Contents

Sentiment Analysis of "The Lion King (2019)" Movie Reviews

Problem Description:

Online reviews are important because they have become a reference point for buyers across the globe and because so many people trust them when making purchase decisions.

Reviews are also important for Search Engine Optimization (SEO). Having positive reviews is also another way through which you can improve a website's Search Engine visibility. The more that people talk about a brand online, the greater its visibility to Search Engines, such as Google, Yahoo and Bing.

For the audience and booking websites, analysing reviews is significant in understanding reviewer opinion about the film. In movie booking websites, 90% of people first check out online reviews before purchasing tickets.
For the production house, analysing negative reviews can be useful for damage control.

In this Competition, you are expected to create an end to end NLP framework to collect, analyse and perform sentiment analysis using supervised learning on user reviews about the latest Hollywood flick - "The Lion King (2019)"

About Data:

Train Data will not be provided, participants are required to collect the data using the instructions provided in the document: Data_Collection_Instructions.docx

Test Data will be provided (columns: 'ReviewID', 'Review'), which should be used for tool evaluations.

Objectives:

In this hackathon, you are expected to:

1. Collect Audience reviews from "www.rottentomatoes.com" for the film:
"The Lion King (2019)"
2. Label the Review Sentiment (Target) – Refer to Main Tasks
3. Perform the Visualizations & EDA on the data gathered.
4. Perform Sentiment Classification using supervised learning
5. Clustering the Reviews – Comparing 'Cluster Label' with Train data Target 'sentiment'

Main Tasks:

1. Collect Audience reviews from "www.rottentomatoes.com" for the film: “The Lion King (2019)” with at least the following features.

a. ReviewID
b. Reviewer Name
c. Review
d. Rating
e. Date-of-Review

The feature names above may not match with the exact tags returned by the response object. Use your intuition to map them and create the above features accordingly.

You are required to collect 3000 reviews only.
You are free to collect any other attributes/features that you think helpful.

“Participants are NOT advised to visit INSOFE office to perform the data collection”
(as the server may block the IP due to many requests from same IP within short interval of time.)

2. Label the Review Sentiment:

You must label the data based on the following condition on Rating:

if ‘Rating’ > 3 then positive review else negative review – Create target attribute with the name: “sentiment” (binary class)

“Drop the Rating attribute once the Target is derived. It should not be part of model building as independent attributes.”

3. Exploratory Data Analysis using visualizations in R Notebook or Jupiter notebook format. (train data to be used for this)

a. What is that the good and bad, people are talking about the film?
(Hint: you may pick any meaningful n-grams and obtain their frequency to emphasize the importance of n-gram)

b. Any other meaningful Insights

4. Perform Sentiment Classification using supervised learning algorithms

a. Identify the best model using traditional Classification ML algorithms like NaiveBayes, Logistic Regression, SVM, Decision Trees, Ensembles etc.

b. Identify the best model using Deep Learning Classification ML algorithms like CNN, RNN/LSTM etc.

Choose the model that outperforms all others for your test predictions and in your submissions.

5. Clustering the Reviews – Comparing 'Cluster Label' with Train data Target 'sentiment'

a. Take only 'reviews' attribute from train data and label them into two clusters using any clustering algorithm of your choice.
b. compare the cluster labels with the train data target attribute - 'sentiment' and Write a brief comparison report with your observations.

Evaluation Metric:

F1-score for Negative reviews.

Important Note for the results submission:

Note: While evaluating the predictions submitted, the system will consider "1" as positive level in target attribute and hence please convert the target attribute accordingly and submit the results. It is very important for this problem as the error metric is "F1 statistic" for the target attribute level "negative review". Refer to the samplesubmission.csv file ('0': positive review and '1': negative review).

Submission Timelines:

Submission No	File	Submission Format	Start Date	End Date
Submission - I	Exploratory data Analysis of data	R Notebook or Jupiter notebook	18-Aug 9:00 (Sun)	23-Aug 20:00 (Fri)
Submission - II	Predictions of test.csv (Target attribute: sentiment)	samplesubmission.csv	24-Aug 9:00 (Sat)	25-Aug 20:00 (Sun)
Submission - III	Improved version of predictions	samplesubmission.csv	26-Aug 9:00 (Mon)	28-Aug 20:00 (Wed)
Submission - IV	final report including all tasks along with clustering and comparison report.	Zip file format or R Notebook or Jupiter notebook	29-Aug 9:00 (Thu)	30-Aug 20:00 (Fri)