

PREDICTIVE ANALYTICS OF POPULARITY OF NEWS IN SOCIAL MEDIA WITH KEY PERFORMANCE INDICATORS USING BIGDATA ANALYTICS

N.Syed Rizwan^{1}, Revathi.R²*

Abstract

It is evident from recent impact from people around the world that everyone was interested in reading news using online applications and tools. The news published in social media platforms contain various information that might provide source to identify the popularity of the person. Key Performance Indicators (KPI) are special features that are capable of measuring the quality of the predictive data using quantifiable components. The major objective of this research paper is to perform predictive analytics of the various parameters used to assess popularity of news being shared in social media using KPI in Bigdata Analytics. The dataset for popularity of news was selected with 61 features and 39644 records to determine the Key Performance Indicators in performing the prediction and analysis of popularity of a person with significant success. The prediction was performed in WEKA using classification and cluster analysis whereas the analytics was performed using Tableau to find out the KPI of the dataset. The research identified six KPI at the end of the outcomes that would be regarded as key to perform any analysis.

Keywords: Bigdata Analytics, Key Performance Indicators, Predictive Analytics, News Popularity, Classification and Cluster Analysis.

I. INTRODUCTION

Popularity Analytics with prediction is one of the growing fields in social media especially among the celebrities as they wish to know the popularity of their career through social media platforms. However, various parameters are significant to analyse and calculate the

popularity of a person in social media. The social media users were virtual and couldn't be directly seen to test their opinions online. Hence few features could be tested to find the popularity of the person. To assess the popularity of the person, it is essential to identify the popularity of the news being spread to different social media applications in different genre for better understanding. The major objective of the paper is to conduct predictive analytics of the popularity of the news in social media and its best features used as key performance indicators through analytics. The prediction of the popularity of the news is assessed using WEKA tool and the analytics of Key Performance Indicators (KPI) [1] to be more specific here as Key Popularity Indicators performed using Tableau Tool for Bigdata analytics respectively. Thus, this research study encompassed the importance of analysing the best features as key performance indicators that enables efficient prediction of popularity of the person or news with predictive analytics.

The Popularity of the person [2] or news in social media varies from time since the popular use of social media in various countries. The popularity depends on various features like likes, shares, comments, reshares, positive and negative feedbacks etc. The popularity could be either positive or negative depending upon its nature of propounding the popularity and the efficiency of prediction. This is because, the expert system must not predict the positive popularity as negative popularity and vice versa. The popularity analysis has been a boon for the current generation and would continue for the next generation as well. Hence this popularity analysis also gains social significance in various sectors like Politics, Entertainment and other Business organizations for the potential users to handle the competitions in the current world.

Department of Computer Science

Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India

*Corresponding Author

offers crucial forensic evidence to security and police investigations [1]. Besides it delivers much-need

II. RELATED WORKS

Numerous literatures were studied and analysed from time to time on the various parameters related to popularity of social media news and also on the individual profile in the dataset. The data analysis and assessments performed in these literatures assisted in finding out the need for making this popularity analysis in the current scenario. Nashaat, M., & Miller, J. (2021) [3] designed a model that assisted in enhancing the prediction and estimation of popularity of the news using supervised learning process. In general, supervised learning assisted in bringing quality solution to the understand popularity of a person in social media. Oschatz, C., Stier, S., & Maier, J. (2021)[4]. Used the concept of popularity analysis in Twitter social media that is predominantly used by celebrities and business men around the world. The author focused on multiple tweets at the same time of covering news at the social media platforms like Twitter. (Ahmed, T., et.al. 2022)[5] developed a model for predicting the popularity of news with the assistance from Machine Learning techniques. To make these models, it is required to understand the importance of popularity prediction in social media. Hence, (Lamot, K., & Van Leuven, S. 2022)[6] studied the importance of popularity analysis on the basis of some of the parameters like communication, linguistics and ethical aspects of analysis. The results were found convincing and encouraging for future research analysis on the popularity. (Hosseini, D., Sood, K., & Bacha, V., 2022)[7] utilised machine learning methods to predict the popularity of news from secondary sources of dataset. Also Fringe News or Fake News detections were performed by (KhudaBukhsh, A. R., Sarkar, R., Kamlet, M. S., & Mitchell, T. M., 2022)[8] to detect the correct and fake news based on the US presidential elections in 2020. The model was successful and had significant impact on the outcome of the detections. (Righetti, N., 2021)[9] also detected Fake News based on the quantitative

analysis using the computational and scientific methods.

Various social media applications were tested for popularity analysis by different authors. (Blassnig, S., Udris, L., Staender, A., & Vogler, D., 2021)[10] analysed the Popularity of leaders in Facebook during the election campaign using simple data mining and analytics techniques. (Yaya, O. S., Ajose, T. S., & Ogbonna., 2021) [11] developed a model for the same to assess the popularity of Nigerian Music popular star using statistical analysis and data science. Some of the specific models like k-Nearest Neighbour models was used by (Mao, X., & Xu, F., 2022)[12] to detect the multimedia-based popularity analysis on the social network people. Similarly, (Clapham, B., Siering, M., & Gomber, P., 2021)[13] analysed whether the popular news was relevant news for the people to make their decisions in their relevant field. This model was developed based on the algorithms. (Bohra, N., & Bhatnagar, V., 2021)[14] tested this model with a group-based popularity prediction in social media along with new models called Adam optimisation.

Few analyses on Popularity Analysis were also carried out by different researchers to encompass the importance of popularity prediction and analysis in social media in recent times. (Bhatia, S., 2021)[15] assessed the evaluation and application of machine learning techniques in assessing the popularity of the news articles being published in a conference. Also, the importance of people requesting to like, comment and share their articles was studied by (Garvey, M. D., Samuel, J., & Pelaez, A., 2021)[16] using artificially intelligent and econometric analysis system. This method assisted in determining the best parameters for prediction of popularity. The importance of metrics and their interpretations were also studied by (Lamot, K., 2022)[17] to indicate that media inputs were recognised as mainstream outlets by social media like Facebook. The best news identification and recommendation model was proposed by (Wang, J., Chen, Y., Wang, Z., & Zhao, W., 2021)[18] that was developed using machine learning methods for multiple views and interests of the persons seeking to know their

popularity in social media. Also, a data driven analysis was performed by (Villosio, M., 2021)[19] that would assist the person to identify their popularity and increase more in future using practical methods. The biased methods to enhance the popularity of a person was also performed by (Zhang, Y., et.al. 2021)[20]to recommend best methods for analysing the popularity of the news in social media. However, in many situations, the popularity of the person in social media is influenced by various factors as it is being identified for the news. Hence analytics with prediction has to be performed to ascertain the prediction of popularity and the key performance or popularity indicators for enhancing the quality of predicting the popularity of the person.

III. MATERIALS AND METHODS

Popularity Prediction and Analysis is identified as one of the complex measures as it varies from time immemorial based on the needs of the people in social media in different situations. However, few parameters are highly in common in all circumstances to predict the popularity of an individual as well as the news circulated in the social media platform. To identify the potential parameters used to determine the best features of predicting the popularity, a dataset from secondary sources called “Popularity News” from Kaggle[21]was selected and used in the research work. The dataset comprised of 61 features among which 2 were non-predictive, 58 were predictive and 1 class feature which are highly significant in assessing the output.

These data in the dataset were rearranged and pre-processed to form a training set in Excel form as shown in Fig.1. The dataset was initially pre-processed using WEKA Attribute Selection methods and grouped together using the feature extraction and elimination methods to identify the best 21 features among the 62 features used for prediction of clusters. These features were trained and modeled to form the prediction model as shown in Fig.2

Fig.1. Preparation of Popularity News Dataset in Excel form from CSV format.

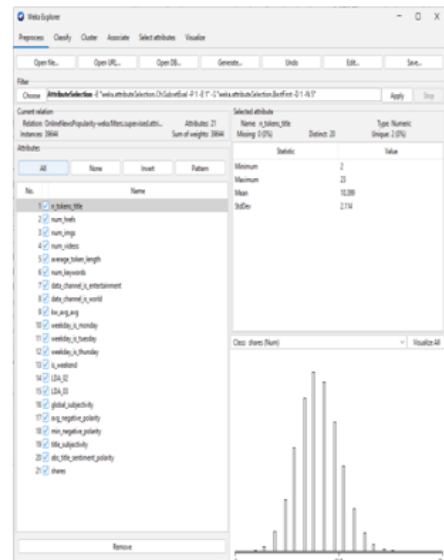


Fig.2. Pre-Processing of dataset in Excel form with selected 21 features

IV. PREDICTION OF CLUSTERS OF POPULARITY

The initial phase of the study originated through loading of the dataset in WEKA tool and completion of the pre-processing stage including Feature extraction and elimination process. After extraction of features, the best features were grouped for popularity analysis as numerated in Table.1.

Table.1. Enumerated Features selected from Popularity Dataset after pre-processing

S.No	Selected Features after Feature Extraction and Elimination
1	n_tokens_title
2	num_hrefs
3	num_imgs
4	num_videos
5	average token length
6	num_keywords
7	data_channel_is_entertainment
8	data_channel_is_world
9	kw_avg_avg
10	weekday_is_monday
11	weekday_is_tuesday
12	weekday_is_thursday
13	is_weekend
14	LDA_02
15	LDA_03
16	global_subjectivity
17	avg_negative_polarity
18	min_negative_polarity
19	title_subjectivity
20	abs_title_sentiment_polarity
21	shares

As mentioned in Table.1., among the total instances of 39,644 records in the dataset forming a bigdata category of dataset and 62 features, 21 features were selected as best features for cluster analysis and visualization. These 21 features contain all predictive except the feature “Shares” which is the class feature of the dataset. The experiment was initiated with the test in cluster mode of WEKA with 66% split of the trained and tested data, the cluster showed good results with classification of clusters into Cluster 0 indicating popular and Cluster 1 with non-popular respectively. The overall classification of clusters with their centroids were presented in Table.2.

Table.2. List of Features and the formation of clusters to predict popularity of news

S. No	Attribute	Full Data (39644.0)	Cluster 0 (32254.0)	Cluster 1 (7390.0)
1	n_tokens_title	10.3987	10.3893	10.4399
2	num_hrefs	10.8837	10.9417	10.6306
3	num_imgs	4.5441	4.5591	4.4788
4	num_videos	1.2499	1.2366	1.3077
5	average token length	4.5482	4.5495	4.5425
6	num_keywords	7.2238	7.2325	7.1857
7	data_channel_is_entertainment	0.178	0.179	0.1739
8	data_channel_is_world	0.2126	0.2133	0.2092
9	kw_avg_avg	3135.8586	3138.3286	3125.0785
10	weekday_is_monday	0.168	0.2065	0
11	weekday_is_tuesday	0.1864	0	1
12	weekday_is_thursday	0.1833	0.2253	0
13	is_weekend	0.1309	0.1609	0
14	LDA_02	0.2163	0.2162	0.2167
15	LDA_03	0.2238	0.2247	0.2198
16	global_subjectivity	0.4434	0.4438	0.4416
17	avg_negative_polarity	-0.2595	-0.2602	-0.2565
18	min_negative_polarity	-0.5219	-0.5238	-0.5139
19	title_subjectivity	0.2824	0.283	0.2797
20	abs_title_sentiment_polarity	0.1561	0.1564	0.1547
21	shares	3395.3802	3439.5725	3202.5007

As mentioned in the Table.2., the complete data was 39644 among which 32254 accounting to 81% of the news were correctly classified as popular news and the remaining 7390 news with 19% were classified as non-popular or irrelevant news. Thus, the features have a significant role in determining the prediction of popularity of news in social media. This was proved at the end of the experiment with k-means analysis and the visualizations were formed as shown in Fig.3.

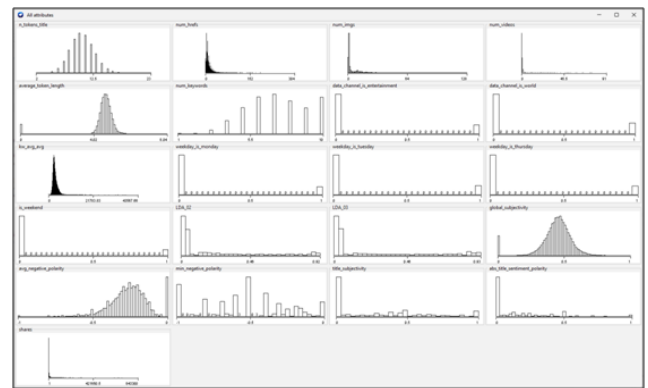


Fig.3. Visualizations of the results obtained after prediction of clusters for popularity analysis

As portrayed in Fig.3., the attribute or feature “shares” influenced the other features to predict the outcomes with the existing model. Some of the features like “global_subjectivity”, “avg_negative_polarity”, “weekdays”, “subjectivity” also influenced the results after the completion of the experiment in WEKA. However, after prediction, the same results have to be verified based on the analytics on the features of the dataset.

V. ANALYTICS OF KEY POINTS IN POPULARITY

Key Performance Indicators (KPI) represent the best features that are capable of predicting the expected results of the experiment. The KPI in this case are identified as Key Popularity Indicators as they determine the best features or key features used in prediction of popularity of the person or news with enhanced performance. To identify the potentially best features of the popularity news dataset, the dataset was loaded into the Tableau Desktop version for analytics as shown in Fig.4.

PREDICTIVE ANALYTICS OF POPULARITY OF NEWS IN SOCIAL MEDIA WITH KEY PERFORMANCE INDICATORS USING BIGDATA ANALYTICS

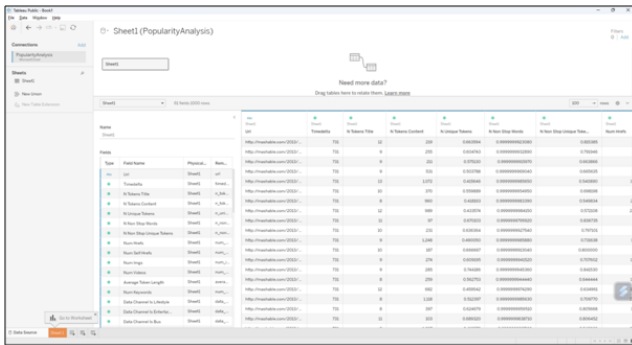


Fig.4. Loading of Popularity News Dataset in Tableau for Analytics after prediction

After loading the dataset, the initial features like Count of Shares, Average Shares, Maximum Shares and Minimum Shares were examined and presented as shown in Fig.5.



Fig.5. Initial Analytics of KPI for Popularity on the basis of shares of the news.

As shown in Fig.5., the shares at the maximum level were 29 lakhs and to the minimum level it was 11 lakhs respectively. The average of the shares for the news was around 17 lakhs. Thus, it was initially tested that the best range of popularity prediction of a news is between 29 Lakhs and 11 Lakhs or an average of 17 Lakhs. The maximum number of counts for the popularity of the shares was 699 at a single news share. In the second level, the next parameter polarity of the news was tested. Both negative and positive polarity were tested as shown in Fig.6(a) and Fig.6(b) respectively.

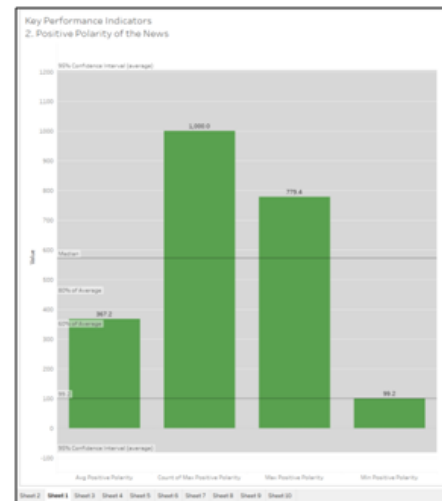


Fig.6(a). Positive Polarity Analytics of Popularity News dataset.



Fig.6(b). Negative Polarity Analytics of Popularity News dataset.

As shown in Fig. 6(a) and Fig.6(b), the positive polarity in many cases showed an uplifted graph with maximum of 1000 and minimum of 99 values. However, the Negative Polarity showed a maximum of 1000 similar to Positive Polarity whereas the negative polarity showed a downfall of -497. This showed the reach of the news among the public in social media. Hence these two features Positive and Negative Polarity could be used as a predictor for popularity of the news.

In the third level, the negative aspects have to be assessed for determining the popularity of the news. One among them

was the impact of banned news in the dataset as shown in Fig.7.

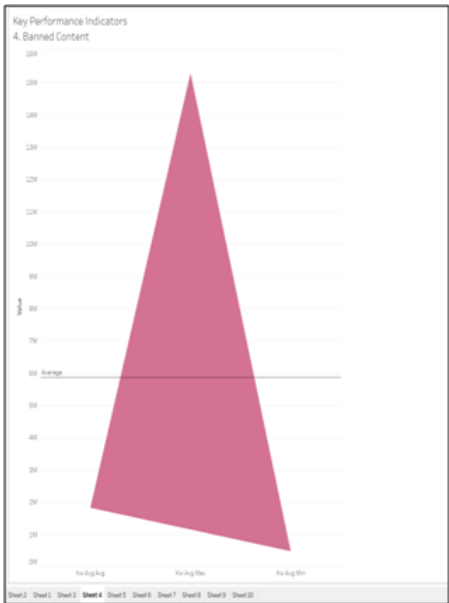


Fig.7. Banned Contents analysis in Popularity Dataset.

The banned contents represented the unwanted or offensive news in the dataset spread among the members in social media. It was identified from the keywords used in the news being published. Thus, the fields “kw_min_min”, “kw_max_min”, “kw_avg_min”, “kw_min_max”, “kw_max_max”, “kw_avg_max”, “kw_min_avg”, “kw_max_avg”, “kw_avg_avg” were observed and formed as performance measures to identify the outcomes. It was found from the analytics that banned or unwanted information were shared with a reach of around 15M to 16M views over a time. This can also be used as a factor to analyze the popularity of the content in negative form. After completing with keyword analysis, the title analysis was also performed. The title and its impact on the audience for reading the content was assessed. It was identified that Title subjectivity has achieved more response than the Title sentiment polarity of the news. It was evident from Fig.8. that Title Subjectivity has a maximum of over 270M reaches in comparison to the Title Sentiment Polarity which created an impact only on over 73M reach at the same time.

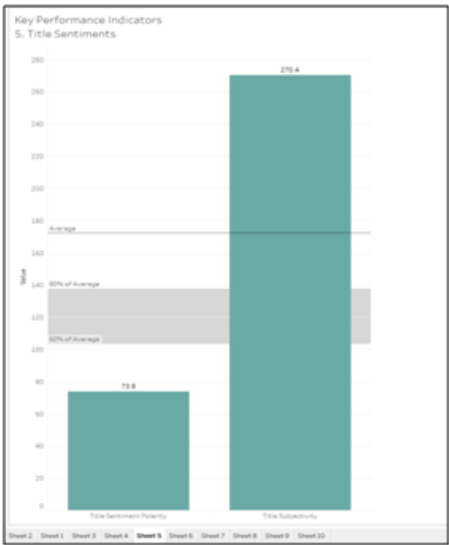


Fig.8. Reach of Title of the news in terms of Subjectivity and Polarity Analysis

In the next level, the Globalization of News was also analyzed as a factor for popularity predictions. It was found that subjectivity of the Global has high impact of 456M views in comparison to other features like sentiment polarity, global rate positive words and global rate negative words as shown in Fig.9.

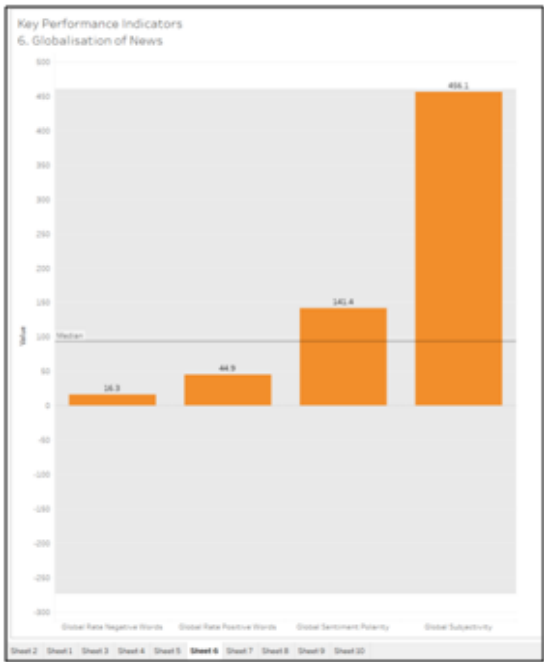


Fig.9. Analytics of Globalization of the News in social media

Some of the interesting features were also considered as KPI for predicting the popularity of the news in social media like the weekdays when the news was released to the viewers in social media as shown in Fig.10.

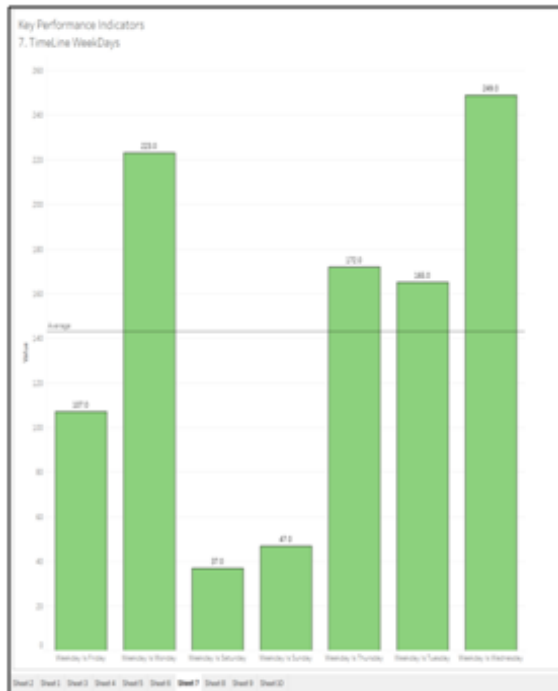


Fig.10. Weekday Analysis of News Popularity on all seven days of the week.

It was evident from Fig.10. that Wednesday has a maximum reach of 249M followed by Monday with 223M, Thursday with 172M and Tuesday with 165M respectively. The content shared on Friday with 107M was moderate whereas Saturday and Sunday showed lowest reach with 37M and 47M respectively. Thus, it was evident that even week day posts indicating the time of post was also a chief factor in determining the popularity of the news in social media. The Analytics further continued with the non-stop words used in the news content and the type of data channel used to transmit the news to the public. The non-stop words represented the continuous words with existing models represented as “non-stop words” and with new words represented as “non-stop unique words” were tested and the results were analysed as shown in Fig.11.

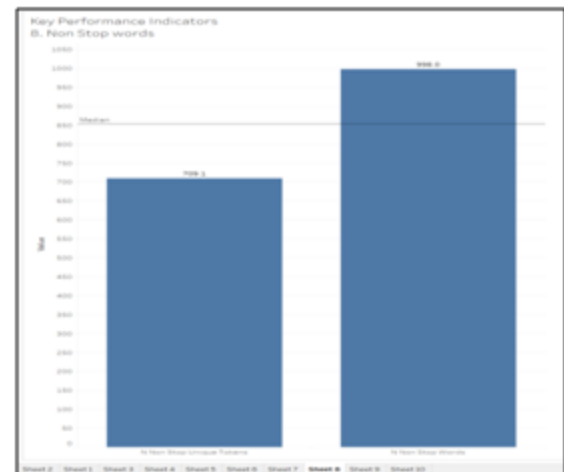


Fig.11. Analytics based on Non-Stop regular words and unique words in popularity analysis

As shown in Fig.11., the Regular non-stop words had high influence (998) over the unique non-stop words. Similarly, the data channel for entertainment, bus, lifestyle, sochmed, tech and world were analyzed as shown in Fig.12.

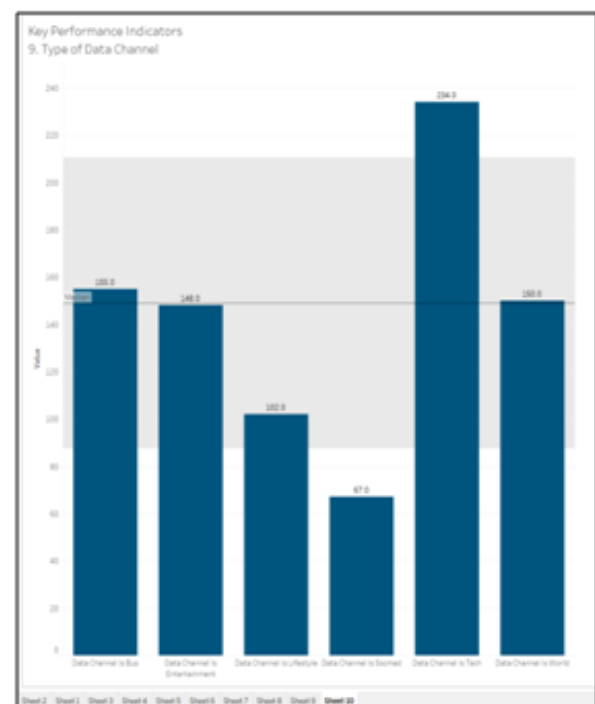


Fig.12. Analytics of various Data Channels in Popularity Analysis

As mentioned in Fig.12., the Technology channel has high popularity with 234M followed by Bus (155M), Entertainment(148M) and World channels(150) respectively. Thus, with the Analytics performed in Tableau after WEKA predictions, it was identified that by selecting the best features or parameters for prediction, the outcomes for Popularity prediction could be achieved with enhanced performance. In this experiment, the parameters like Shares, Polarity, Type of Channel, Weekdays, Globalization of News, Banned News, non-stop words, title reach were considered as Key Popularity Performance parameters to predict the popularity of the news in social media with maximum performance.

VI. CONCLUSION

The research paper analysed and identified that by predicting the Key Performance Indicators of the Popularity dataset, the prediction of popularity of a person is highly significant. Also, the prediction can be performed for bigdata using analytic tool like Tableau where dashboards are used to analyse the features of the dataset to test for best ranked features. Thus, in this experiment, it is found that Shares, Polarity, Title, Days, Channel type and other parameters could predict the popularity of a person similar to the news as studied in this paper. This model could be employed in the study of a person's popularity in the social media in the near future.

REFERENCES

- [1] Chen, Q., & Wang, Y. (2021). Key-performance-indicator-related state monitoring based on kernel canonical correlation analysis. *Control Engineering Practice*, 107, 104692.
- [2] Xie, C., Jauhari, S., & Mora, A. (2021). Popularity and performance of bioinformatics software: the case of gene set analysis. *BMC bioinformatics*, 22(1), 1-16.
- [3] Nashaat, M., & Miller, J. (2021). Improving news popularity estimation via weak supervision and meta-active learning.
- [4] Oschatz, C., Stier, S., & Maier, J. (2021). Twitter in the news: An analysis of embedded tweets in political news coverage. *Digital Journalism*, 1-20.
- [5] Ahmed, T., Paul, R. R., Alam, A., Hasan, M., & Rab, R. (2022). Bangla News Popularity Prediction Using Machine Learning Techniques. Raqeebir, Bangla News Popularity Prediction Using Machine Learning Techniques (April 28, 2022).
- [6] Lamot, K., & Van Leuven, S. (2022). Paywalled popularity? Exploring differences in audience engagement in front and behind the paywall. In *A True and Fair View: Communicative, Linguistic and Ethical Aspects of Expressing Point of View in Journalism*.
- [7] Hosseini, D., Sood, K., & Bacha, V. (2022, October). News Popularity Prediction with Machine Learning. In *2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)* (pp. 474-480). IEEE.
- [8] KhudaBukhsh, A. R., Sarkar, R., Kamlet, M. S., & Mitchell, T. M. (2022, June). Fringe news networks: dynamics of US news viewership following the 2020 presidential election. In *14th ACM Web Science Conference 2022* (pp. 269-278).
- [9] Righetti, N. (2021). Four years of fake news: A quantitative analysis of the scientific literature. *First Monday*.
- [10] Blassnig, S., Udris, L., Staender, A., & Vogler, D. (2021). Popularity on Facebook during election campaigns: An analysis of issues and emotions in parties' online

- communication. *International Journal of Communication*, 15, 21.
- [11] Yaya, O. S., Ajose, T. S., & Ogbonna, A. E. Modelling the Popularity of Some Selected Nigerian Music Top Stars with Long-Range Dependence.
- [12] Mao, X., & Xu, F. (2022). Research on the Popularity Prediction of Multimedia Network Information Based on Fast K Neighbor Algorithm. *Journal of Electrical and Computer Engineering*, 2022.
- [13] Clapham, B., Siering, M., & Gomber, P. (2021). Popular news are relevant news! How investor attention affects algorithmic decision-making and decision support in financial markets. *Information Systems Frontiers*, 23(2), 477-494.
- [14] Bohra, N., & Bhatnagar, V. (2021). Group level social media popularity prediction by MRGB and Adam optimization. *Journal of Combinatorial Optimization*, 41(2), 328-347.
- [15] Bhatia, S. (2021, September). Application and evaluation of Machine Learning for news article popularity prediction. In *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSSES)* (pp. 1-5). IEEE.
- [16] Garvey, M. D., Samuel, J., & Pelaez, A. (2021). Would you please like my tweet?! An artificially intelligent, generative probabilistic, and econometric based system design for popularity-driven tweet content generation. *Decision Support Systems*, 144, 113497.
- [17] Lamot, K. (2022). What the metrics say. The softening of news on the Facebook pages of mainstream media outlets. *Digital Journalism*, 10(4), 517-536.
- [18] Wang, J., Chen, Y., Wang, Z., & Zhao, W. (2021, October). Popularity-Enhanced News Recommendation with Multi-View Interest Representation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 1949-1958).
- [19] Villosio, M. (2021). Data-driven Analysis of Interactions and Popularity Increase in Online Social Networks (Doctoral dissertation, POLITECNICO DI TORINO).
- [20] Zhang, Y., Feng, F., He, X., Wei, T., Song, C., Ling, G., & Zhang, Y. (2021, July). Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 11-20).
- [21] Quaranta, L., Calefato, F., & Lanubile, F. (2021, May). KGTorrent: A dataset of python Jupyter notebooks from kaggle. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)* (pp. 550-554). IEEE.