BUZZ IN SOCIAL MEDIA

Ashish kumar

01419011921

9310335443

Data Set Link

https://archive.ics.uci.edu/dataset/248/buzz+in+social+media

ABSTRACT

This report has two parts - one which has been used for Classification and the other which has been used for regression. The goal of this project is to mainly find out the optimum machine learning model or a ensemble of models that can accurately classify a social media data-point to be a buzz or not. The dataset has 77 features and over 1/2 million data points.

Social media interaction happens in a broad variety of context and magnitude. The vast majority of posts cause little to no discussion, while some start trends and become viral. We study the virality, explicitly of "Buzzes" - posts that evoke intense interaction over a short period of time, as they have been observed frequently, some- times with severe consequences for individuals and companies in the physical world. Early detection of a Buzz may help mitigate or prevent negative consequences of large scale social media outrage against companies or persons, by giving them a chance to react at an early stage. Collecting a labeled set of over 100,000 posts on Facebook pages, we first explore properties that define a Buzz using logistic regres- sion. This method helps us to interpret the results and derive prac- tical recommendations. We subsequently train classifiers and apply machine learning based classification techniques to demonstrate the potential capabilities of automated prediction. We achieve high recall with moderate precision, where feature boosting on broad feature sets yields the most promising results. Our study reveals that Buzzes are well described by a high num- ber of comments from previously passive users, a high number of likes given to comments, and a prolonged discussion period - properties that can be used to distinguish inconsequential posts from potentially volatile ones.

INTRODUCTION

This report aims to provide an analysis of the buzz in social media surrounding Buzz in social media. Social media platforms have become powerful tools for communication, information sharing, and content dissemination, making it essential to understand the level of attention and engagement generated by specific topics.

PROPOSED MYTHOLOGY

DATASETS

We use the buzz in social media dataset from the UCI Machine Learning Repository, which contains 14000 samples of and 97 features. We download the dataset from https://archive.ics.uci.edu/dataset/248/buzz+in+social+media and load it into a pandas data frame.

	NCD_0	NCD_1	NCD_2	NCD_3	NCD_4	NCD_5	NCD_6	NCD_7	BL_0	BL_1	...	AS(NA)_6	AS(NA)_7	AS(NAC)_0	AS(NAC)_1	AS(NAC)_2	AS(NAC)_3	AS(NAC)_4	AS(NAC)_5
0	0	0	0	0	0	0	0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0	0	0	0	0	0	0	0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0	0	0	0	0	0	0	0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0	0	0	0	0	0	0	0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0	0	0	0	0	0	0	0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

```
# DATA LOADING
# LIST OF COLUMN NAMES WHICH ARE GIVEN IN DATASET_TOWSHARDWARE_NAMES FILE
column_names=['NCD_0', 'NCD_1', 'NCD_2', 'NCD_3', 'NCD_4', 'NCD_5', 'NCD_6', 'NCD_7',
                  'BL_0', 'BL_1', 'BL_2', 'BL_3', 'BL_4', 'BL_5', 'BL_6', 'BL_7',
                  'NAD_0', 'NAD_1', 'NAD_2', 'NAD_3', 'NAD_4', 'NAD_5', 'NAD_6', 'NAD_7',
                  'AI_0', 'AI_1', 'AI_2', 'AI_3', 'AI_4', 'AI_5', 'AI_6', 'AI_7', 'AI_8' ,
                  'NAC_0', 'NAC_1', 'NAC_2', 'NAC_3', 'NAC_4', 'NAC_5', 'NAC_6', 'NAC_7',
                  'ND_0', 'ND_1', 'ND_2', 'ND_3', 'ND_4', 'ND_5', 'ND_6', 'ND_7',
                  'CS_0', 'CS_1', 'CS_2', 'CS_3', 'CS_4', 'CS_5', 'CS_6', 'CS_7',
                  'AT_0', 'AT_1', 'AT_2', 'AT_3', 'AT_4', 'AT_5', 'AT_6', 'AT_7',
                  'NA_0',' NA_1', 'NA_2', 'NA_3', 'NA_4', 'NA_5', 'NA_6', 'NA_7'
                  'ADL_0', 'ADL_1', 'ADL_2', 'ADL_3', 'ADL_4', 'ADL_5',' ADL_6', 'ADL_7',
                  'AS(NA)_0', 'AS(NA)_1', 'AS(NA)_2', 'AS(NA)_3', 'AS(NA)_4','AS(NA)_5', 'AS(NA)_6', 'AS(NA)_7',
                  'AS(NAC)_0', 'AS(NAC)_1', 'AS(NAC)_2', 'AS(NAC)_3', 'AS(NAC)_4','AS(NAC)_5', 'AS(NAC)_6', 'AS(NAC)_7',]
df=pd.read_csv('/content/drive/MyDrive/PROJECT_FILE/TomsHardware.data',names=column_names)
```

DATA ANALYSIS & VISUALIZATION

We perform some steps on the dataset, such as: Handling missing values: We check for any missing values in the dataset and find none.

```
# missing value
df.isnull().sum()
```

```
NCD_0        0
NCD_1        0
NCD_2        0
NCD_3        0
NCD_4        0
            ..
AS(NAC)_2    0
AS(NAC)_3    0
AS(NAC)_4    0
AS(NAC)_5    0
AS(NAC)_6    0
Length: 95, dtype: int64
```

Statistical measures: We describe the dataset for better data analyzing.

	count	mean	std	min	25%	50%	75%	max
NCD_0	28179.0	1.137904	5.352677	0.0	0.0	0.0	0.000000	118.000000
NCD_1	28179.0	1.165442	5.222078	0.0	0.0	0.0	0.000000	118.000000
NCD_2	28179.0	1.192910	5.153191	0.0	0.0	0.0	0.000000	118.000000
NCD_3	28179.0	1.187409	4.983439	0.0	0.0	0.0	0.000000	118.000000
NCD_4	28179.0	1.169310	4.810775	0.0	0.0	0.0	1.000000	118.000000
...	...	...	...	...	...	...	...	...
AS(NAC)_2	28179.0	0.001663	0.005968	0.0	0.0	0.0	0.000567	0.153209
AS(NAC)_3	28179.0	0.001729	0.006051	0.0	0.0	0.0	0.000647	0.153209
AS(NAC)_4	28179.0	0.001762	0.005992	0.0	0.0	0.0	0.000683	0.147003
AS(NAC)_5	28179.0	0.001844	0.006233	0.0	0.0	0.0	0.000782	0.179334
AS(NAC)_6	28179.0	0.001834	0.006464	0.0	0.0	0.0	0.000711	0.187696

DATA PREPROCESSING

Checking missing values that is no missing values found and Re-named the column names based on the feature labels and Cleaned the buzz column to convert it to binary and Generated X and y – kept all the features in X dataset and kept the buzz column in y

MODEL TRAINING

Fit the selected regression model to the training data. This involves estimating the coefficients or parameters of the model that minimize the difference between the predicted values and the actual values of the dependent variable.

MODEL EVALUATION

Evaluate the performance of the trained regression model using appropriate evaluation metrics. Common metrics for regression models include mean squared error (MSE), mean absolute error (MAE), R-squared, or root mean squared error (RMSE). Compare the model's performance on the test set against the evaluation metrics.

RESULT & DISCUSSION

We can see the r2_score for Random Forest Regression is more that is 0.9721892380024579 and bad in support vector regression that is -0.01560834962326005

CONCLUSION

We conclude that Random Forest Regression is the best model for predicting the buzz in social media.

References

Buzz in Social Media | Companion Proceedings of the The Web Conference 2018 (acm.org)