BT5153 Applied Machine Learning for Business Analytics, Spring 2023

Kaggle Project

Sentiment Analysis of Movie Reviews

Submitted by:

Pragati Sangal

A0262745Y

E0997968

Introduction

Sentiment analysis is a branch of natural language processing that involves extracting subjective information from text data. It has become increasingly important in recent years due to the exponential growth of online reviews and social media content. In this report, we will be discussing our analysis of sentiment in movie reviews using machine learning techniques. Specifically, we will be exploring the process of data pre-processing, feature engineering, and model selection to develop accurate and robust sentiment analysis models for predicting the polarity of movie reviews. Additionally, we will be discussing how we validated the predictions of our models and arrived at the final model for our top two submissions. Overall, the goal is to perform sentiment analysis on movie reviews and predict whether the sentiment is positive or negative and provide a comprehensive overview of the steps we took and the insights we gained in performing sentiment analysis on movie reviews.

Data Understanding

The dataset for the analysis is taken from the Kaggle website (www.kaggle.com). It provided train and test dataset. The training dataset consists of pairs of sentences in English and their corresponding sentiments, has 24,995 records and 3 features including target variable "Sentiment". The target variable is a categorical variable, containing 0 or 1 so it is a binary classification of Movie Reviews based on given text. Each observation in the train dataset represents a text and corresponding sentiment label. The test dataset only contains a list of sentences on in which we need to perform inference.

In sentiment analysis, data pre-processing of given text is a major part as it helps to remove irrelevant or noisy data and focus on the most important words in the text. To enhance the quality of our text data and prepare it for further analysis, we devised a personalized pipeline for cleaning the text using the texthero package which involves replacing missing values with empty strings, converting all characters to lowercase, and eliminating any leading or trailing whitespace. We further removed several other elements from the text data such as URLs, digits, diacritics, punctuation marks, stopwords, round brackets, HTML tags, and various types of brackets. By carrying out these steps, we were able to standardize and refine the text data to a greater extent.

Based on data exploration, we observe that the train dataset is balanced as target feature has equal distribution of 0 and 1 classes *(figure 1 in Appendix).* For further analysis, after performing feature engineering on Text column, we plotted the chart on top ten most common words in the text, which gave us a quick overview of the most frequently used words in the dataset (*figure 2 in Appendix).* It is also useful to see any additional stop words that we can add to the stop words lists.

Model Building and Evaluation

1 Logistic Regression (Baseline model)

As a baseline model, we used logistic regression to classify the sentiment of text data. Logistic regression is easy to implement, interpret, and very efficient to train. Additionally, It's designed for balanced datasets. For this model, the data is split into 80% training and 20% validation dataset. The text data, received from feature engineering (using texthero package), is further transformed into a matrix of word counts using BoW (Bag of words) approach. The accuracy score of the model on the validation set is 0.88, indicating that the model is able to correctly predict the sentiment of the text with a high degree of accuracy. However, the score on test data is reduced to 0.86 (table 1 below).

2 Bert (Bidirectional Encoder Representations from Transformers) Model

After baseline model, we used pre-trained Bert model as It is pre-trained on a massive amount of text data and can generate high-quality contextualized embeddings for each word in a sentence, which can be used as input to downstream NLP tasks. For this model, the data is split into 80% training and 20% validation dataset. For cleaning the text data, instead of utilizing 'texthero' package, we employed the Bert Tokenizer directly because it results in better performance compared to other traditional text pre-processing techniques. Then the model is trained using the AdamW optimizer with a learning rate of 2e-5 and a batch size of 32 for different epochs. Model with 5 epochs gives the better result on validation and test datasets as compared to 3 or 7 epochs. After 5 epochs, the training loss gets increased. The accuracy score, we get on the validation set is 0.88 as same as logistic regression. However, the score on test data is increased to 0.88 (table 1 below) in comparison of logistic regression. It depicts that the BERT model performs better than logistic regression in sentiment classification.

3 TensorFlow Bert Model

To improve previous Bert model (based on the Transformer architecture) performance, we used TensorFlow BERT model further which is an implementation of BERT using the TensorFlow framework. TensorFlow's implementation of BERT has been optimized for performance, allowing it to process large amounts of data quickly and efficiently. It also provides a set of pre-trained models and tools to fine-tune BERT for specific NLP tasks. For this task, in pre-processing section, we performed transforming a review to the three embeddings, and formatting inputs that can be consumed by the model in training and testing. We set the maximum sequence length to 500. During the model training, addition to previous Bert model, we add one fully connected layer which has 768 ReLU activation units and dropout = 0.1. We also add an output layer which has two softmax functions. As a result, in just one epoch training, the nlp model has already achieved 93% accuracy on both validation and test datasets.

Metric (Accuracy)	Logistic Regression	Bert	TensorFlow Bert
Validation data	0.88	0.88	0.93
Test data	0.86	0.88	0.93

table 1: Model Evaluation Metric

Conclusion

Our sentiment analysis model, which utilized the BERT algorithm with TensorFlow framework, achieved an impressive accuracy of 0.93 on both the validation and test sets. The model's performance was further confirmed through the high precision, recall, and F1 score metrics in the classification report generated on the validation set, indicating its ability to effectively classify both positive and negative sentiments. Moreover, the sentiment results of the first five texts in the test dataset were visualized in the Appendix and confirmed the model's consistency with both BERT models.

Overall, this study demonstrates the efficacy of the BERT algorithm in sentiment analysis tasks and highlights the significance of data cleaning and visualization techniques in preparing text data for machine learning applications. With additional fine-tuning and optimization, this model can be utilized for sentiment analysis in different domains and datasets.

In summary, this project successfully tackled the sentiment analysis task of movie reviews and provided valuable insights into the structure of the dataset, while also achieving high accuracy and performance.

Appendix

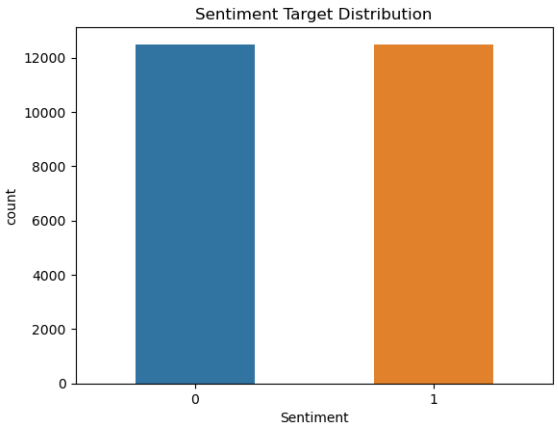

figure 1

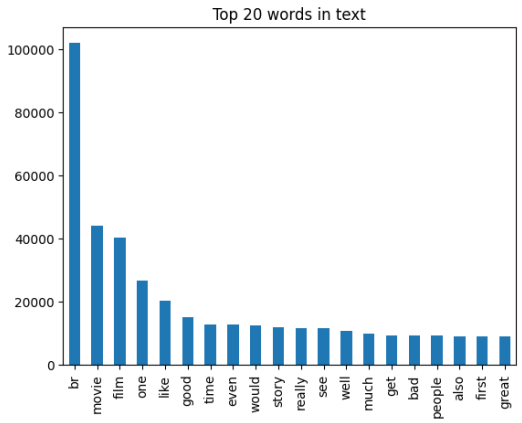

figure 2

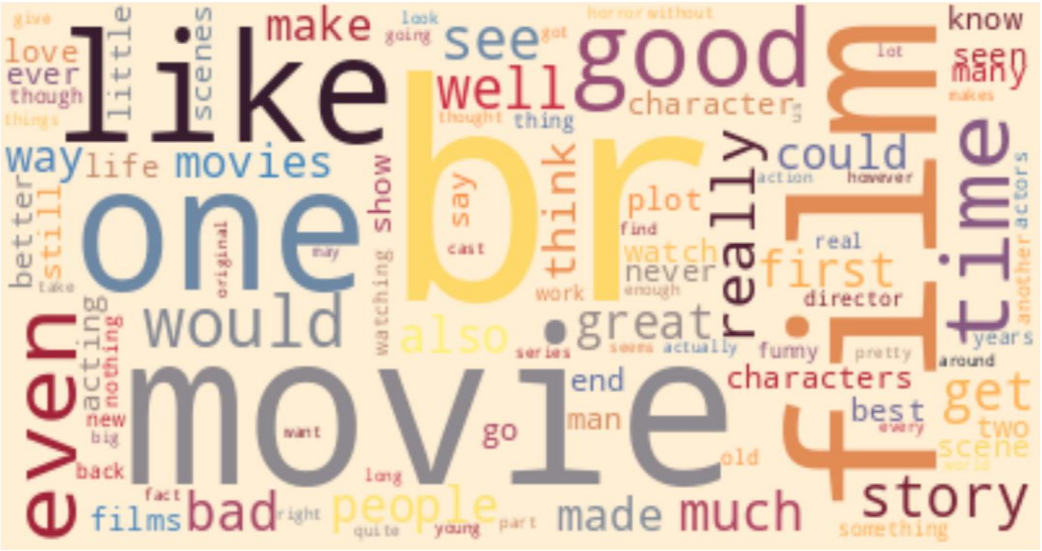

figure 3: Top 100 words in Text

Classification report

	precision	recall	f1-score	support
Negative	0.90	0.86	0.88	2466
Positive	0.87	0.90	0.89	2533
accuracy			0.88	4999
macro avg	0.89	0.88	0.88	4999
weighted avg	0.89	0.88	0.88	4999

figure 3: Logistic Regression Model

	precision	recall	f1-score	support
0	0.90	0.86	0.88	2541
1	0.86	0.90	0.88	2458
accuracy			0.88	4999
macro avg	0.88	0.88	0.88	4999
weighted avg	0.88	0.88	0.88	4999

figure 4: Bert Model

	precision	recall	f1-score	support
0	0.91	0.96	0.93	2521
1	0.96	0.90	0.93	2474
accuracy			0.93	4995
macro avg	0.93	0.93	0.93	4995
weighted avg	0.93	0.93	0.93	4995

figure 5: TensorFlow Bert Model

Model Results on Test Data

Logistic Regression:

	Id	Text	Sentiment
0	0	This is probably the best movie from director ...	1
1	1	It's particularly hard for a director to captu...	0
2	2	A very good movie about anti-semitism near the...	0
3	3	Interesting story and sympathetic treatment of...	1
4	4	There are films that are not released in theat...	0

Bert Model:

	Id	Text	Sentiment
0	0	This is probably the best movie from director Hector Babenco. It shows a Brazilian reality unknown by foreigners, which is the same reality that haunts all of the Latin American countries, poverty and a survival instinct. The most affected in this reality is the children usually left orphans, or abandoned by their poor parents have to make it in a "dog eat dog" society many times falling into the gap of delinquency, prostitution and crime. Very well acted and with a "no frills" approach, this movie will get to you, Great story plot, a must have movie on anybody's collection. The starring role went to Fernando Ramos da Silva, a young boy who fell into the crime wave, killed some years later during a robbery. I would suggest people to watch the movie "Who killed Pixote?" so you can have a more in depth idea of the lives of these characters. Some other Characters from the movie had a similar fate, some died and others are in jail. None the less this movie will last for a long time in ...	1
1	1	It's particularly hard for a director to capture film-making without getting precious, inbred, over-dramatic, or all three. Breillat ably demonstrates the instinctive, lizard-brain methods of a female auteur in extracting from two "cattle" (as Hitchcock called actors) a love-scene of searing intimacy. Her main battle is with her leading man ("an actor is really a woman" she opines), although, naturally, it is the leading lady who will steal the show. I disagree that this is Breillat's first comedy. 'Romance' was at various points hilarious, but I accept that the French sense of humour can be elusive for foreigners; indeed, dozens of IMDb reviewers detected no comedy in Romance. By contrast, Sex Is Comedy raises plenty of laughs, mainly by using an actor's prop that goes back thousands of years to Plautus and the ancient Greeks. We wondered, leaving the theatre, whether Roxane's "beard" was a wig. A lovely performance from Anne Parillaud as Breillat wrestling with her own script, lo...	1
2	2	A very good movie about anti-semitism near the end of WWII. The scene that really speaks loudly of the ignorance of these people is the meeting at the church when the priest is giving his speech against the "international money grubbers and communists". It sounds amazingly like the speeches that Adolph Hitler used to force down his peoples' throats, yet none of the meeting attendees seem to make this comparison.	1
3	3	Interesting story and sympathetic treatment of racial discrimination, Son of the Gods is rather too long and contains some hammy acting, but on the whole remains a fascinating film. Story about a Chinese passing as White (Rchard Barthelmess) starts as Barthelmess leaves college after being insulted by a trio of brainless co-eds. He embarks on a world tour to discover himself and ends up as secretary to a British playwright (Claude King). In Monte Carlo he meets beautiful Alanna Wagner (Constance Bennett) and they fall in love. But when she discovers he is Chinese she goes berserk in a memorable scene. Plagued by guilt and love, Alanna goes into a mental spiral and makes a few attempts to contact Barthelmess. After his father dies he takes over the business (banking?) and dons Chinese garb as a symbol of his hatred of the White race that has spurned him. After a San Francisco detective tells him the truth about his birth, Barthelmess makes the decision to honor...	1
4	4	There are films that are not released in theaters but on video. This one should be allowed to age and disintegrate the way old nitrate film stock does. No story, inept violence, over acted, badly written and the sorry thing is that the star was not the only bad part in the film. And I did like and enjoyed some of Siegel's other movies.	0

TensorFlow Bert Model:

	Id	Text	Sentiment
0	0	This is probably the best movie from director Hector Babenco. It shows a Brazilian reality unknown by foreigners, which is the same reality that haunts all of the Latin American countries, poverty and a survival instinct. The most affected in this reality is the children usually left orphans, or abandoned by their poor parents have to make it in a "dog eat dog" society many times falling into the gap of delinquency, prostitution and crime. Very well acted and with a "no frills" approach, this movie will get to you, Great story plot, a must have movie on anybody's collection. The starring role went to Fernando Ramos da Silva, a young boy who fell into the crime wave, killed some years later during a robbery. I would suggest people to watch the movie "Who killed Pixote?" so you can have a more in depth idea of the lives of these characters. Some other Characters from the movie had a similar fate, some died and others are in jail. None the less this movie will last for a long time in ...	1
1	1	It's particularly hard for a director to capture film-making without getting precious, inbred, over-dramatic, or all three. Breillat ably demonstrates the instinctive, lizard-brain methods of a female auteur in extracting from two "cattle" (as Hitchcock called actors) a love-scene of searing intimacy. Her main battle is with her leading man ("an actor is really a woman" she opines), although, naturally, it is the leading lady who will steal the show. I disagree that this is Breillat's first comedy. 'Romance' was at various points hilarious, but I accept that the French sense of humour can be elusive for foreigners; indeed, dozens of IMDb reviewers detected no comedy in Romance. By contrast, Sex Is Comedy raises plenty of laughs, mainly by using an actor's prop that goes back thousands of years to Plautus and the ancient Greeks. We wondered, leaving the theatre, whether Roxane's "beard" was a wig. A lovely performance from Anne Parillaud as Breillat wrestling with her own script, lo...	1
2	2	A very good movie about anti-semitism near the end of WWII. The scene that really speaks loudly of the ignorance of these people is the meeting at the church when the priest is giving his speech against the "international money grubbers and communists". It sounds amazingly like the speeches that Adolph Hitler used to force down his peoples' throats, yet none of the meeting attendees seem to make this comparison.	1
3	3	Interesting story and sympathetic treatment of racial discrimination, Son of the Gods is rather too long and contains some hammy acting, but on the whole remains a fascinating film. Story about a Chinese passing as White (Rchard Barthelmess) starts as Barthelmess leaves college after being insulted by a trio of brainless co-eds. He embarks on a world tour to discover himself and ends up as secretary to a British playwright (Claude King). In Monte Carlo he meets beautiful Alanna Wagner (Constance Bennett) and they fall in love. But when she discovers he is Chinese she goes berserk in a memorable scene. Plagued by guilt and love, Alanna goes into a mental spiral and makes a few attempts to contact Barthelmess. After his father dies he takes over the business (banking?) and dons Chinese garb as a symbol of his hatred of the White race that has spurned him. After a San Francisco detective tells him the truth about his birth, Barthelmess makes the decision to honor...	1
4	4	There are films that are not released in theaters but on video. This one should be allowed to age and disintegrate the way old nitrate film stock does. No story, inept violence, over acted, badly written and the sorry thing is that the star was not the only bad part in the film. And I did like and enjoyed some of Siegel's other movies.	0