Sentiment Analyzer Case Study

by using IMDB Dataset of Movies Reviews

By: Anmol Aman

IMDB Dataset of Movies Reviews

Data Overview

Data Source : https://www.kaggle.com/datasets/crisbam/imdb-dataset-of-65k-movie-reviews-and-translation

Basic Statistics of Data:

IMDB review dataset contained four columns: Ratings, Reviews, Movies, Resenhas

- **Number of Reviews:** 149780
- **Number of Movies:** 14205

Attribute Information:

1. **Review:** User review in English Language
2. **Ratings:** rating between 1 to 10
3. **Movies:** Movie names
4. **Resenhas:** User review translation in Portuguese Language

Objective

The IMDB dataset goal is to provide a base for sentiment analysis. Usually, datasets of this kind provide binary classes telling whether the text is positive or negative however this approach fails in some points. We don't know how the data was linked to those labels and the queries used

Our Goal:
In this case study, we will focus only on rating, review attribute to develop a sentiment analyzer model to categorize review either positive or negative.

1. Map rating with 1 to 10 to binary classes (positive and negative).
2. Try to solve real time challenges related to sentiment Analyzer : Sarcasm, multi polarity and negation.
3. Develop NLP and ML model to categorize English text revies either positive or negative.
4. Deploy Sentiment Analyzer model on AWS cloud by using test API.

Mapping to Data Science Problem

Buisness problem: Our task is to find sentiment of end user by using NLP based model. Our NLP model outcome would be binary class either positive or negative. But, we have rating attribute value 1 to 10 scale. So, we need to convert this problem into binary class text classification problem.

Map Ratings range to Binary classes:

- 1 to 4: Negative
- 5 to 6 Neutral
- 7 to 10: Positive

As, we are interested in positive and negative sentiment. So, we will remove neutral sentiment data from our dataset.

Final Outcome:

- Sentiment Value (either positive or negative).
- Movie attribute liked by majority of users.
- Movie attribute not liked by majority of users.

Challenges while developing Sentiment Analyzer

Sarcasm : Some time user provide their negative review by using positive words. Simple Sentiment analysis model can easily fool by this kind of sarcasm. So, we need to design our sentiment analyzer to detect sarcasm of user review.

Example:

- Simple Review: My car has an awesome mileage of 25 KM per liter.
- Sarcastic Review: My car has an awesome mileage of 4 KM per liter.

Polarity: Sometime user provide multipolarity review then traditional Sentiment analyzer system may not work. So, we need to design Sentiment analyzer to get valuable information from multipolarity reviews.

Example:

- The screenplay, cuts, action and sound effect deserve a standing ovation but the story is not too good.

Negation: Some user use non, less, not, never, cannot to reverse the polarity of review or word. So, we need to use state of the art ML model to handle this kind of scenarios.

Example:

- I don't call this movie a good movie.

Thank You.