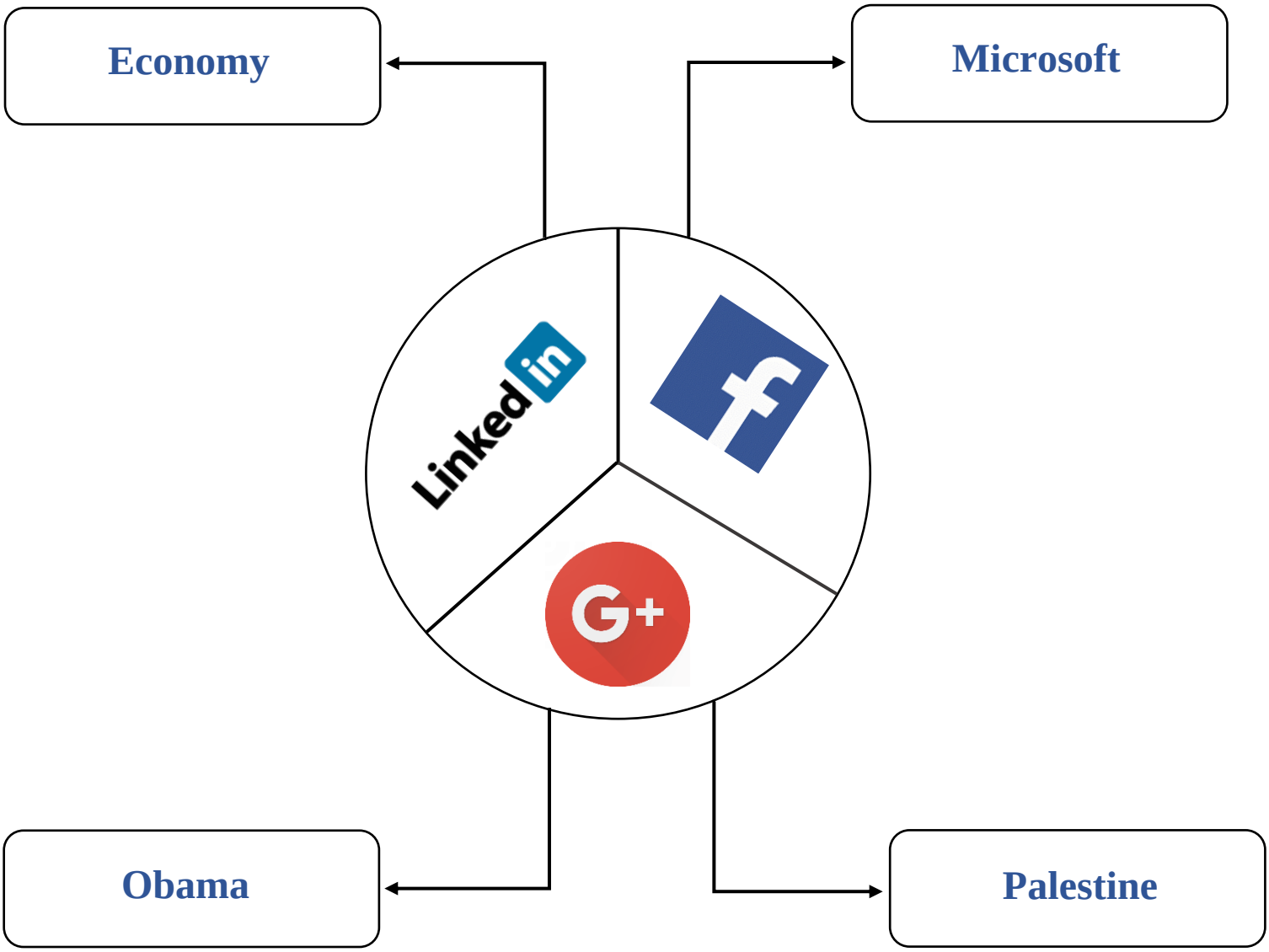
TEAM NEURON

Predicting the News Popularity in Multiple Social Media Platforms



Team Member

- Shivangi Patel
- Ravi Gohil
- Harsh Thakkar
- Kirti Pandey

Mentor Name

- Rohit Bhadauriya

Publish Date

- 20-10-2022

Abstraction

News organizations have increasingly come to rely on media analytics as a way to attract and retain readers. It's become vital for media companies to know which news articles resonate with readers, and which articles don't. This project is aimed at finding out what makes a news article popular or unpopular on social media. In this project, the focus is mainly on EDA, where we have picked out trends and patterns while figuring out what might be the most important predictors for a machine learning model.

The dataset used for this project 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.

Acknowledgement

The successful completion of this report would not have been possible without the co-operation and support of all the people who supported us, the professors, Group members and our institute.

I extend my thanks to respected Prof. Rohit Bhadauriya and all our teachers for imparting guidance and co-operation in the making of this research. We are also thankful to the Gujarat University and Department of Animation BIA for introducing Dissertation as a significant subject and aspect of Master in Business intelligent and analytics study without which we would have no practical exposure of real time environment.

TEAM NEURON

Index

1.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.

2.Data Introduction

Technically this data set shows the insights were in November 2015 to July 2016 were these four news (Microsoft, Obama, Palestine, economy) were effectively shared in these three platforms (linked in, google plus, Facebook.)

This dataset describes about social media in which people get influenced by news and share it via various social media platforms.

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¶

3. Brief About work

There were two target variables

Target variable: sentimental title and sentimental heading

•We took sentimental title as target variable. Via title we get rate in which it is rated according to the influence of news and numbers of time it shared.

We as human get influenced by title or the heading. But here we took title because we think that title is most catchy word to easily get influence.

Then we have done cleaning.

- In which we did
- Change the data type
- Column name rename
- Null value drop (In the entire data set there were point percentage of data which holds null value so we dropped the null value which did not made much changes in dataset)

Value was in continues form so we used algorithm of linear regression. but we didn't get proper output regarding accuracy purpose.

So, we used ensemble technique: in that we used decision tree.

Target variable: round using numpy (round is numpy function)

With that technique target value will divided in to three parts:

- -1 not influence by people.
- 0 moderate influence by people.
- 1 highly influence by people.

4.Methodology

The existing methodologies for predictions are Leaner regression and Decision Tree.

4.1 Data pre-processing and data transformation

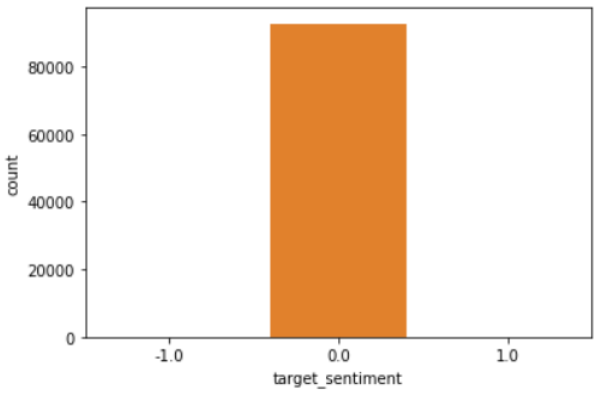
- Instead of date column we added publish day, publish month and season.

  Reason: For better performance to machine learning model.

- When we applied round: data was misbalance

  So, for that we performed over sampling and under sampling.

```
 0.0    92640
-1.0      172
 1.0      133
```



4.2 Developing and optimizing Linear regression

In liner regression we need a target variable in continues form so based on target variable we created a linear regression model where we have chosen sentiment title as a target variable. But we got 0.035 accuracy on our model the reason behind is, in dataset target variable is not co-relate with any column.

4.3 Developing and optimizing Decision Tree

After performing linear regression, we got very less accuracy so, we move forward with ensemble technique. Where, we round the target variable and get three categories (-1,0,1).

Then we performed decision tree on data and created a model. Which ended up with 0.86 percentage accuracy in our model.

5 Conclusion

1. Through our analysis we could figure that amongst all the four news topics, news items related to Economy were most popular on social media.

2. Secondly, we figured that Facebook has higher reach as compared to GooglePlus and LinkedIn.

3. Next we found out that the news items related to the topic Obama trended more on Facebook and Google plus. Also, the news items related to Microsoft trended more on LinkedIn.

4. In multivariate analysis we could conclude that 51% of Facebook news are also shared on Google plus and vice-versa.