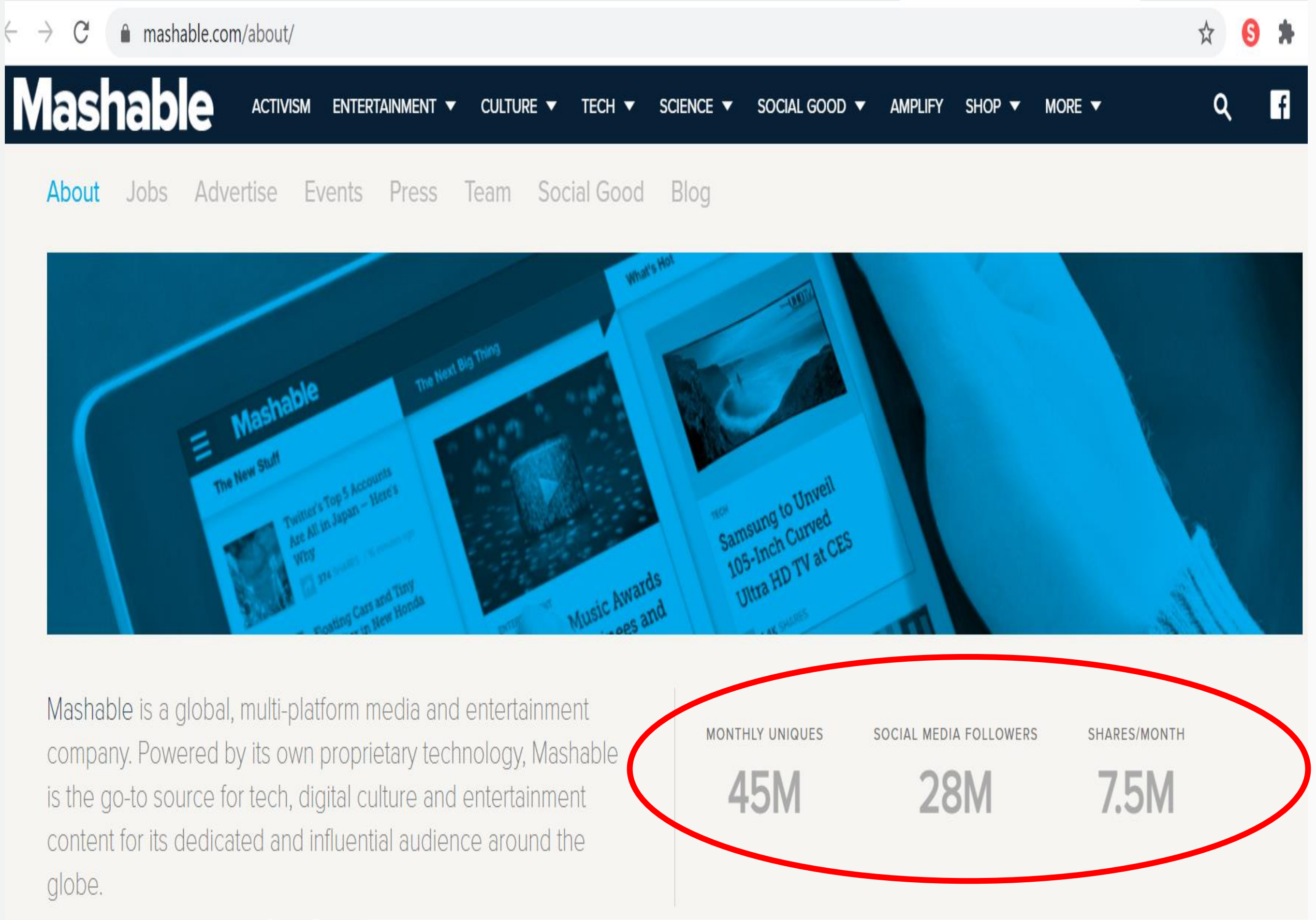
Capstone Project
News Polarity

Woon Yong & JinMei

About Mashable

- Mashable is the largest independent online news site dedicated to covering digital culture, social media and technology.
- With more than 40 million unique monthly visitors, Mashable has one of the most engaged online news communities.
- Mashable current primary competitors are BuzzFeed, Verge and TechCrunch.
- Other than the numbers or how many articles can be shared, polarity of news also play an important role that helps people to choose the ideal articles directly based on their search.



Mashable is a global, multi-platform media and entertainment company. Powered by its own proprietary technology, Mashable is the go-to source for tech, digital culture and entertainment content for its dedicated and influential audience around the globe.

MONTHLY UNIQUES	SOCIAL MEDIA FOLLOWERS	SHARES/MONTH
45M	28M	7.5M

Problem Statements

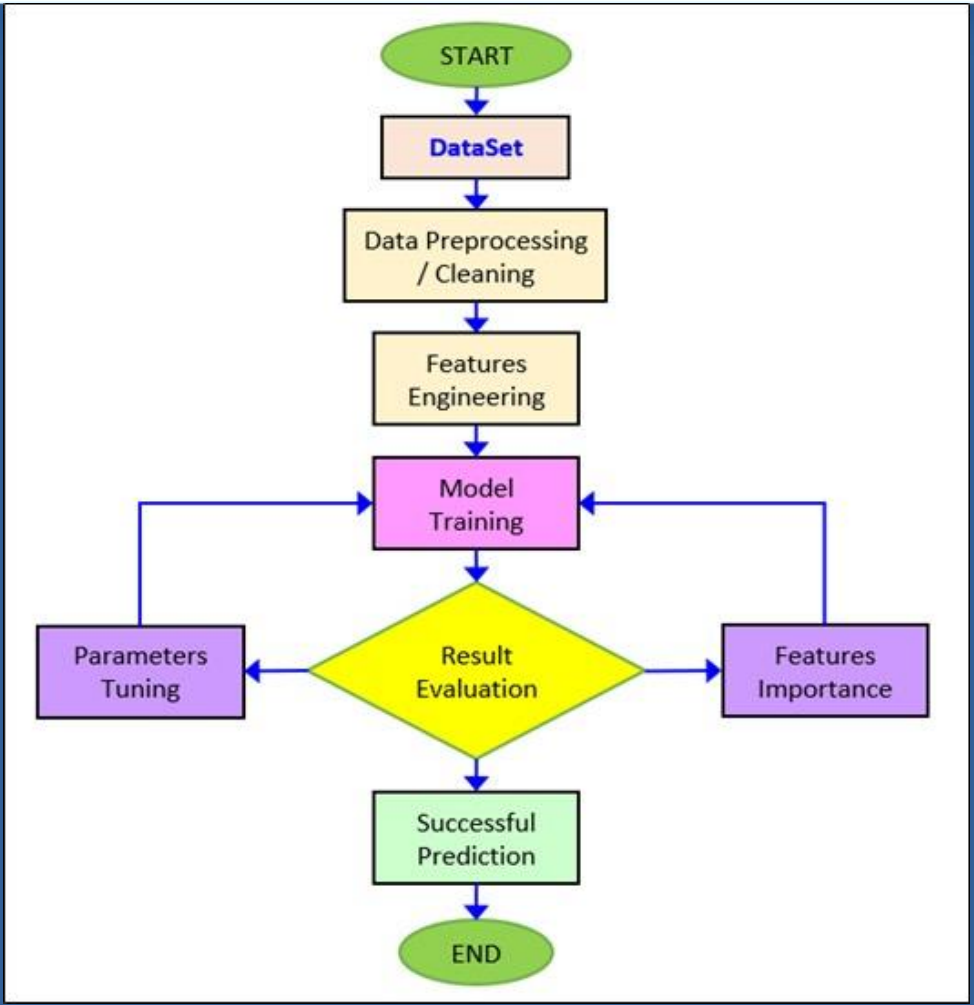
- Huge Number of Articles published Daily
- Time Consuming to check Published Articles
- Slow Response to re-act if articles are Negative

Objectives

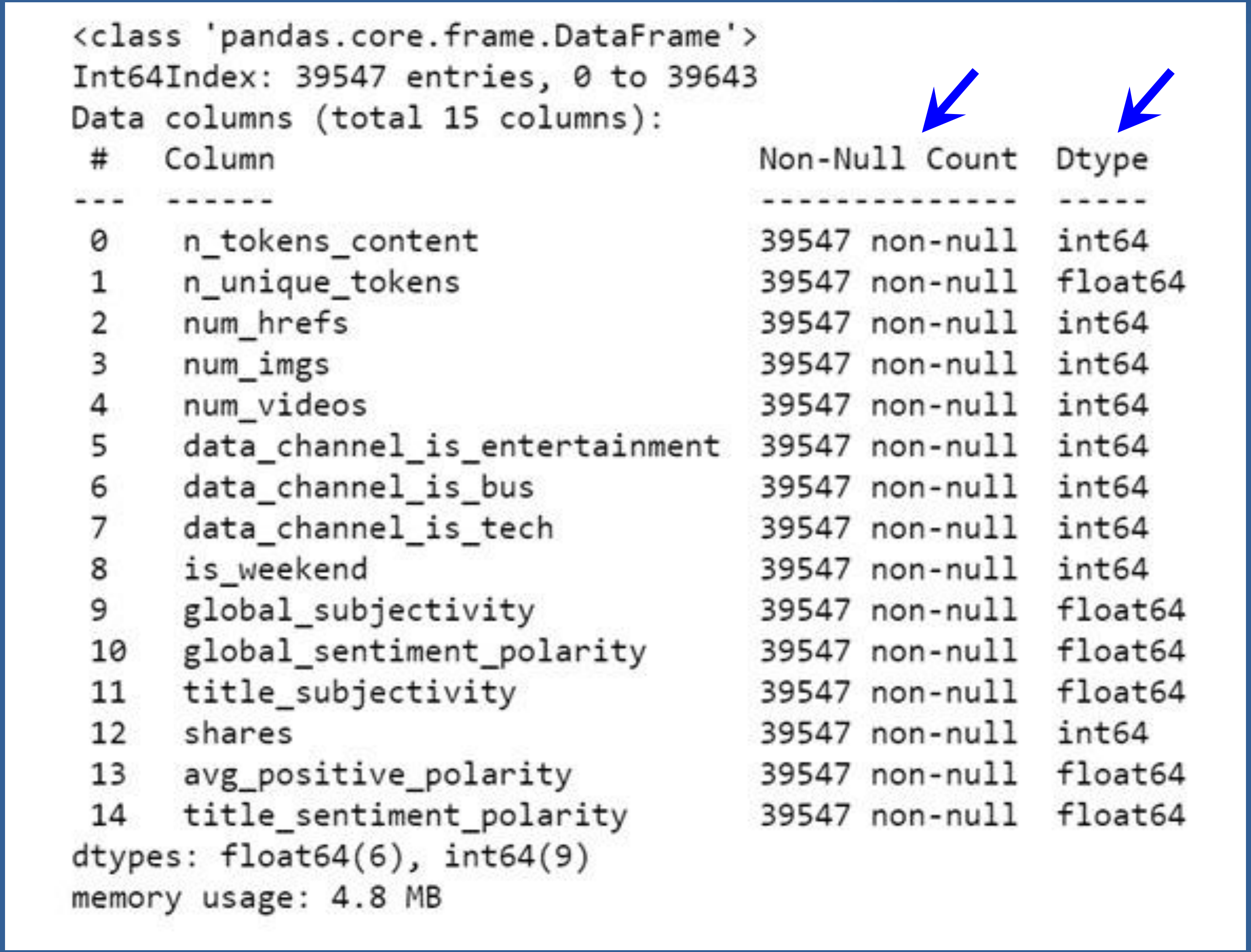
- Understanding the Factors that influence the sentiments of the Articles
- Using Machine Learning to automate the news articles polarity
- To keep up with the fast-paced news environment
- To push positive news out to the targeted audiences

Data Definition

No	Data	Description
1	n_tokens_content	Number of words in the content
2	n_unique_tokens	Rate of unique words in the content
3	num_hrefs	Number of links
4	num_imgs	Number of images
5	num_videos	Number of videos
6	data_channel_is_entertainment	Data channel 'Entertainment'
7	data_channel_is_bus	Data channel 'Business'
8	data_channel_is_tech	Data channel 'Tech'
9	is_weekend	Article published on the weekend
10	global_subjectivity	Text subjectivity
11	global_sentiment_polarity	Text sentiment polarity (above 0 is towards positive, below 0 is towards negative)
12	title_subjectivity	Title subjectivity
13	shares	Number of shares
14	avg_positive_polarity	Avg. polarity of positive words
15	title_sentiment_polarity	Title polarity



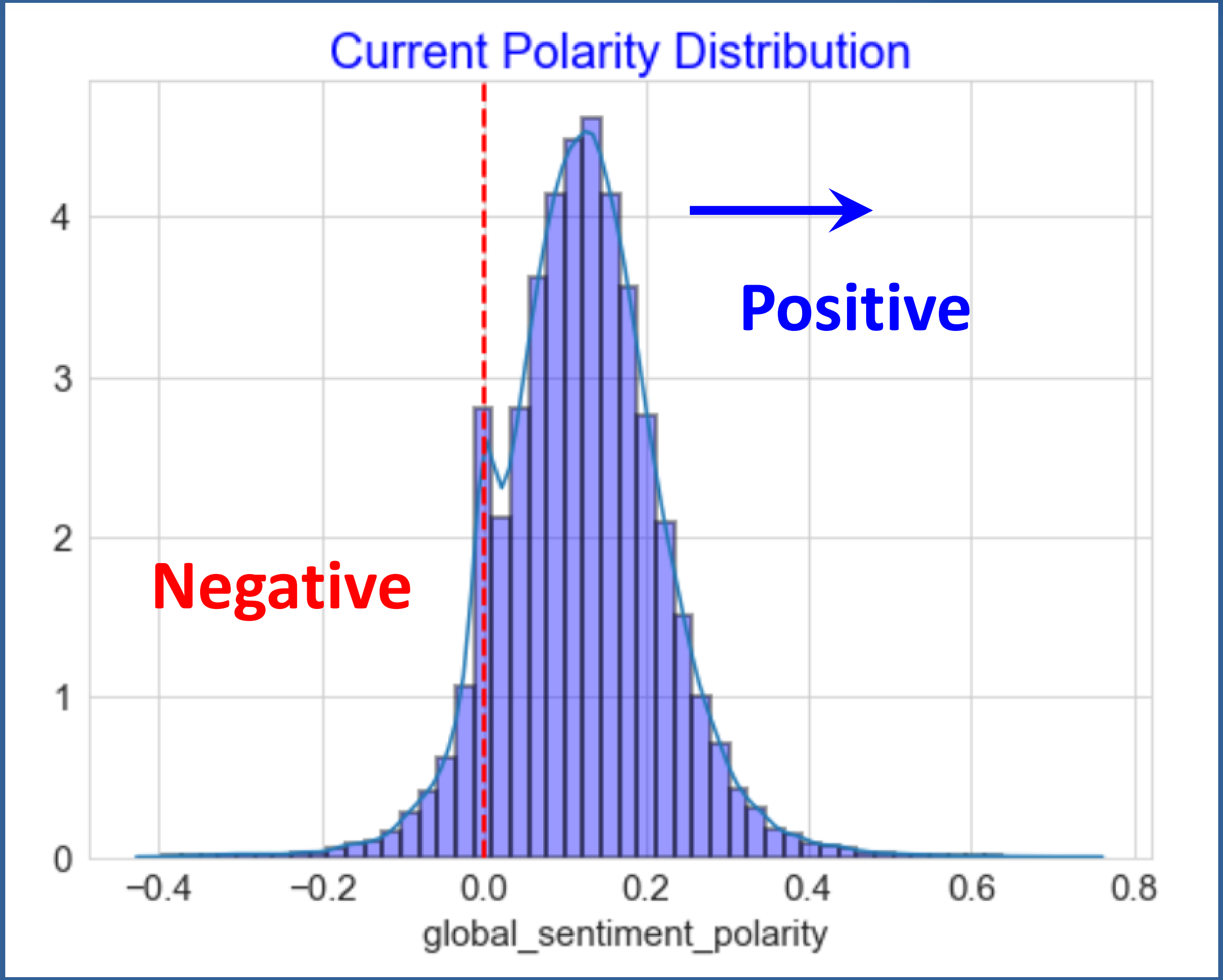
Data Understanding



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 39547 entries, 0 to 39643
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0    n_tokens_content             39547 non-null  int64  
 1    n_unique_tokens              39547 non-null  float64
 2    num_hrefs                    39547 non-null  int64  
 3    num_imgs                     39547 non-null  int64  
 4    num_videos                   39547 non-null  int64  
 5    data_channel_is_entertainment  39547 non-null  int64  
 6    data_channel_is_bus          39547 non-null  int64  
 7    data_channel_is_tech         39547 non-null  int64  
 8    is_weekend                   39547 non-null  int64  
 9    global_subjectivity          39547 non-null  float64
 10   global_sentiment_polarity    39547 non-null  float64
 11   title_subjectivity           39547 non-null  float64
 12   shares                       39547 non-null  int64  
 13   avg_positive_polarity        39547 non-null  float64
 14   title_sentiment_polarity     39547 non-null  float64
dtypes: float64(6), int64(9)
memory usage: 4.8 MB
```

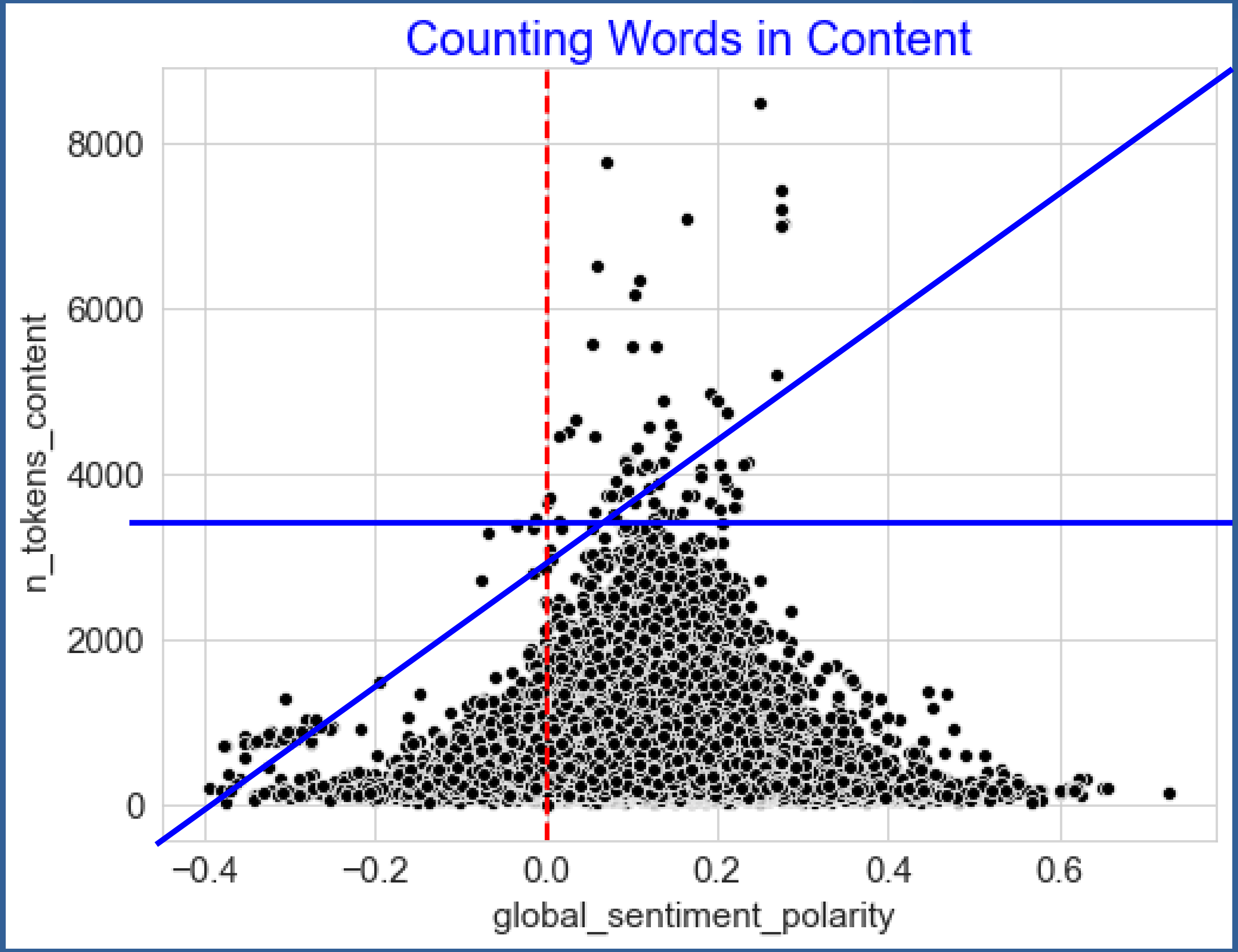
- Dataset is clean
- No missing values and all numerical type
- Apply One-Hot Encoding and Standard Scaler

Polarity Distribution



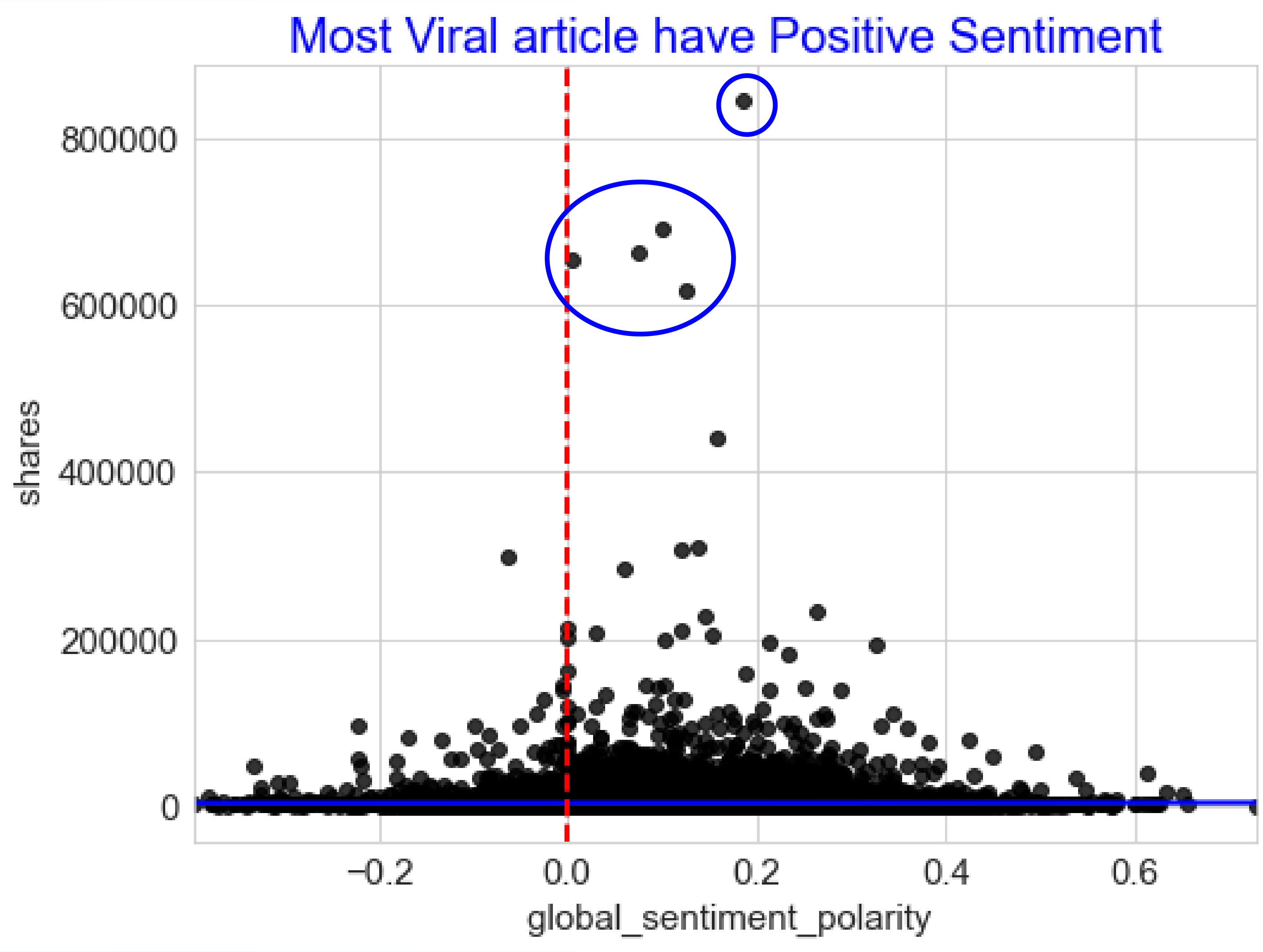
- Normal distribution (bell curve) indicating the model result could have good Quality for prediction.
- Majority of the articles are positive

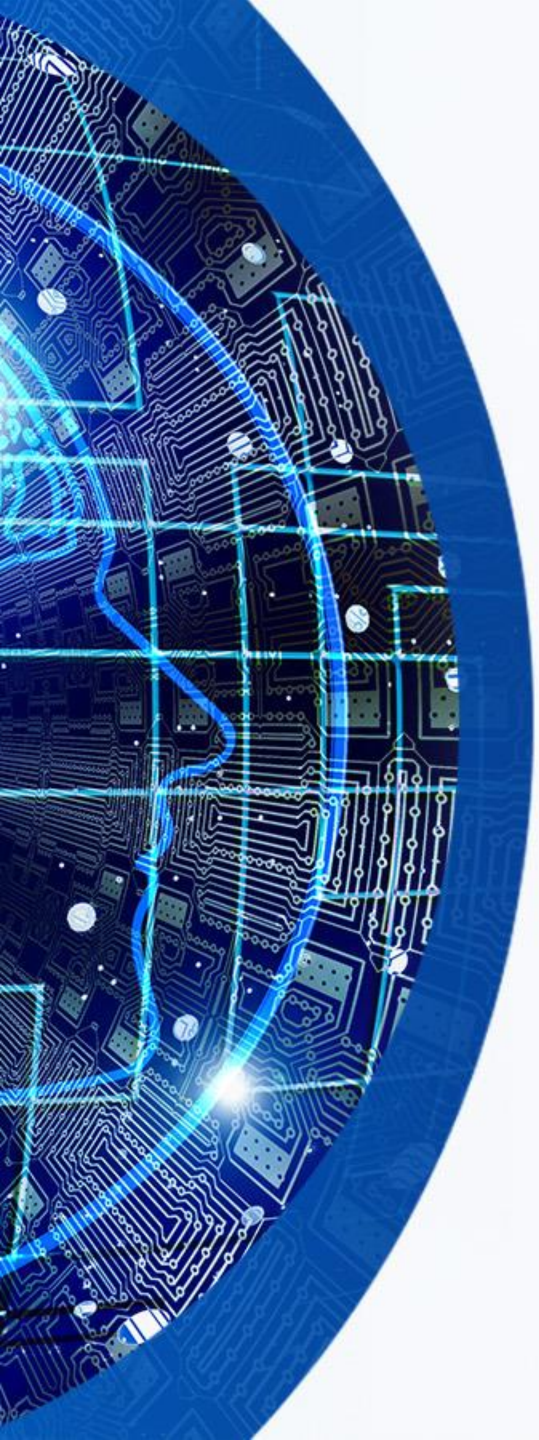
Counting Words in Content



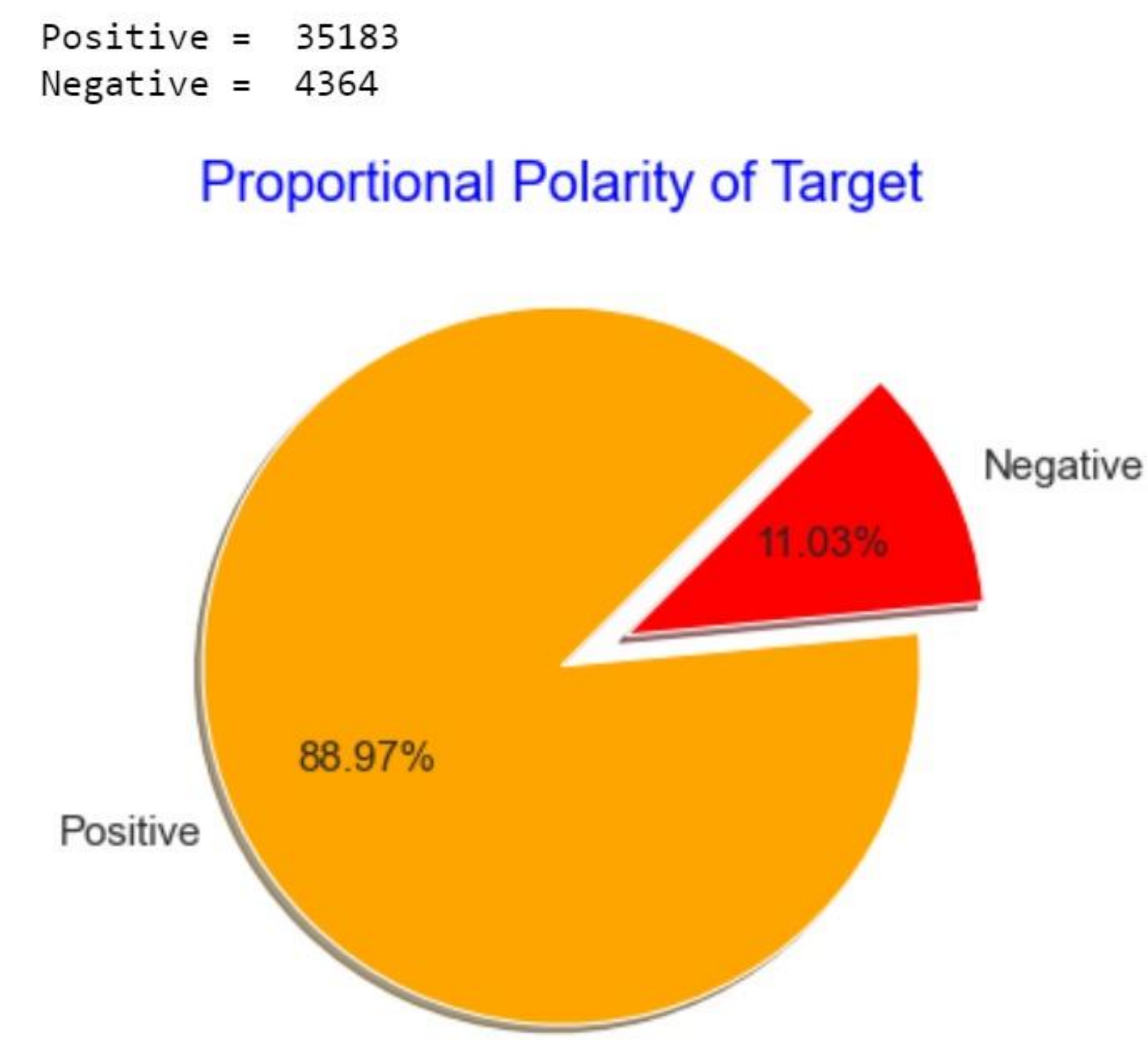
- Increase of positive when increasing the number of words in the content
- Articles within 3500 words are mostly positive

Most Viral Articles are Positive



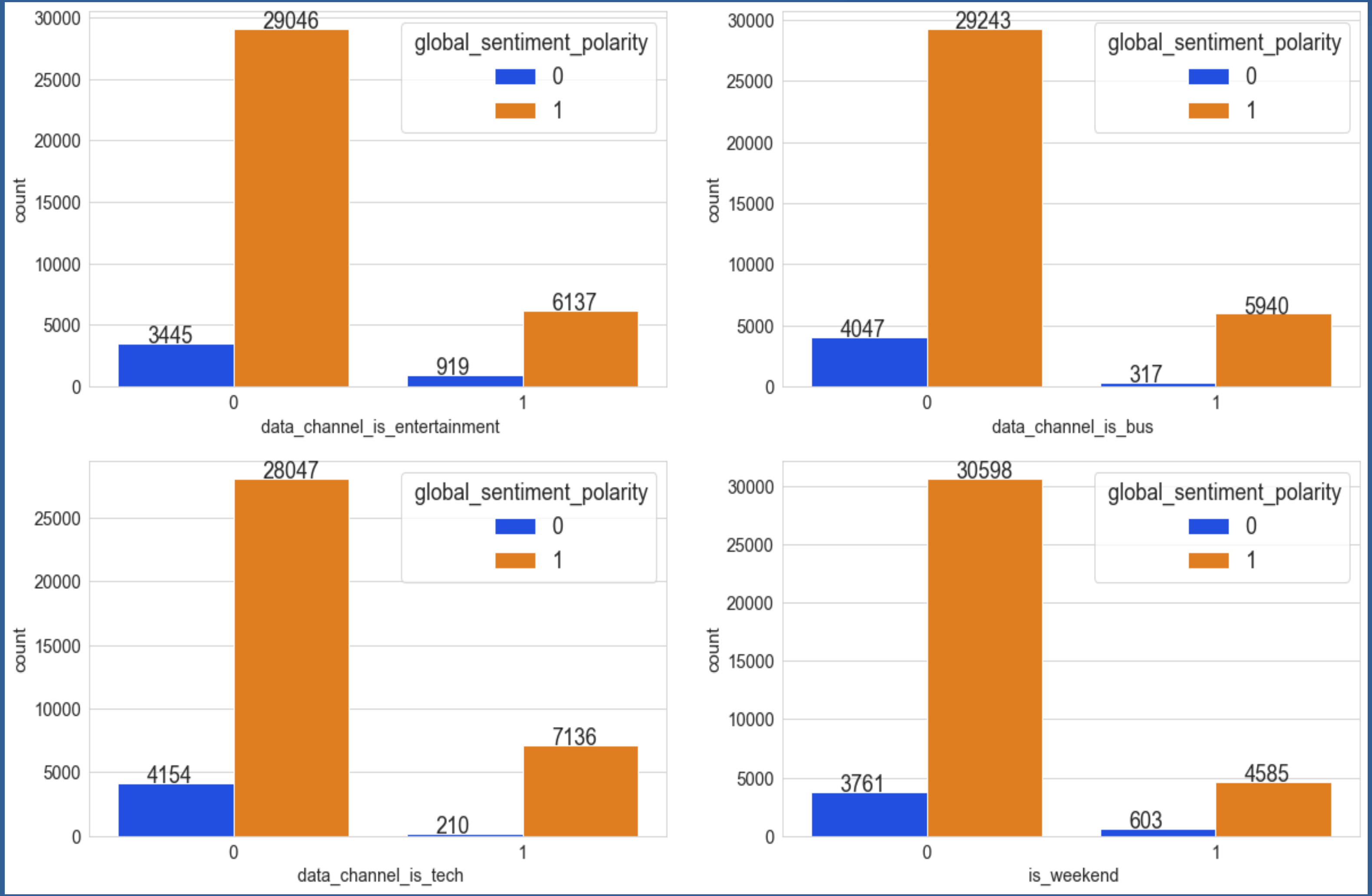


88% of Target are Positive



```
Positive =  35183
Negative =  4364
```

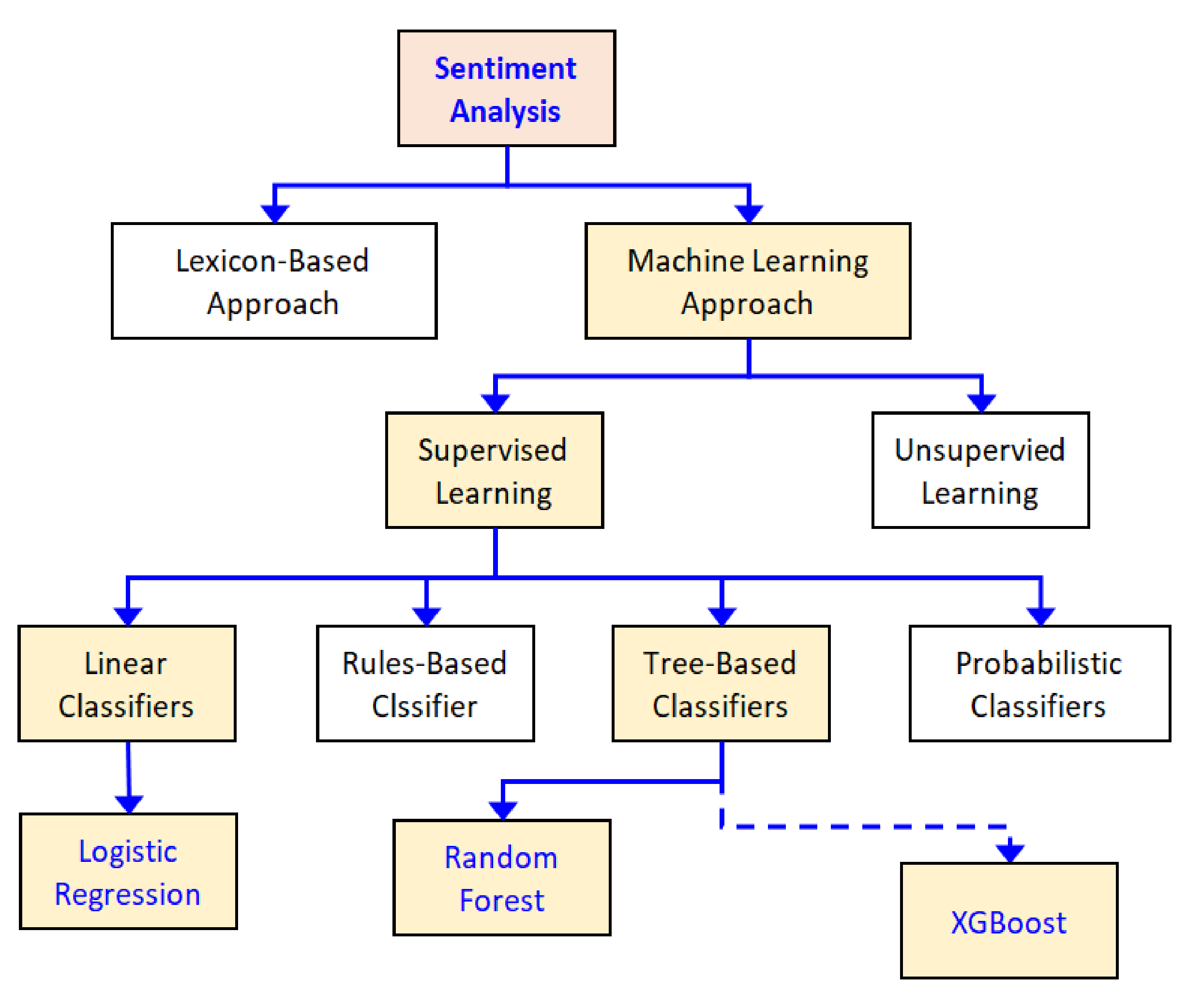
Different Relationship



- Articles published are mostly positive. (Entertainment, business, technology and etc)
- Articles published on weekend / weekday are positive.

Model Selection

- Logistic Regression
- Random Forest
- XGBoost



Sentiment Analysis

Lexicon-Based Approach

Machine Learning Approach

Supervised Learning

Unsupervied Learning

Linear Classifiers

Rules-Based Clssifier

Tree-Based Classifiers

Probabilistic Classifiers

Logistic Regression

Random Forest

XGBoost

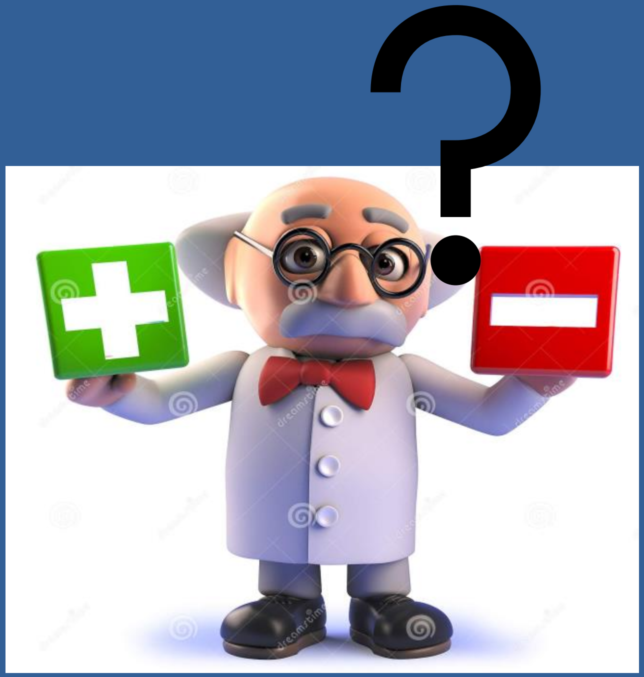
Evaluation of Model

Precision

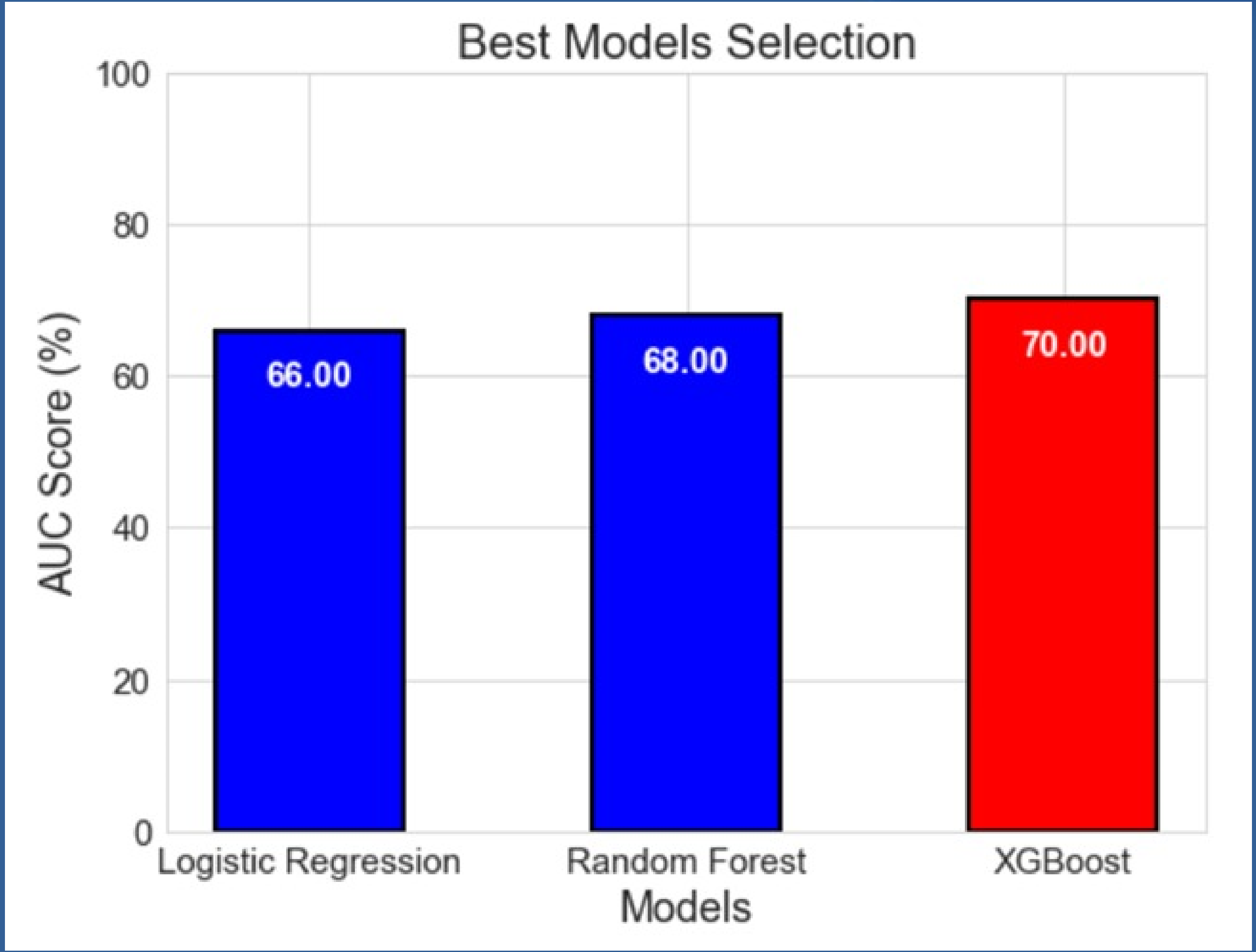
Recall

AUC-ROC

- Predicting the Positive
- Capable of the Model



Model Selection

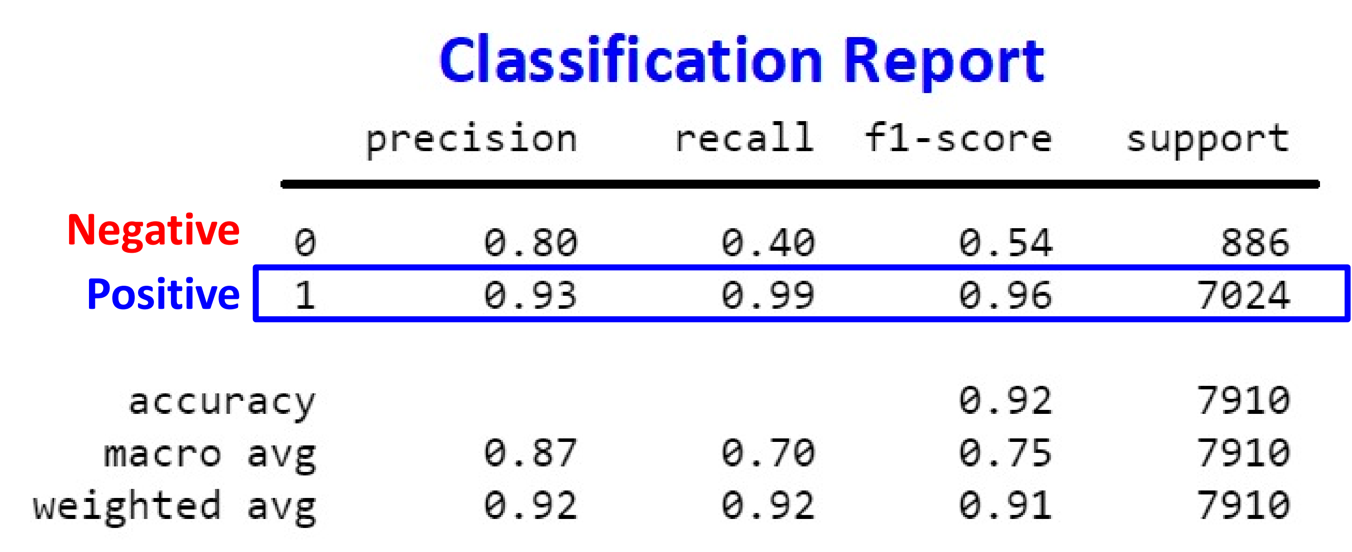


Result :

Based on the AUC-Score

- Logistic Regression = 66.00
- Random Forest = 68.00
- XGBoost = 70.00

Performance Measurement

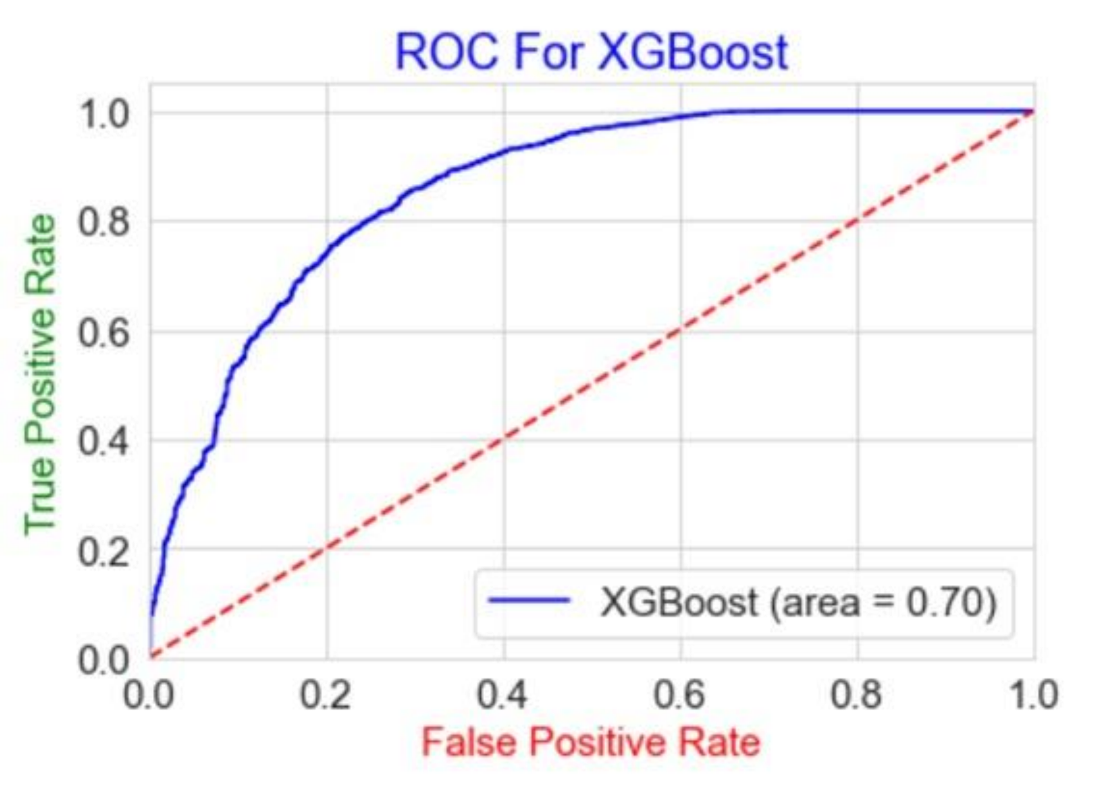


Classification Report

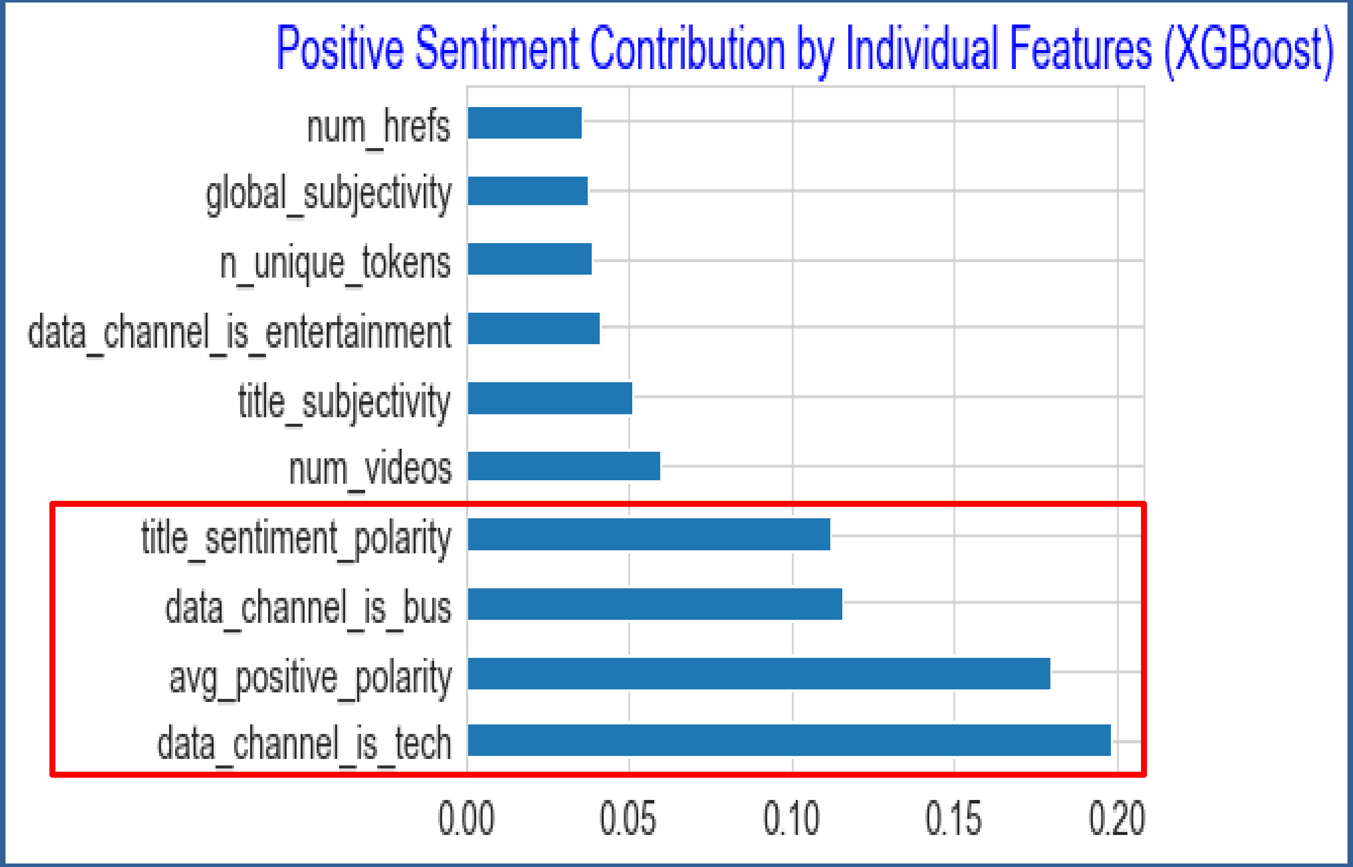
```
                  precision    recall  f1-score   support

Negative   0           0.80      0.40      0.54       886
Positive   1           0.93      0.99      0.96      7024

    accuracy                               0.92      7910
   macro avg           0.87      0.70      0.75      7910
weighted avg           0.92      0.92      0.91      7910
```



Feature Importance



- 4 features with high score are importance for prediction
- Features related to technology and business, having good title and more positive polarity

Conclusion

- Majority of the Articles are Positive (around 88%).
- Length of articles within 3500 words, mostly are positive and most positive articles are viral.
- Conclude that XGBoost is providing the best results with the AUC-ROC score at 70%.
- With the automated predicted model in place, Mashable can easily use it to determine the polarity of the articles.
- This helps to save time and keep up with the fast-paced digital media environment.
- Positive news can also be used for targeted audiences of Mashable.
- Negative news can be filter and action can be taken before it spreads.

Recommendation

Maintain

- To maintain the current segregation of the articles in the channel (technology, entertainment and business) as result in the pie chart indicating strong positive (88%).

Improve

- With the automated model and available resources, further improvement of the other features which are weak.

Accelerate

- Explore other markets and include new business model like Eshop or partner with competitors to target different segment / services / products.

Thank You!

Q&A