PREDICTING THE NEWS POPULARITY ON

MULTIPLE SOCIAL MEDIA PLATFORMS

BY :- ASHWANI BIST

ABSTRACT:

Consuming news articles is an integral part of our daily lives and news agencies such as The Washington Post (WP) expend tremendous eﬀort in providing high quality reading experiences for their readers. Journalists and editors are faced with the task of determining which articles will become popular so that they can eﬃciently allocate resources to support a better reading experience.

The reasons behind the popularity of news articles are typically varied, and might involve contemporariness, writing quality, and other latent factors.

INTRODUCTION:

With the advancement in technology, news organizations have begun to rely more on online social platforms and media analytics as a way to attract readers. So, for news publishing sources, it’s become very important to know which kind of news articles will appeal more to the readers. In this project, firstly we have a news dataset which contains around 100000 news items published on three social media platforms: Facebook, Google Plus and LinkedIn, between November 2015 to July 2016 on four topics: Obama, Economy, Palestine, Microsoft. And we also have 12 social feedback dataset which contains the popularity level of news items in incremental time slices of 20 min after publication.

PROBLEM STATEMENT :

**This is a large data set of news items and their respective social feedback on multiple platforms: Facebook, Google+ and LinkedIn. The collected data relates to a period of 8 months, between November 2015 and July 2016, accounting for about 100,000 news items on four different topics: Economy, Microsoft, Obama and Palestine.**

**DATA DESCRIPTION**

### Attribute Information:

### IDLink (numeric): Unique identifier of news items

### Title (string): Title of the news item according to the official media sources

### Headline (string): Headline of the news item according to the official media sources

### Source (string): Original news outlet that published the news item

### Topic (string): Query topic used to obtain the items in the official media sources

### PublishDate (timestamp): Date and time of the news items' publication

### SentimentTitle (numeric): Sentiment score of the text in the news items' title

### SentimentHeadline (numeric): Sentiment score of the text in the news items' headline

### Facebook (numeric): Final value of the news items' popularity according to the social media source Facebook

### GooglePlus (numeric): Final value of the news items' popularity according to the social media source Google+

### LinkedIn (numeric): Final value of the news items' popularity according to the social media source LinkedIn

### VARIABLES OF SOCIAL FEEDBACK DATA

### IDLink (numeric): Unique identifier of news items

### TS1 (numeric): Level of popularity in time slice 1 (0-20 minutes upon publication)

### TS2 (numeric): Level of popularity in time slice 2 (20-40 minutes upon publication)

### TS... (numeric): Level of popularity in time slice ...

### TS144 (numeric): Final level of popularity after 2 days upon publication

### EXPLORATARY DATA ANALYSIS:

As the first step of our experiment, we performed Data Cleaning by removing trash and duplicate data, applied null value treatment and removed outliers in the data set using the **90th percentile quantile method**. We further applied the **Standardization** technique for feature scaling.

In **Exploratory Data Analysis**, we categorized SentimentTitle, SentimentHeadline and sources to extract some meaningful insights from the data. Then we compared popularity between the social media platforms using multiple plots.

We applied text-preprocessing techniques to transform the headline and title of the news items.

For feature selection, we used **ExtraTreeRegressor** and **Correlation matrix** to obtain results on features.

For model prediction, we used supervised machine learning algorithms like Decision Trees, Catboost, LightGBM, Gradient Boosting, KNN and then applied hyperparameter tuning techniques to obtain better accuracy and to avoid overfitting.

CONCLUSION:

Through **Exploratory Data Analysis**, we observed that on weekends the number of news items published gradually decreases.

News Items related to the topic of Obama gain higher popularity on Facebook. And, Economy news gains equal popularity on GooglePlus and Facebook whereas Microsoft-related news items gain higher popularity on LinkedIn.

We observed that the number of news items related to Palestine is comparatively less. We also observed that news Items published at midnight gain less popularity.

The best model score obtained from our prediction modeling is **92%** and the most impacting features are popularity level in time slices variables.

### REFERENCES:

[https://towardsdatascience.com](https://towardsdatascience.com/)

[https://www.analyticsvidhya.com](https://www.analyticsvidhya.com/)

[https://machinelearningmastery.com](https://machinelearningmastery.com/)