

Sentiment Analysis

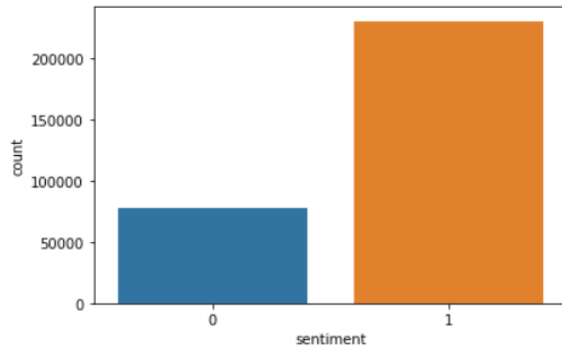
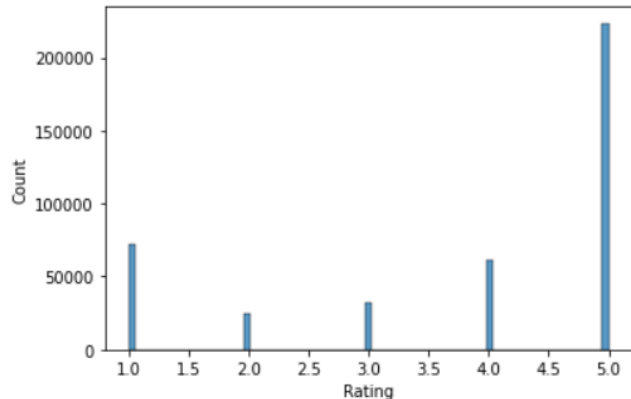
Abstract:

The study of public opinion can provide us with valuable information. The analysis of sentiment on social networks, such as Twitter or Facebook, has become a powerful means of learning about the users' opinions and has a wide range of applications. Sentiment analysis is the task of classifying the polarity of a given text. For instance, a text can be categorized into either "positive", "negative", or "neutral". Given the text and accompanying labels, a model can be trained to predict the correct sentiment.

The efficiency and accuracy of sentiment analysis is being hindered by the challenges encountered in natural language processing (NLP).

Dataset:

We are working with the [Amazon Reviews](#) dataset(413840 rows \times 6 columns)



Methods:

Data cleaning

- Converted into lower cases
- Excluded the characters using RE.
- Removed the stop words
- Performed Stemming and Lemmatization

Vectorizing the data

❖ Counter Vectorizer

Convert a collection of text documents to a matrix of token counts. This implementation produces a sparse representation of the counts using `scipy.sparse.csr_matrix`.

❖ Tf-idf Vectorizer

Convert a collection of raw documents to a matrix of TF-IDF features.

Equivalent to [CountVectorizer](#) followed by [TfidfTransformer](#).

Machine Learning models

❖ ML Naive bayes

Naive Bayes classifier for multinomial models.

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

❖ Logistic Regression

Logistic Regression (aka logit, MaxEnt) classifier.

Deep Learning models

❖ LSTM

Long short-term memory is an artificial [recurrent neural network](#) (RNN) architecture used in the field of [deep learning](#). Unlike standard [feedforward neural networks](#), LSTM has feedback connections. It can process not only single data points, but also entire sequences of data.

Deployment using flask

Results:

*Classification LSTM				
	precision	recall	f1-score	support
0	0.94	0.95	0.95	19500
1	0.98	0.98	0.98	57570
accuracy			0.97	77070
macro avg	0.96	0.96	0.96	77070
weighted avg	0.97	0.97	0.97	77070

