y_train)
```

```
▾ LogisticRegression
LogisticRegression()
```

```
[ ] # Accuracy and the confusion matrix
    from sklearn.metrics import confusion_matrix
    y_valid_pred = classifier.predict(X_valid)
    cm = confusion_matrix(y_valid, y_valid_pred)

    print('Accuracy:',classifier.score(X_valid,y_valid))
    print('Confusion Matrix', cm)
```

```
Accuracy: 0.7172619047619048
Confusion Matrix [[142  17]
 [ 78  99]]
```

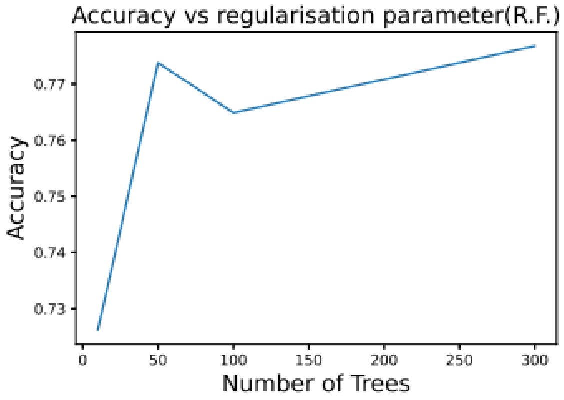
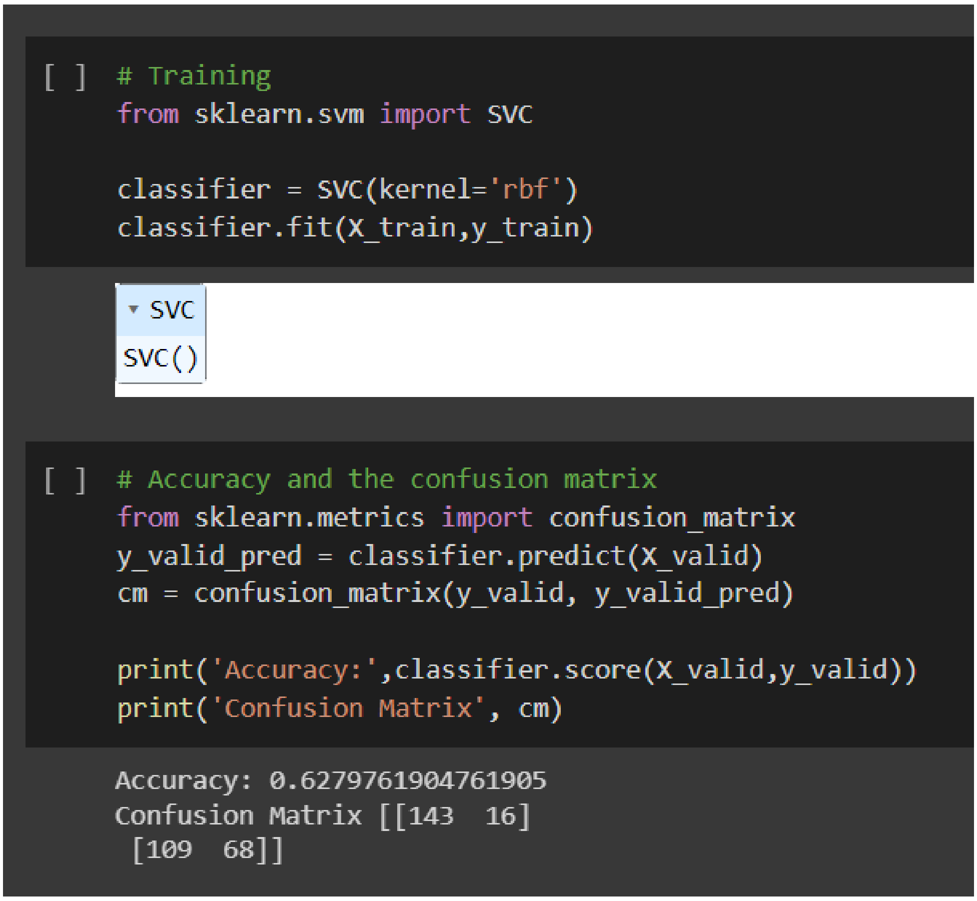


Fig 4.1



```
[ ]  # Training
     from sklearn.svm import SVC

     classifier = SVC(kernel='rbf')
     classifier.fit(X_train,y_train)
```

```
SVC()
```

```
[ ]  # Accuracy and the confusion matrix
     from sklearn.metrics import confusion_matrix
     y_valid_pred = classifier.predict(X_valid)
     cm = confusion_matrix(y_valid, y_valid_pred)

     print('Accuracy:',classifier.score(X_valid,y_valid))
     print('Confusion Matrix', cm)
```

```
Accuracy: 0.6279761904761905
Confusion Matrix [[143  16]
 [109  68]]
```

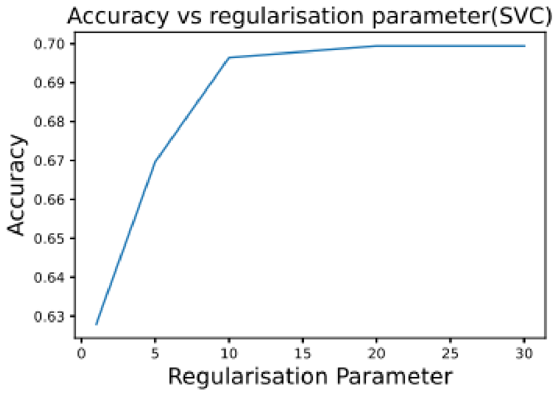
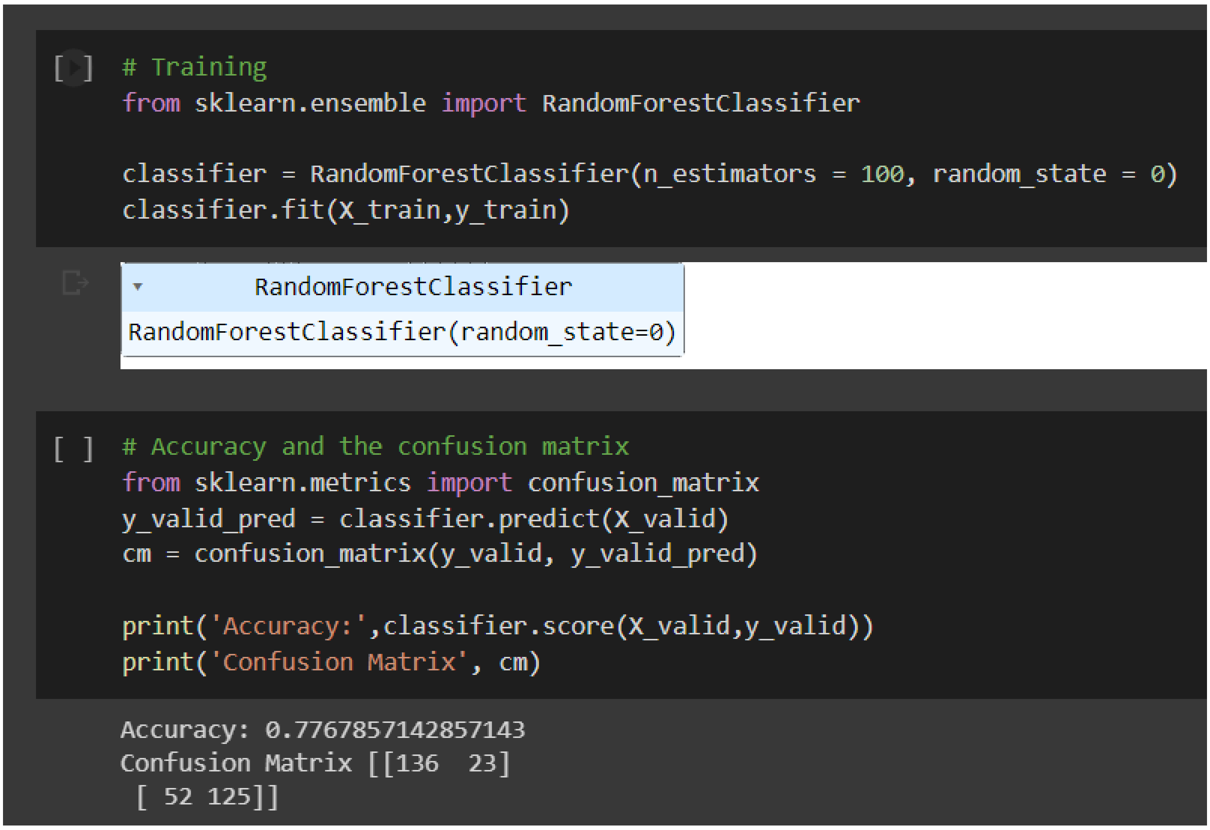


Fig 4.2



```
[ ] # Training
    from sklearn.ensemble import RandomForestClassifier

    classifier = RandomForestClassifier(n_estimators = 100, random_state = 0)
    classifier.fit(X_train,y_train)
```

```
▾        RandomForestClassifier
RandomForestClassifier(random_state=0)
```

```
[ ] # Accuracy and the confusion matrix
    from sklearn.metrics import confusion_matrix
    y_valid_pred = classifier.predict(X_valid)
    cm = confusion_matrix(y_valid, y_valid_pred)

    print('Accuracy:',classifier.score(X_valid,y_valid))
    print('Confusion Matrix', cm)
```

```
Accuracy: 0.7767857142857143
Confusion Matrix [[136  23]
 [ 52 125]]
```

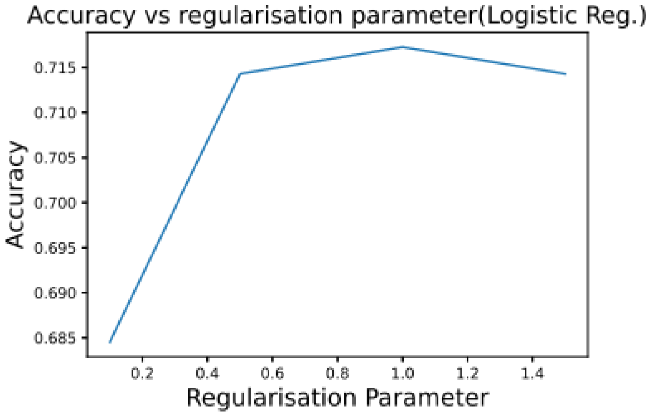


Fig 4.3

4.2 Discussions

In Figure 3.1, we see that most of the articles are published on Thursday, in anticipation of the Weekend. From Figure 3.2, we see that most of the popular articles portray a neutral sentiment. In Figure 3.3, we see that Culture and Politics are the most popular topics. From Figure 13, we see that Technological articles use a lot of images, whereas, Journalism is almost image free as compared to other categories. In Figure 3.4, we see that Excellent articles have, on average, roughly, 50 more words than others. Number of claps and number of responses has a high correlation. That makes sense as more the number of people who like the article, more are they likely to respond in the form of a comment. The correlation amongst different tokens is also high - no. of tokens, no. of unique tokens, no. of non-stop words. Tokens also have a high correlation with the reading time. This makes sense as more the number of words, the longer the reading time.

4.3 Conclusions:

We have looked at the Online News Popularity Dataset and looked at prediction through classification. Then, we have then, translated these onto a new Dataset that has been scraped from Medium.com. We see similar results in some places, and dissimilar in others, due to the time difference as well as the difference in the websites themselves. One is online news platform, other is an open platform for publishing articles open to all.

References

[1] F. Namous, A. Rodan, and Y. Javed, “Online news popularity prediction,” Nov. 2018, pp. 180–184. DOI: 10.1109/CTIT.2018.8649529.

[2] M. T. Uddin, M. J. A. Patwary, T. Ahsan, and M. S. Alam, “Predicting the popularity of online news from content metadata.”

[3] A. Tatar, J. Leguay, M. D. de Amorim, A. Limbourg, S. Fdida, and P. Antoniadis. (2011). “Predicting the popularity of online articles based on user comments,” [Online]. Available: https:// doi. org/ 10 . 1145 / 1988688 . 1988766. (accessed: 13.12.2020).

[4] E. Hensinger, I. Flaounas, and N. Cristianini. (2013). “Modelling and predicting news popularity,” [On- line]. Available: https : / / www . researchgate . net / publication / 257471988 Modelling and predicting news popularity. (accessed: 13.12.2020).

[5] S. Petrovic, M. Osborne, and V. Lavrenko. (2011). “Rt to win! predicting message propagation in twitter,” [On- line]. Available: https : / / www . aaai . org / ocs / index . php / ICWSM / ICWSM11 / paper / viewFile / 2754 / 3209. (accessed: 13.12.2020).