Online News Popularity Analysis

1st Aruja Khanna
Mechanical Dept.
IIT Bombay
190100024

2nd Neilabh Banzal
Mechanical Dept.
IIT Bombay
170010014

3rd Shashank Singh
Aerospace Dept.
IIT Bombay
190010061

***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.**

I. 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.

A. Previous Work

Various other works have been done in prediction of content popularity. In [1], the authors have compared different ML models on the Online News Popularity Dataset. [2] introduces another approach for prediction of online news popularity. In [3] , the comments of different users are analyzed to predict the popularity of a online article. [4] defines the popularity in terms of a competition where the popular articles are those which were the most visited on that particular day. Support Vector Machine (SVM) is used to classify the popularity/unpopularity of online news article. In [5], the number of retweets is predicted using both the features of the retweet content (length, words, number of hashtag, etc.) and the features of author (number of followers, friends, etc.).

B. Problem Statement

In this project, we use machine learning techniques for a classification problem, which is to predict if an online news article will become popular before its publication. The number of shares is a continuous variable, any article with shares more than the median of the shares is labelled as popular, and the rest are labelled as unpopular. Four machine learning algorithms are implemented and compared, which are Logistic Regression, SVC (with Gaussian Kernel), Random Forest and Neural Network. The best model will be selected based on the metric.

This is done for 2 different datasets, one is the standard Online News Popularity Dataset, and then, for a dataset which has been scraped from *Medium.com*.

Aspects	Features
Words	Number of words of the title/content; Average word length; Rate of unique/non-stop words of contents
Links	Number of links; Number of links to other articles in Mashable
Digital Media	Number of images/videos
Publication Time	Day of the week/weekend
Keywords	Number of Keywords; Worst/best/average keywords(#shares); Article category
NLP	Closeness to five LDA topics; Title/Text polarity/subjectivity; Rate and polarity of positive/negative words; Absolute subjectivity/polarity level
Target	Number of shares at Mashable

TABLE I
Aspects and Features of the Online News Popularity - Mashable

II. Online News Popularity Dataset

The first data set used for this project, is the Online News Popularity Data Set from the University of California Irvine's Machine Learning Repository. The dataset consist of data points on 39644 news articles from an online news website called Mashable collected from Jan 2013 to Jan 2015. For each instance of the dataset, it has 61 attributes which includes 58 predictive attributes, 2 non-predictive attributes, and 1 goal field. A preliminary analysis is already done on these articles to find the different attributes for each article.

III. 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 four classification algorithms, which are, Logistic Regression, SVC with Gaussian kernel, Random Forest and Neural Network. We tune our models with different value of hyper-parameter to find the the model with the highest accuracy.

1) Exploratory and Descriptive:

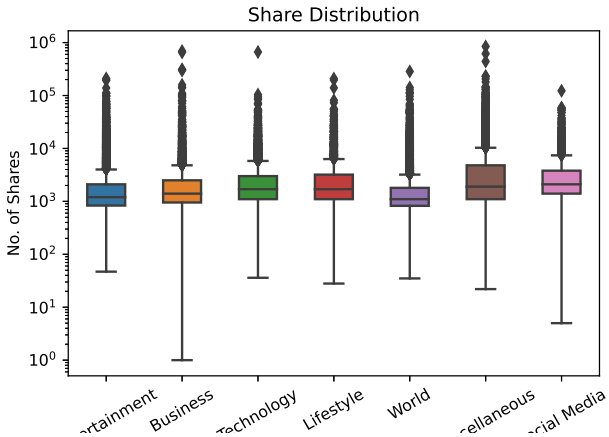

Fig. 1. No. of Shares vs Channels

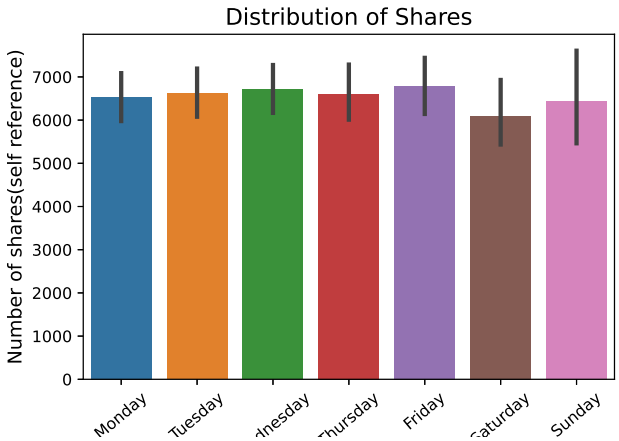

Fig. 2. No. of Shares vs Day of the Week

2) Predictive: For the purpose of classification, the dataset is split into three parts, namely training set to train the model, validation set to tune the model and the test set to find the final accuracy of the model. First, we train our logistic regression model with the training dataset. This is then tuned using different values of the regularisation parameter to maximise the accuracy. The similar process is then done using SVC, Random Forest and the Neural Network. In the end, we get the best model for this classificaiton problem.

IV. Results

Figure 5 shows that for the logistic regression model,the accuracy is best without having any regularisation parameter,

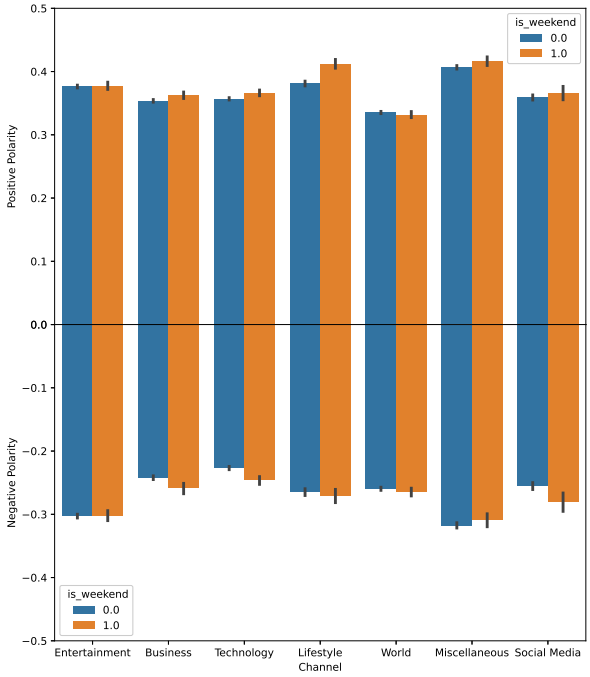

Fig. 3. Polarity vs Channels

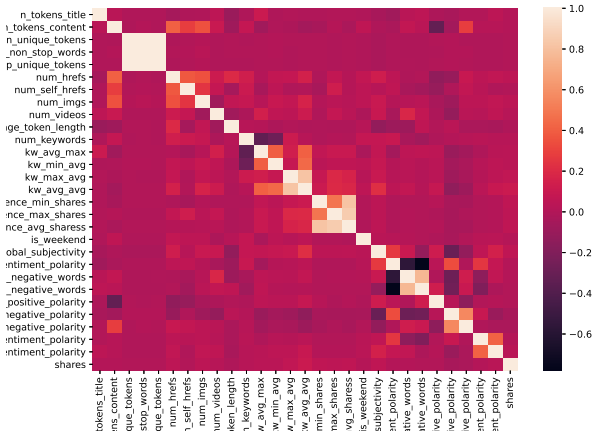

Fig. 4. Heat Map

i.e. by setting regularisation parameter as zero. The accuracy achieved using logistic regression is 66 % . Figure 6 shows that for the SVC model with gaussian kernel, the accuracy is achieved by setting the regularisation parameter as 1. The accuracy achieved using the SVC(with the gaussian kernel) is 67.5 % . Figure 7 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

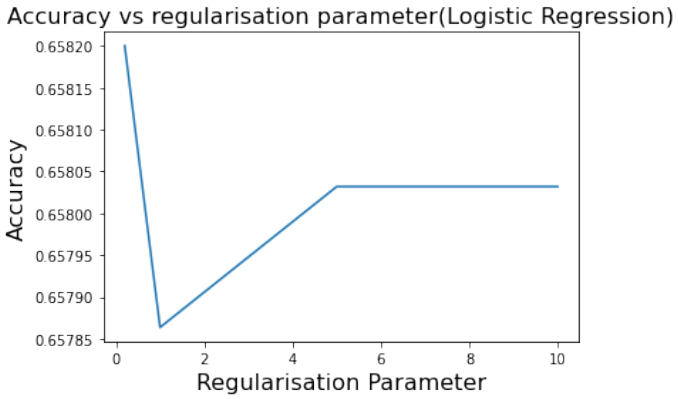

Fig. 5. Accuracy vs Regularisation parameter (Logistic Regression)

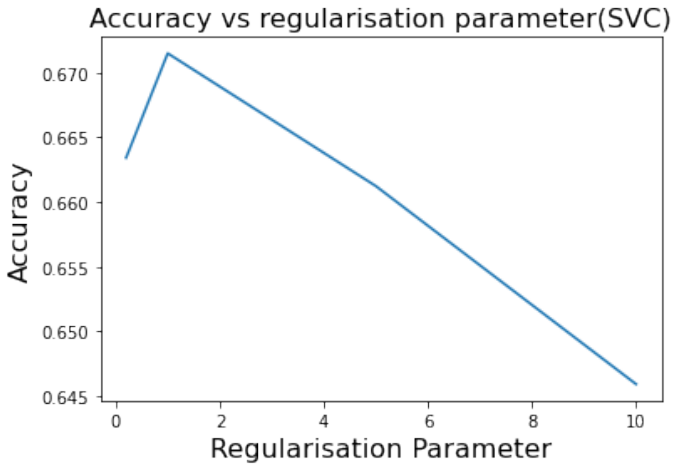

Fig. 6. Accuracy vs Regularisation parameter (SVM)

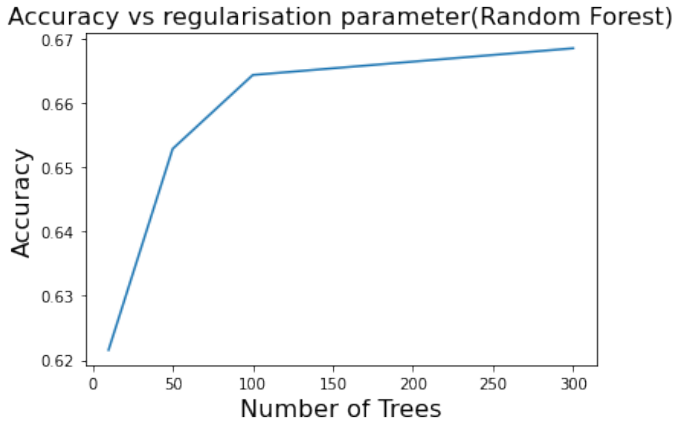

Fig. 7. Accuracy vs Regularisation parameter (Random Forest)

trees is 68 % . Figure 8 shows that for the Neural Network achieves a accuracy of about 65 %. Here, the neural network has one hidden layer and we can see in Figure 9 that fastest learning takes place with layer size equal to 60. Overall, Random Forest gives the best accuracy, therefore it is the best

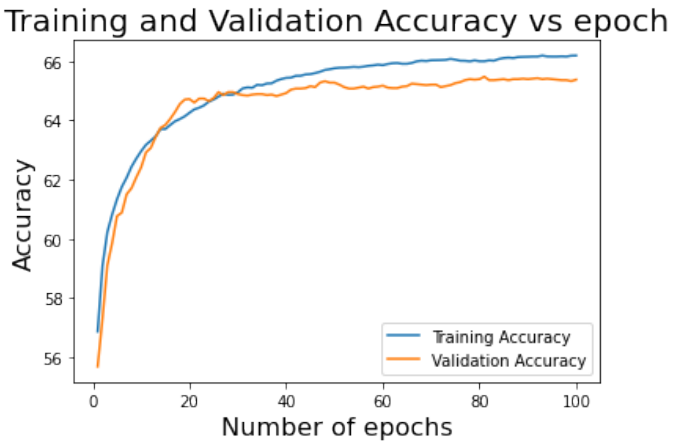

Fig. 8. Training and Validation loss vs epoch

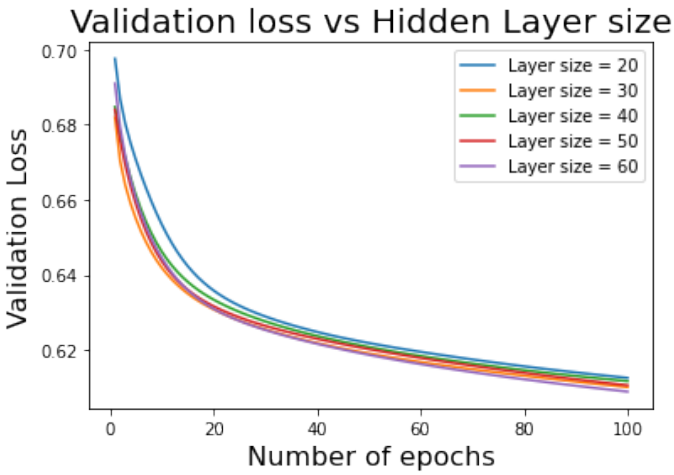

Fig. 9. Training and Validation Accuracy vs epoch

algorithm to model this classification problem.

V. Discussion

In Figure 1, we see that number of shares for almost all channels is of the same order. From Figure 2, we see that weekends have less shares. In Figure 3, we see that more articles are positive than negative. The heat map, as expected, shows high correlation between different tokens.

VI. Medium Articles Dataset (Scraped)

Note that this data on the News Popularity is 4 years old. So, we set out to recreate a similar data set. For this, we used BeautifulSoup4 and requests libraries to scrape data from the website *Medium.com* by using their archives. In total, we scraped 2238 Articles across 7 genres - News, Politics, Culture, Music, History, Journalism and Technology. Then, we used the TextBlob and nltk Libraries to perform various computations like Tokenisation, calculation of number of unique tokens, number of non-stop unique words, etc. to get a similar dataset to the one seen earlier, although on *Medium.com* instead.

VII. 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 four classification algorithms, which are, Logistic Regression, SVC with Gaussian kernel, Random Forest and Neural Network. We tune our models with different value of hyper-parameter to find the the model with the highest accuracy.

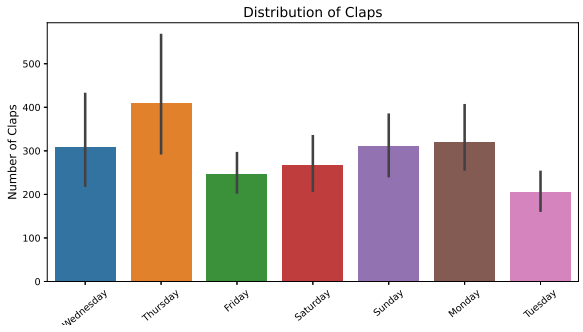

Fig. 10. No. of Claps vs Weekday

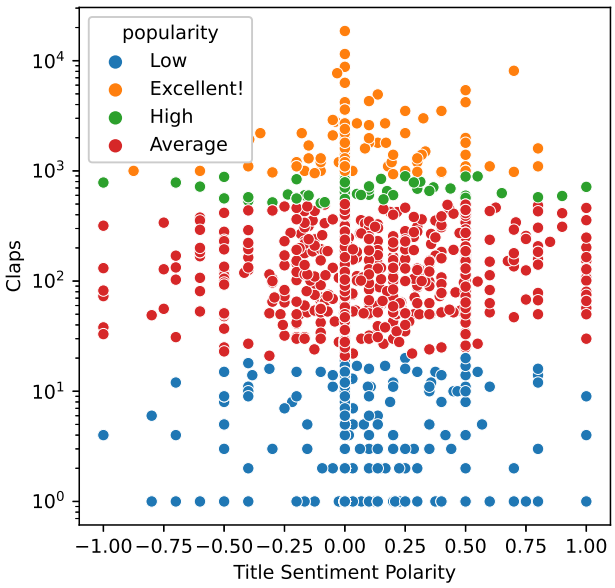

Fig. 11. No. of Claps vs Sentiment

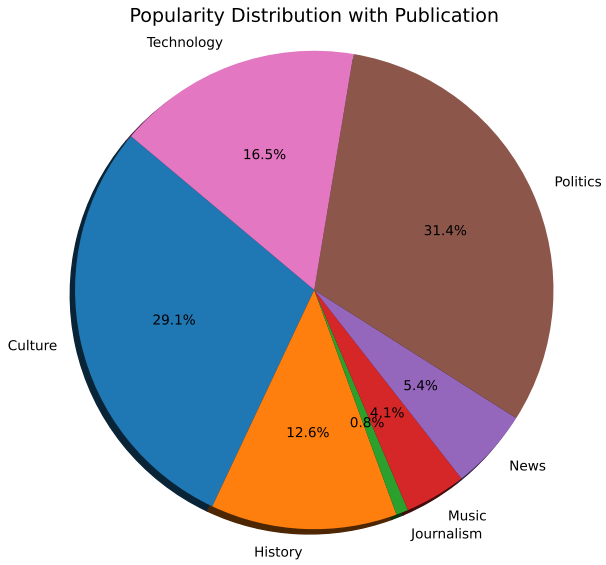

Fig. 12. Popularity (No. of Claps) vs Publication

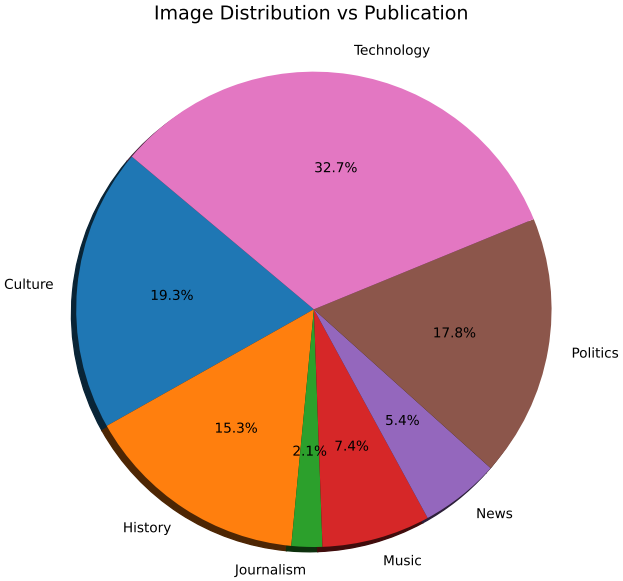

Fig. 13. Image Distribution with Publications

VIII. Results

Figure 19 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 20 shows that for the SVC model with gaussian kernel, the accuracy increases with the regularisation parameter and then become constant after 10. The accuracy achieved using the SVC(with the gaussian kernel) is 70 % . Figure 21 shows that for the Random Forest model, the accuracy keeps increasing with the increase in the number of trees. The accuracy achieved with

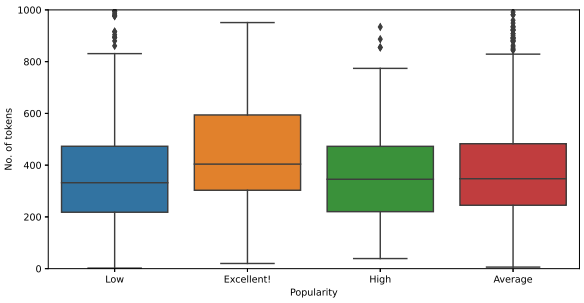

Fig. 14. No. of Tokens vs Popularity

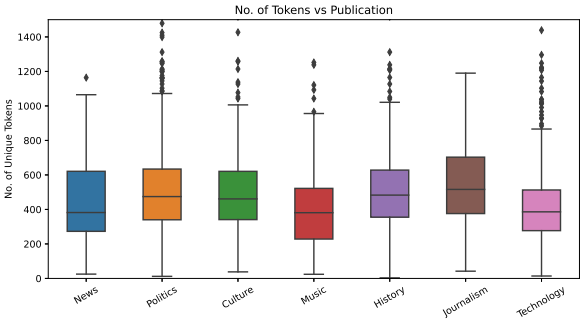

Fig. 15. Tokens vs Publication

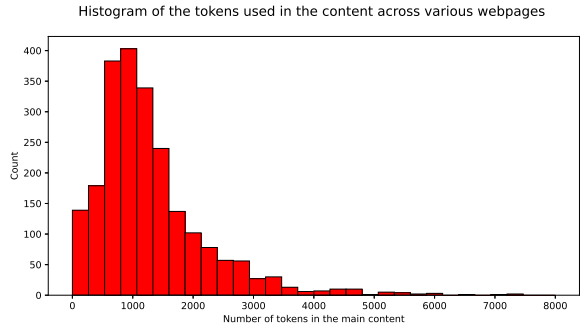

Fig. 16. Histgram of no. of tokens

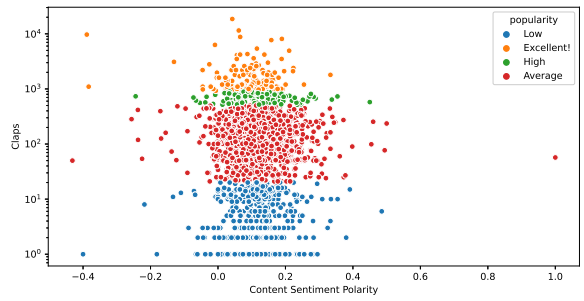

Fig. 17. Polarity vs Claps

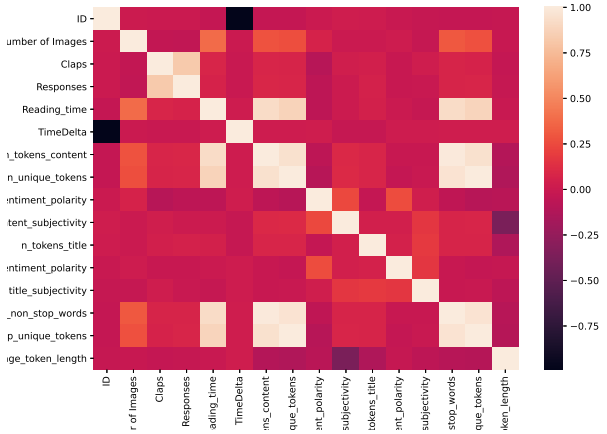

Fig. 18. Heatmap

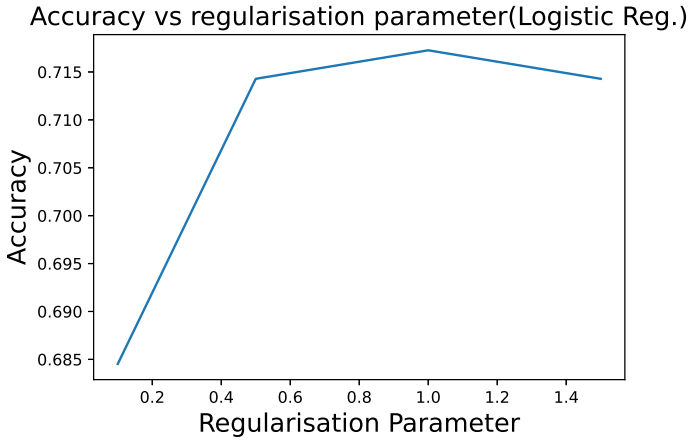

Fig. 19. Accuracy vs Regularisation parameter (Logistic Regression)

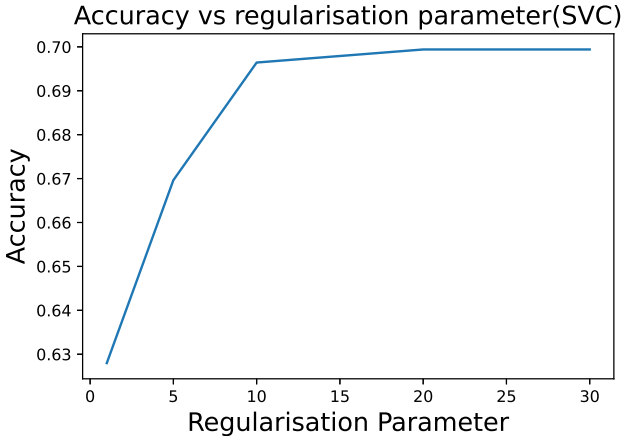

Fig. 20. Accuracy vs Regularisation parameter (SVC)

Random Forest, using 300 trees is 78 % . Figure 22 shows that for the Neural Network achieves a accuracy of about 63 %. Here, the neural network has one hidden layer and we can see

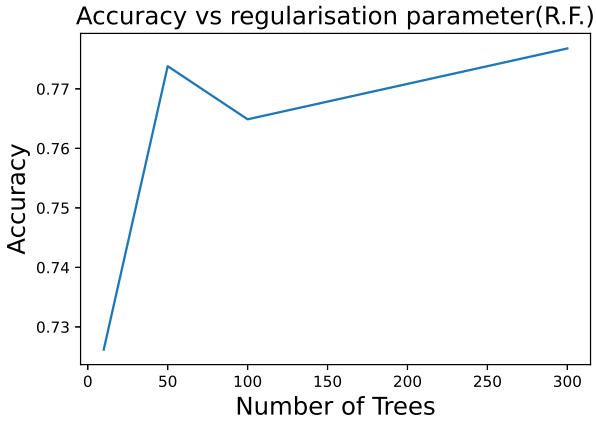

Fig. 21. Accuracy vs Regularisation parameter (Random Forest)

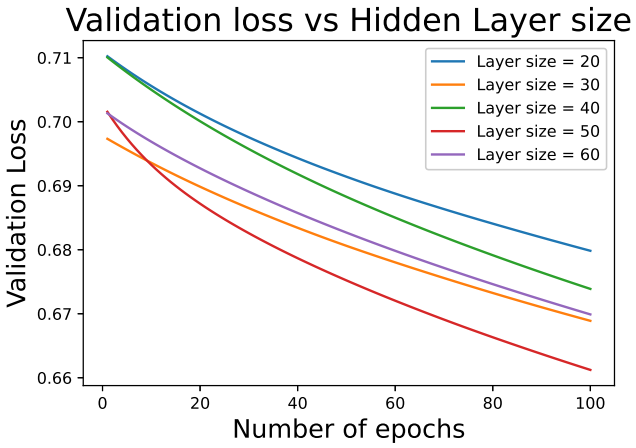

Fig. 22. Training and Validation loss vs epoch

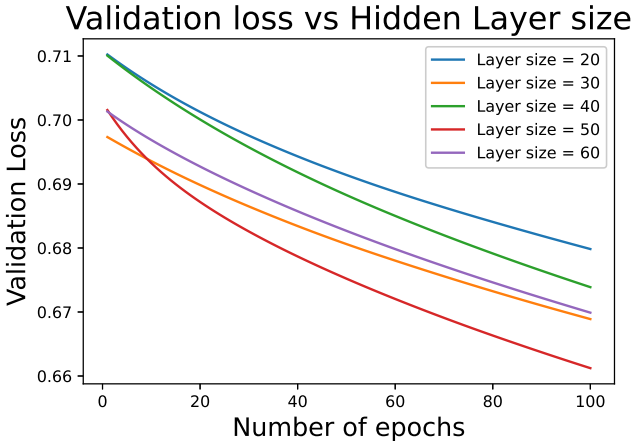

Fig. 23. Training and Validation vs epoch

in Figure 23 that fastest learning takes place with layer size equal to 50. Overall, Random Forest gives the best accuracy, therefore it is the best algorithm to model this classification problem.

IX. Discussion

In Figure 10, we see that most of the articles are published on Thursday, in anticipation of the Weekend. From Figure 11, we see that most of the popular articles portray a neutral sentiment. In Figure 12, 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 14, we see that Excellent articles have, on average, roughly, 50 more words than others. From Figure 15, we see that technological articles have less unique tokens. This can be attributed to high number of images as well as repetition of the same jargon. From Figure 16, we see that the tokens are distributed like a chi-squared distribution. Also, the mode is at 1000 tokens. In Figure 17, we see that in general, the content is slanted towards positive polarity. From the Heat Map in Figure 18, we see that most of the variables are independent of each other. 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 token,s 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.

X. Conclusion

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.

Acknowledgment

We would like to thank Prof. Manjesh Hanawal, Prof. Amit Sethi, Prof. Sunita Sarawagi and Prof. S. Sudarshan as we have learnt almost everything we have presented in this report from them. We would also like to thank the TA Arjit Jain, who provided feedback on our project ideas, and helped us streamline the project.

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," Oct. 2016, pp. 1–5. DOI: 10.1109/ICISET.2016.7856498.

[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," [Online]. 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," [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2754/3209. (accessed: 13.12.2020).