New York Times

Comments analysis

May 2019

Paul-Renaud Raymond

Image source: NYTimes

Why are certain **article comments highlighted by NYTimes editors**? The actual content, or something else?

Objective: **Develop and compare text and feature classification models to predict whether a user-submitted article comment is selected by NYTimes editors**

New York Times Company

Executive summary

Datasets included ~2M rows of records

Used two datasets provided by The New York Times: Articles written and Comments submitted during both Jan-April 2017 & Jan-April 2018

Difference in number of articles and comments between 2017 and 2018

There is marked uptick in total number of articles written in the first months of 2017 relative to 2018. Comments have slight increase as well.

High performance for classification using non-text features with oversampling of classes

Classification with Pipelines indicate high Accuracy, Precision, Recall, F1, and Precision – after data is resampled for balance

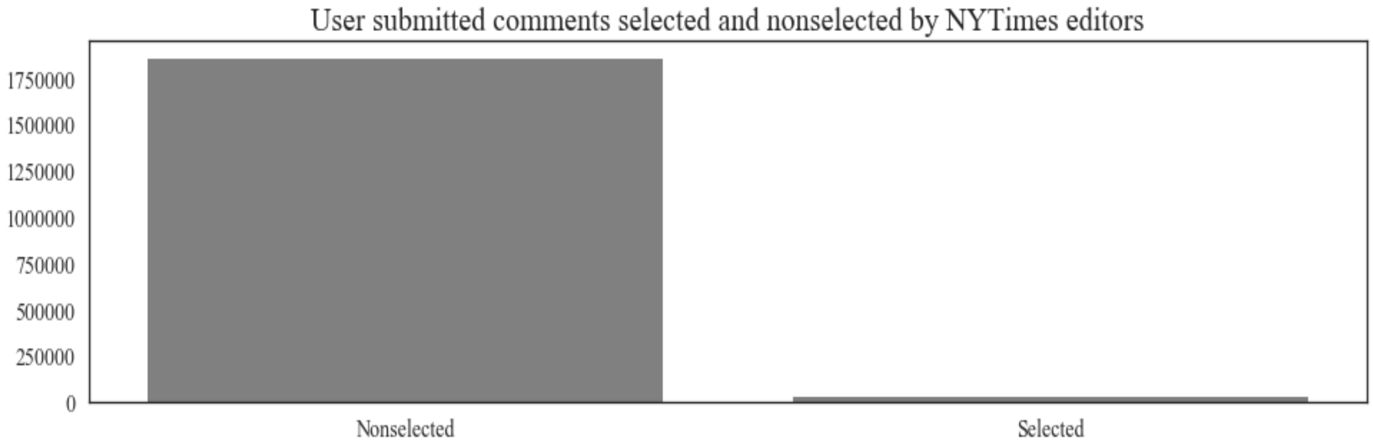
High performance in classification of “Editors Selected” comments with NLP analysis

NLP classification model with scikit-learn produces strong results with synthetic samples (XX% recall). Results for undersampled classification with much less accuracy yet similar recall (XX%) and AIC-BIC(XX%)

Exploratory data analysis: Some data and meta-comments

	Articles ***during Jan – April (2017 & 2018)***	**Comments**
2 Datasets:	• **8,339 articles** written • **17 features** for each article	• **1,899,975 comments** written • **38 features** for each comment
Range:	• **Longest article:** “On the shooting in Florida, Student Activism..” **(16,336 words)** • **Shortest article:** Unknown **(11 words)** • **Most articles: Deb Amlen** • **Most words: Editorial Board**	• **Longest comment:** “Beautifully written and deeply touching. It makes me wonder how many of us also have college classmates who died while homeless and suffering from mental illness..” (452 words) • **Shortest comment:** (4232 with one word)
	• **News** was type of material with **most articles (5,596)** • **OpEd** was category with **most articles (1,528)**	• **News** was type of material with **most comments (1M+)** • **OpEd** was category with **most comments (671893)**

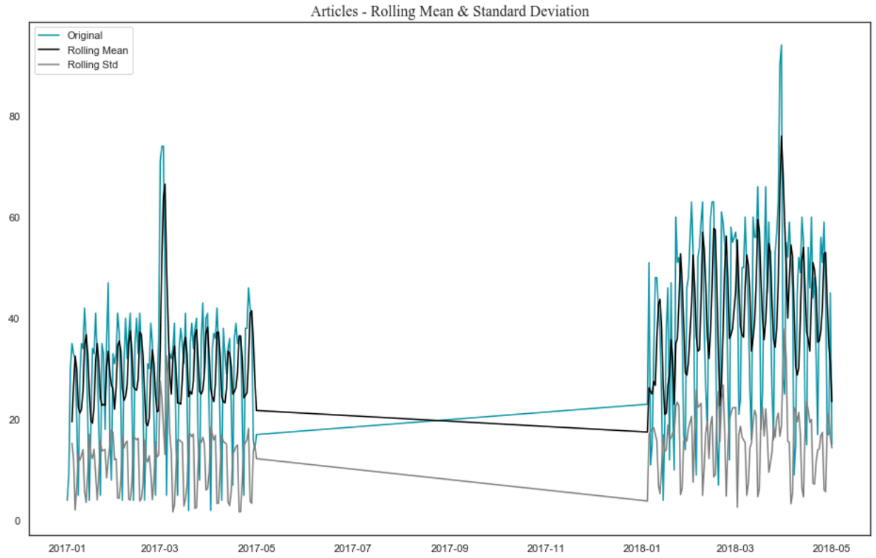
Feature class breakdown: Selected vs non-selected comments



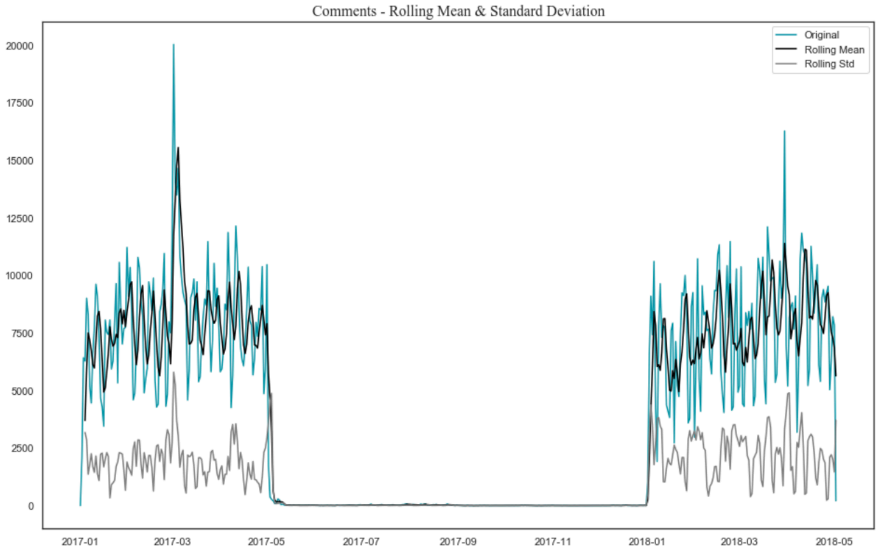
Less than 2% of the comments are selected by NYTimes editors. As the target feature, there will be a class imbalance issue to address for accurate classification models

Question 1: Is there a statistically significant difference in the number of articles and comments in Q1 2017 and Q1 2018?

Total articles



Total comments



Visually, there appears to be a considerable difference in articles between the two periods; not as clear for comments. Dickey-Fuller test for both features over time verifies lack of stationary trends.

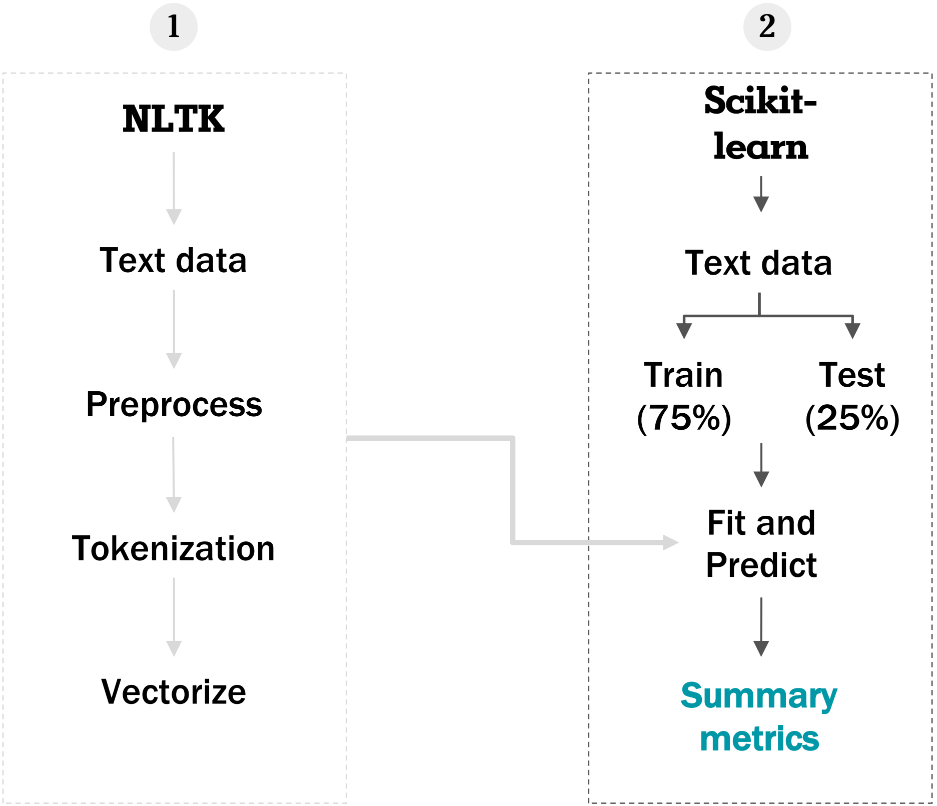
With data and trends over time is established, the key question is why editor selected comments are deemed better than non-selected comments.

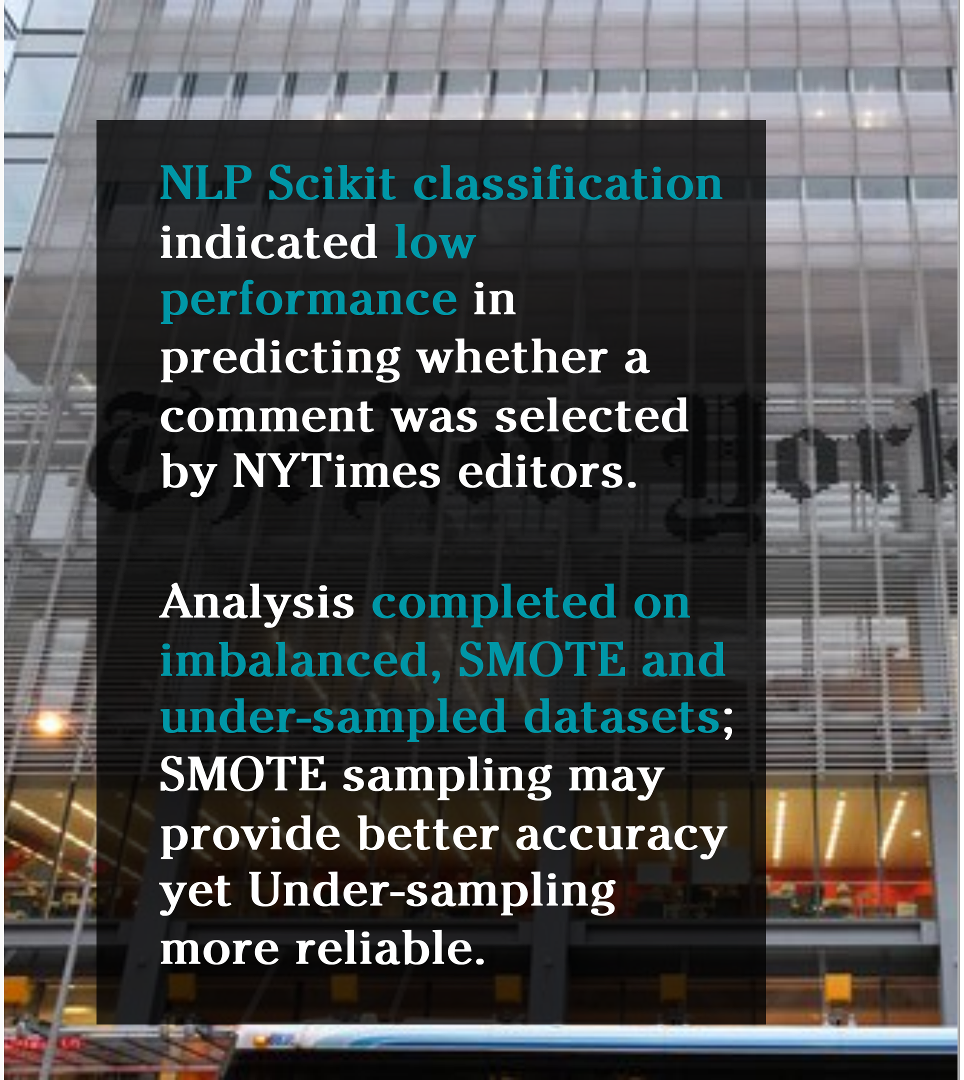
This question will be explored with two lenses: 1. content of comment, and 2. non-content features of comment

Driver tree for analysis

Question 2:
How well can we predict whether a comment is highlighted by NYT editors, based solely on its content?

NLP Analysis (In two phase approach)



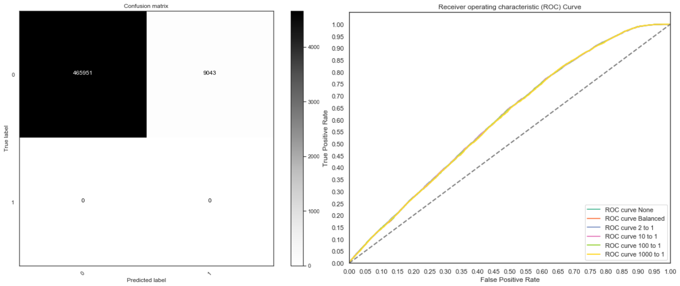


NLP Scikit classification indicated low performance in predicting whether a comment was selected by NYTimes editors.

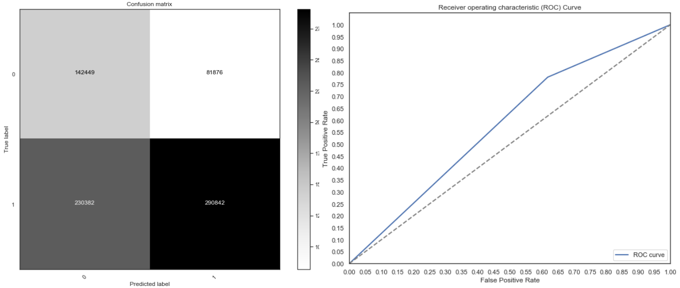
Analysis completed on imbalanced, SMOTE and under-sampled datasets; SMOTE sampling may provide better accuracy yet Under-sampling more reliable.

scikit-learn model output

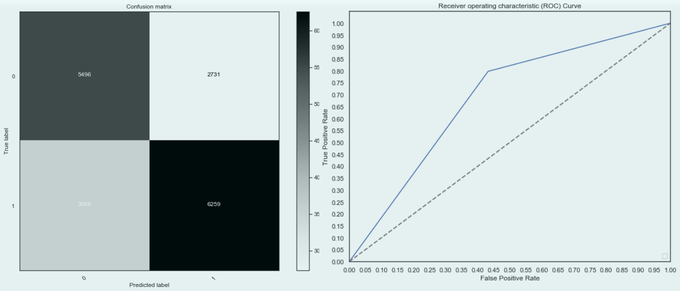
Baseline model



Resampled (SMOTE) model



Under-sampled model

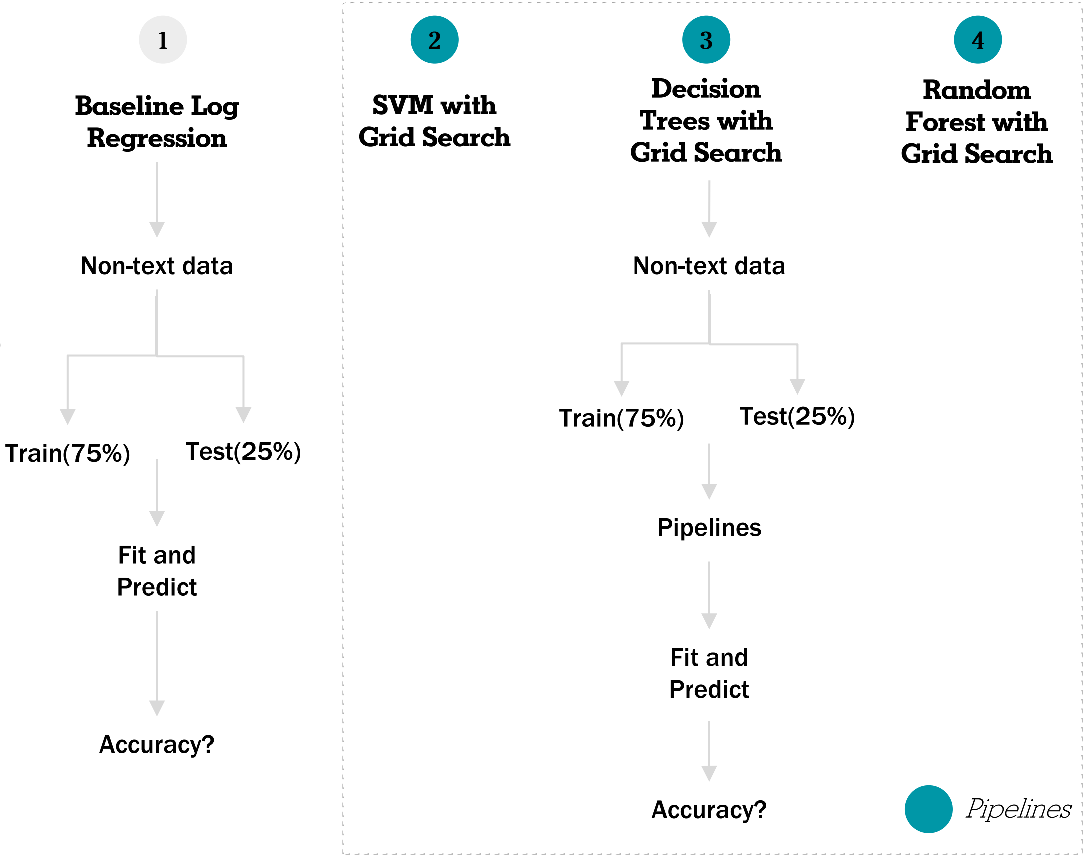


Best results

Question 3:

How well can we predict whether a comment is highlighted by NYT editors, based on features other than its content?

Four Classification Approaches



1 Baseline Log Regression

Non-text data

Train(75%)

Test(25%)

Fit and Predict

Accuracy?

2 SVM with Grid Search

3 Decision Trees with Grid Search

Non-text data

Train(75%)

Test(25%)

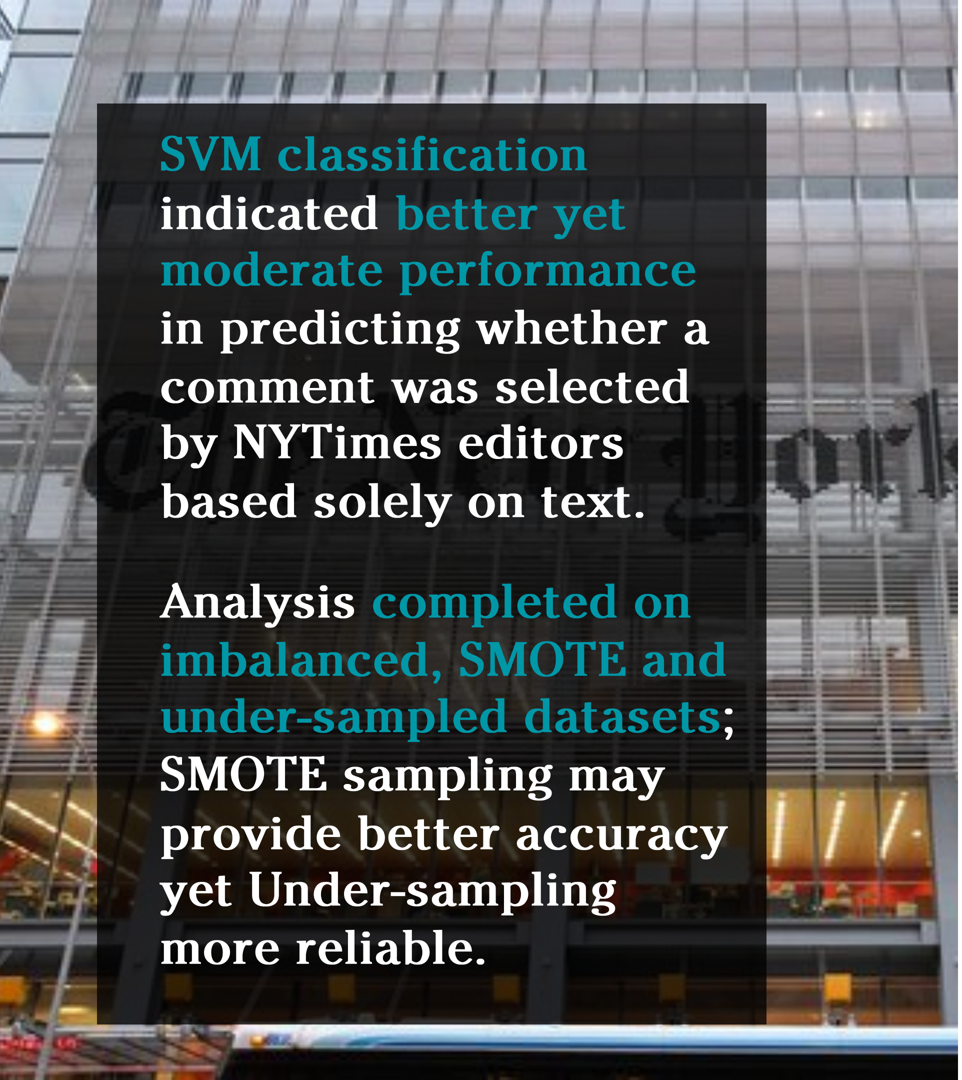
Pipelines

Fit and Predict

Accuracy?

4 Random Forest with Grid Search

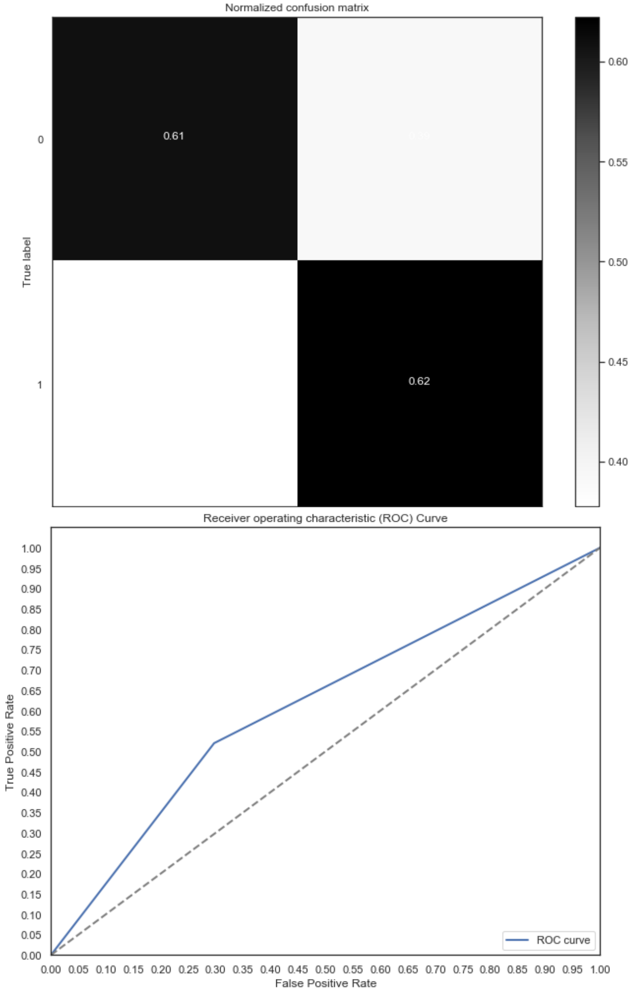
● *Pipelines*



SVM classification indicated **better yet moderate performance** in predicting whether a comment was selected by NYTimes editors based solely on text.

Analysis **completed on imbalanced, SMOTE and under-sampled datasets**; SMOTE sampling may provide better accuracy yet Under-sampling more reliable.

Top model (SVM) output



Future analyses that can unlock additional insights

Evident trends in article and comment submissions by category and type of material

Improvement in model performance with neural network models

Sentiment analysis of articles and comment submissions over time