Problem

In this project, I am aiming to find the target platform to post specific news.

Finding the best platform to post can be more efficient, for the users in different platforms may have different preference on the topic of news. Also, the number of comments (feedback) can represent the popularity of news. People seek a high volume of feedback to spread the news. Thus, I will predict the number of feedbacks for news on different platforms and conclude the best platform for posting different type of news.

To be more specific, this research wants to predict the best platform (among Facebook, GooglePlus, LinkedIn) for posting a certain type of news (among 'Obama', 'Microsoft', 'Economy', 'Palestine'). That is to say, if someone wants to spread any news about Obama, he/she can find where to post for the highest popularity. After the model has been developed, we can add more news type as well as online platform to generalize the result. The model can be used for political activities or business publicity.

Data

News Popularity in Multiple Social Media Platforms Data Set

Download: Data Folder, Data Set Description

Abstract: Large data set of news items and their respective social feedback on multiple platforms: Facebook, Google+ and LinkedIn.

Data Set Characteristics:	Multivariate, Time-Series, Text	**Number of Instances:**	93239	**Area:**	Computer
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	11	**Date Donated**	2018-02-20
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	44066

The data I will be using is real numbers. It is reasonable to use regression model to make prediction. There are 44066 observations with 11 features (3 out of 11 may not be needed).

Firstly, I am not going to use "Title" "Headline" and "Source" because knowing their topic will be enough.

Then, for the rest data, I will make the following changing:

1. IDLink: used as index, no need to process.
2. Topic, a categorical variable. Use OneHotEncoder:
   ['x0_economy', 'x0_microsoft', 'x0_obama', 'x0_palestine'] = [1,0,0,0]
   if df['Topic']== 'Economy'
   ['x0_economy', 'x0_microsoft', 'x0_obama', 'x0_palestine'] = [0,1,0,0]
   if df['Topic']== 'Microsoft'
   ['x0_economy', 'x0_microsoft', 'x0_obama', 'x0_palestine'] = [0,0,1, 0]
   if df['Topic']== 'Obama'
   ['x0_economy', 'x0_microsoft', 'x0_obama', 'x0_palestine'] = [0,0,0,1]
   if df['Topic']== 'Palestine'
3. PublishDate: the post time and date for the news, no need to process.
4. SentimentTitle: sentiment score for the title, range from (-1,1), no need to process.
5. SentimentHeadline: sentiment score for the headline, range from (-1,1), no need to process.
6. Facebook: the final popularity for Facebook

7. GooglePlus: the final popularity for GooglePlus
8. LinkedIn: the final popularity for LinkedIn

Lastly, the target variable will be a new variable “Platform”. When Facebook, GooglePlus and LinkedIn have the same popularity, Platform = 0. Else, Platform = the platform that has the largest popularity.

We can find several researches online that use the same data. The data can be used to research on topic detection, sentiment analysis and first story detection.

Data address:

https://archive.ics.uci.edu/ml/datasets/News+Popularity+in+Multiple+Social+Media+Platforms

GitHub:

https://github.com/zshen15/data1030_project.git