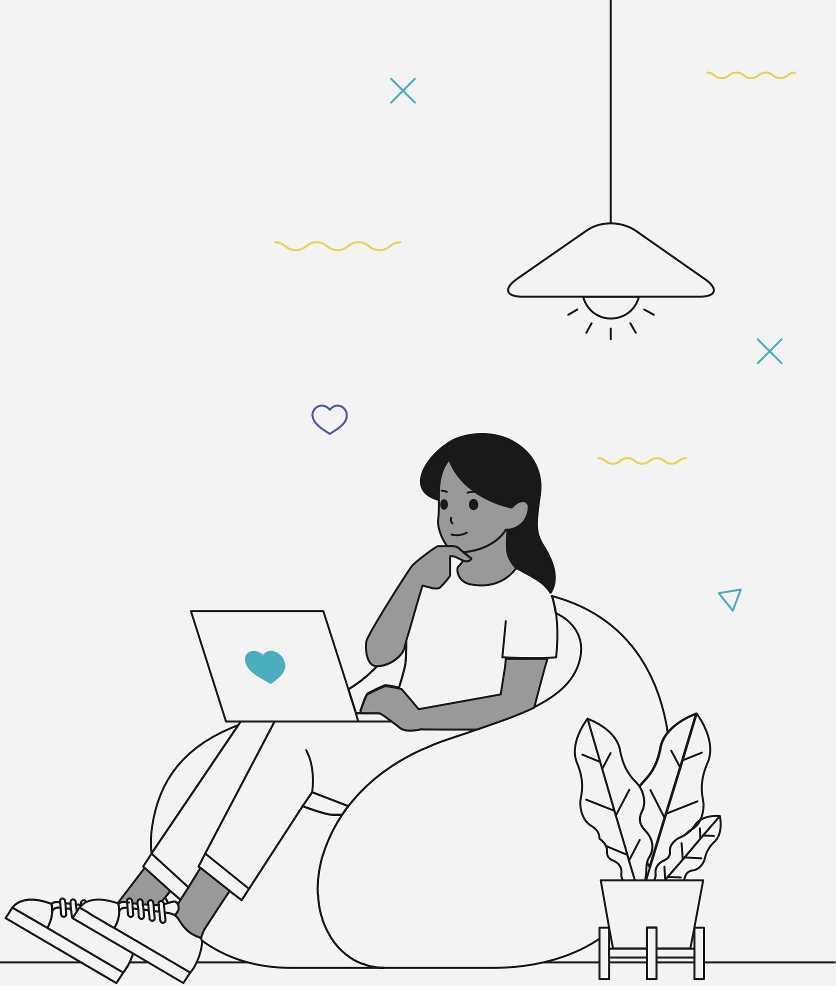
Sentiment Analysis

Andric and Maha



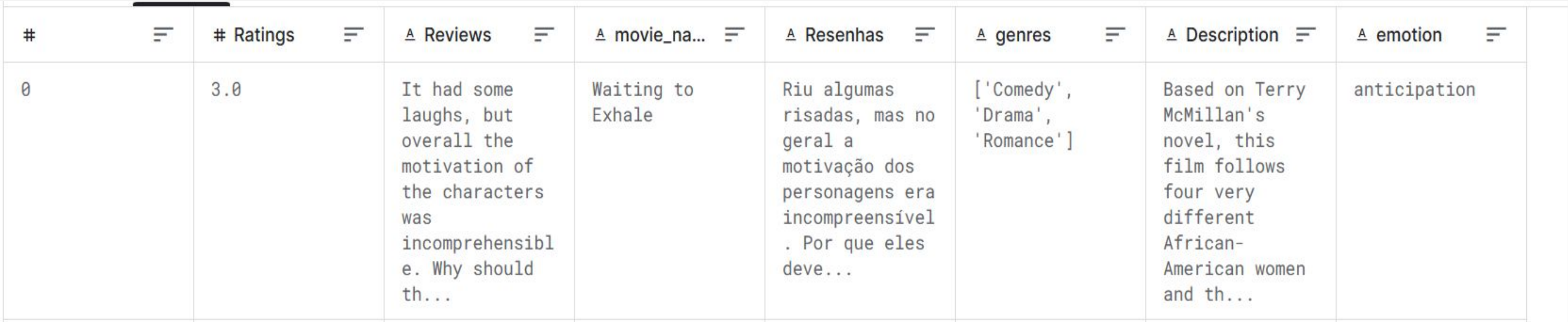
Movie Reviews

Goals

- Produce two models:
  - Predicting a movie’s determined emotion
  - Predicting a movie’s binary sentiment based on its ratings
- Evaluate whether a review expresses the movie’s sentiment and its emotion.
- Identify the correlation between a movie’s rating and its determined emotion.

Dataset

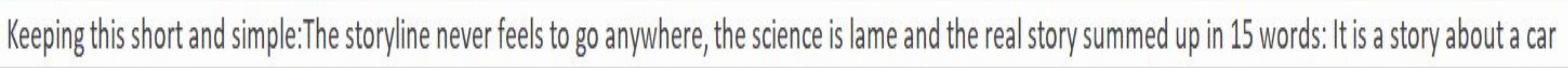
- An expansion of the 50k reviews from IMDB
- Main Focus:
  - Ratings
  - Reviews
  - Emotions



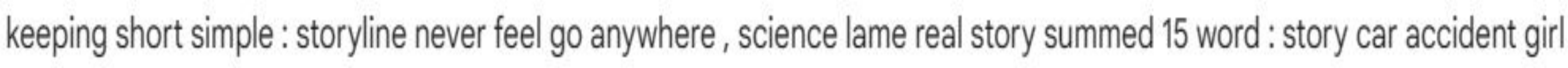
#	# Ratings	A Reviews	A movie_na...	A Resenhas	A genres	A Description	A emotion
0	3.0	It had some laughs, but overall the motivation of the characters was incomprehensible. Why should th...	Waiting to Exhale	Riu algumas risadas, mas no geral a motivação dos personagens era incompreensível. Por que eles deve...	['Comedy', 'Drama', 'Romance']	Based on Terry McMillan's novel, this film follows four very different African-American women and th...	anticipation

Preprocessing the Data

1. Dropped columns “Resenhas”, “genres”, “description”
2. Removed duplicated reviews
3. Converted ratings column to binary
    a. Pos=1 label for every rating > 5
    b. Neg=–1 label for every rating <= 5
4. Filtered each review
    a. Tokenization, Lemmatization, and Stop Word Removal
5. Vectorized each text using Bag of Words
6. Shuffled and split the data

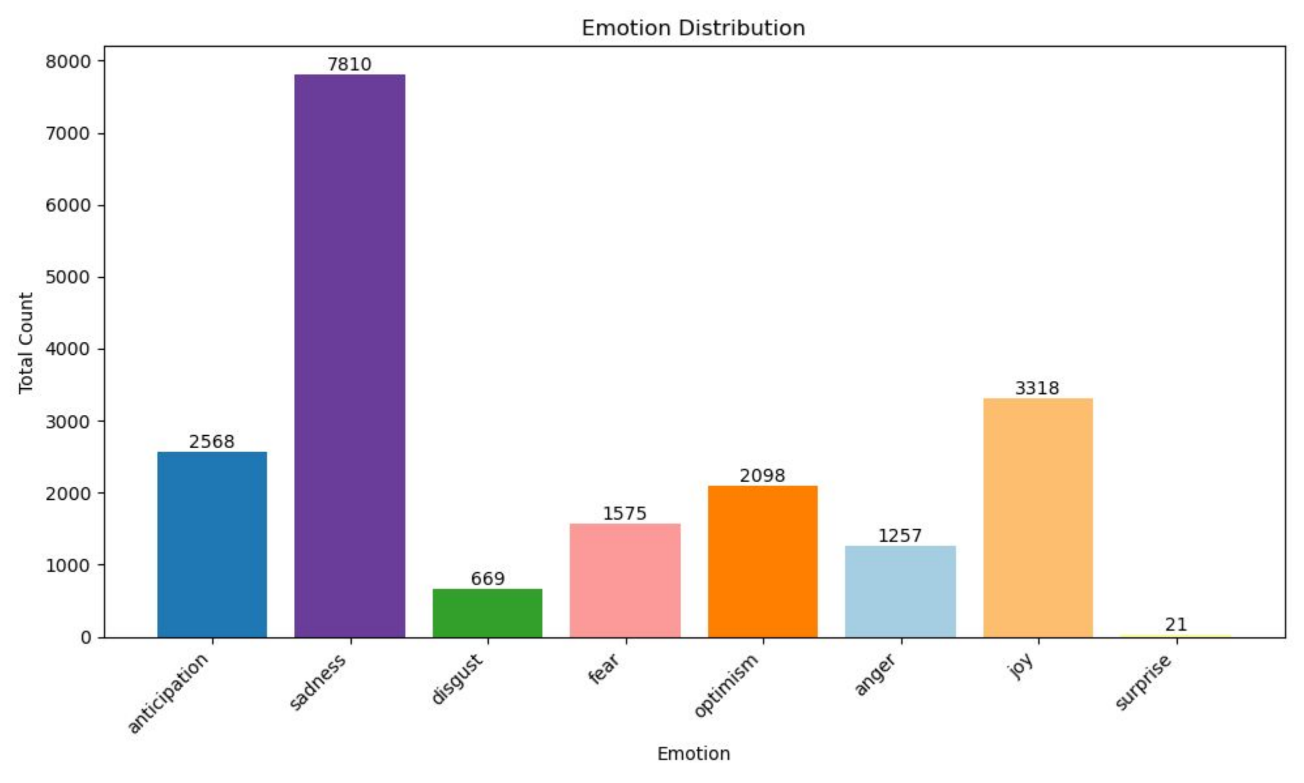


Keeping this short and simple:The storyline never feels to go anywhere, the science is lame and the real story summed up in 15 words: It is a story about a car

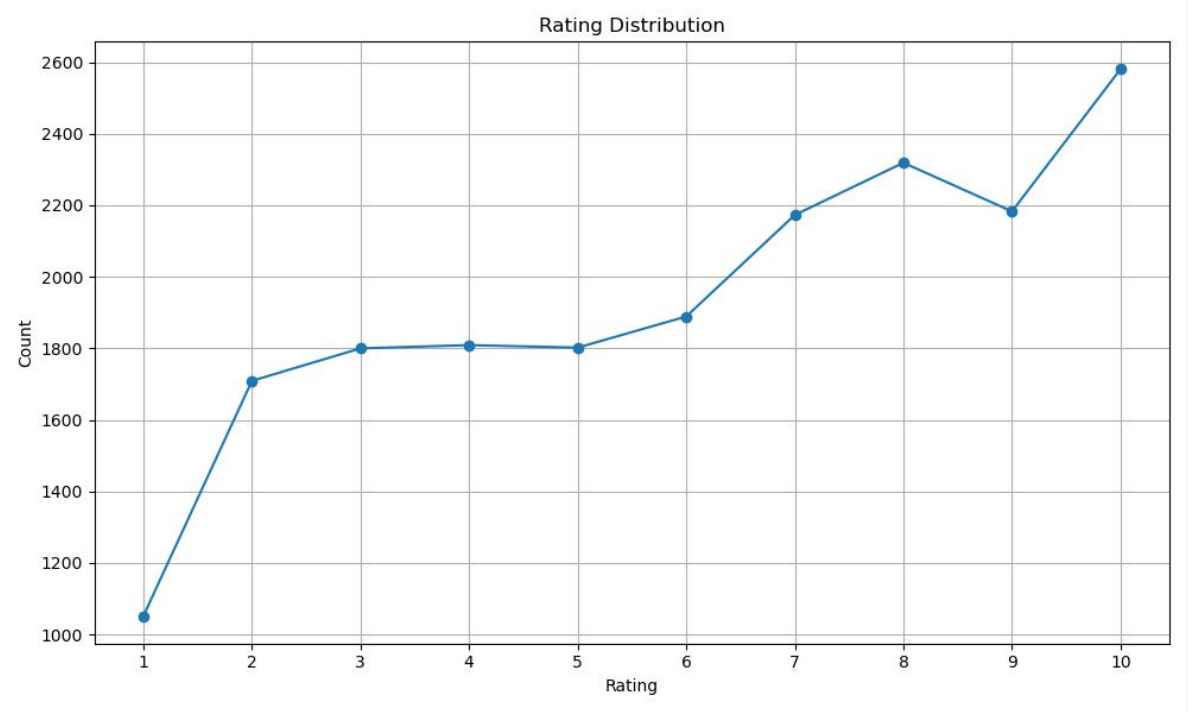


keeping short simple : storyline never feel go anywhere , science lame real story summed 15 word : story car accident girl

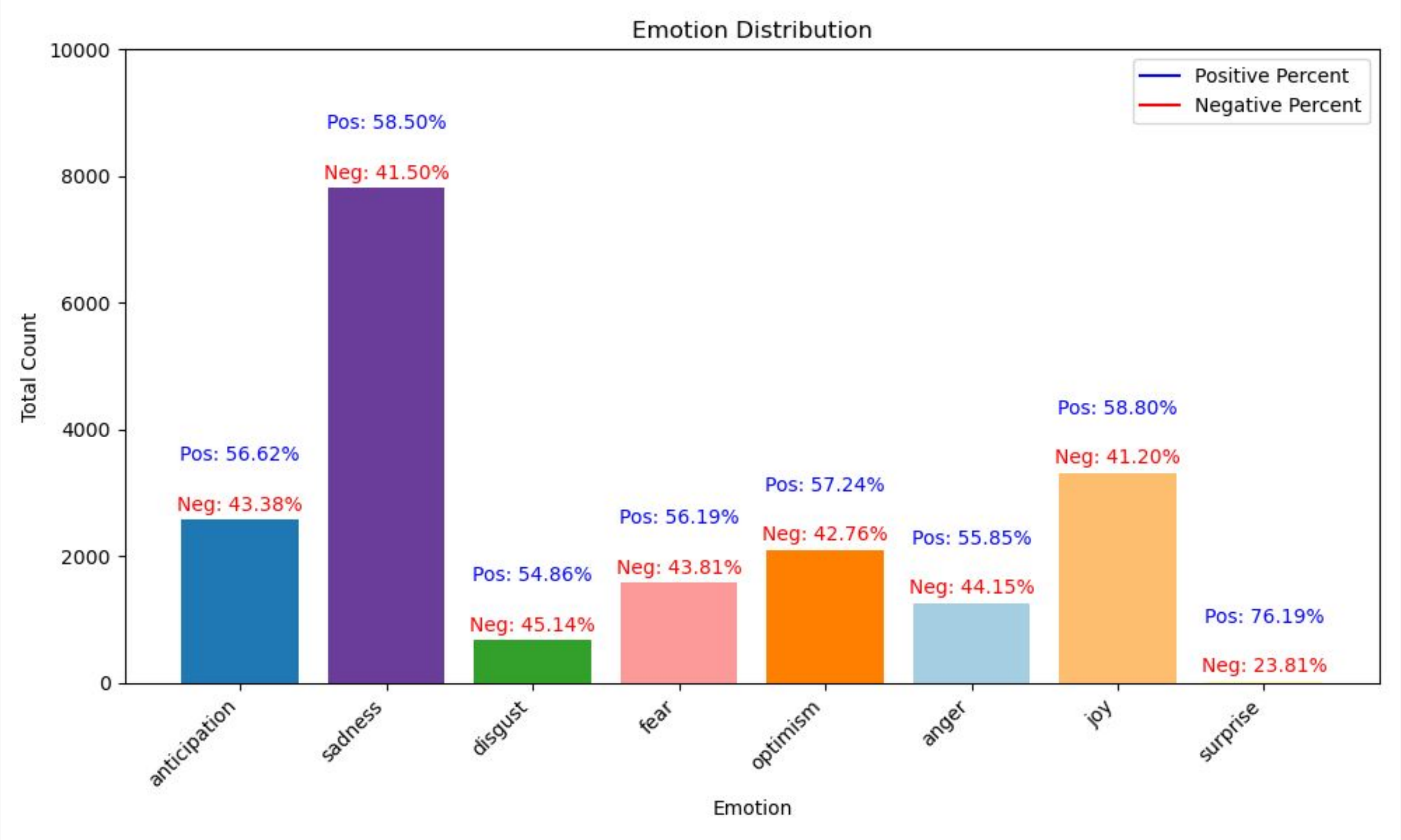
Preliminary analysis



Preliminary analysis



Preliminary analysis

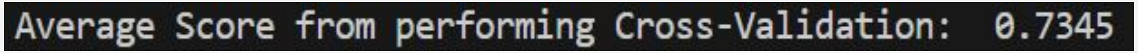


SVM Model for binary classification

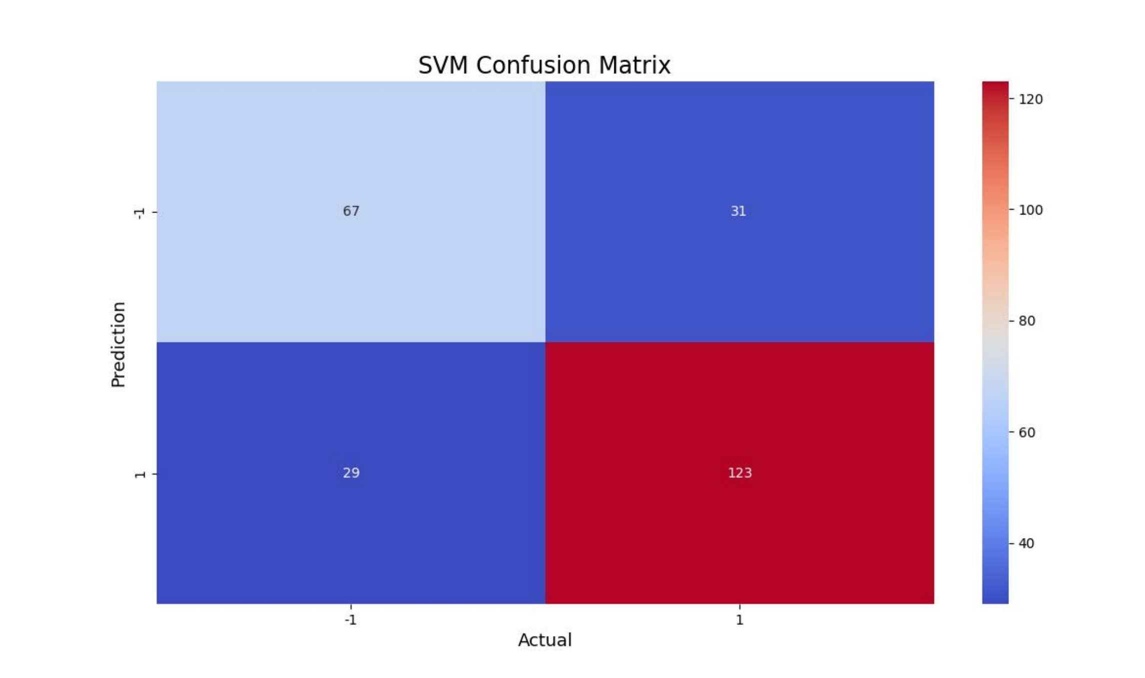
- **Effective in a high-dimensional feature space:**
  Each word can be treated as a feature for classifying the sentiment (reflected in the ratings).
- **Resistant to overfitting:**
  Regularization controls the trade-off between maximising the margin between classes and minimising classification error
- **Does not need data to be linearly separable:**
  Kernel functions

Classification for a dataset of 2500 examples

Accuracy = 76% on the testing dataset (250 examples)

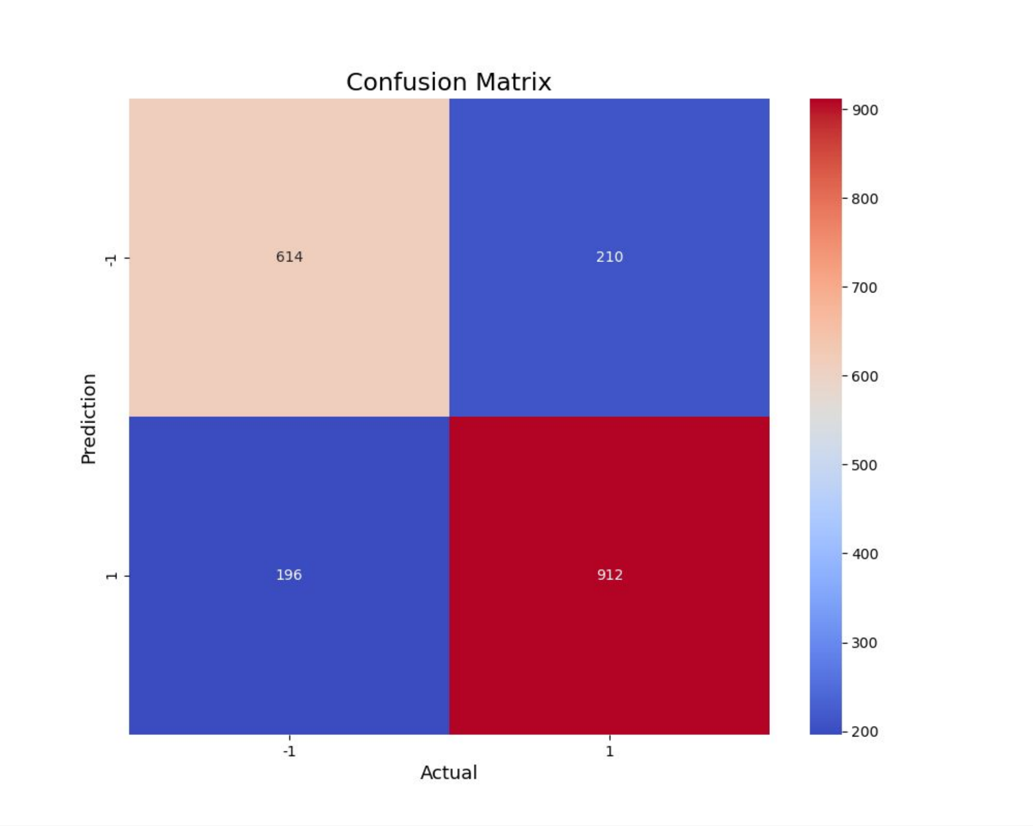


```
Average Score from performing Cross-Validation:  0.7345
```



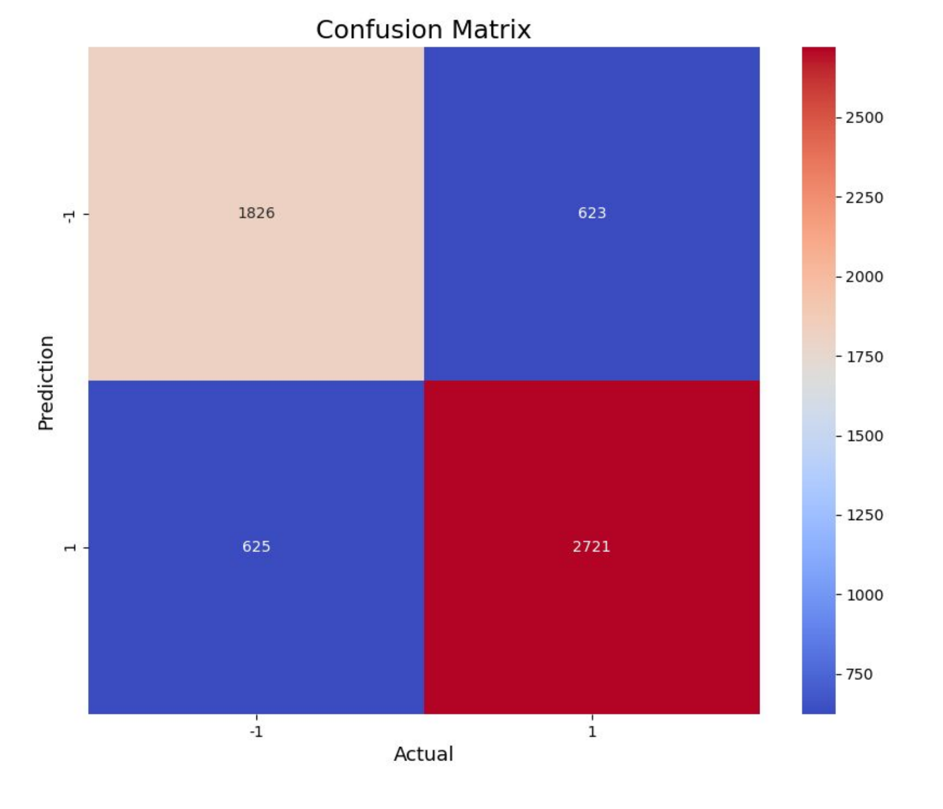
Classification for a dataset of 19317 examples

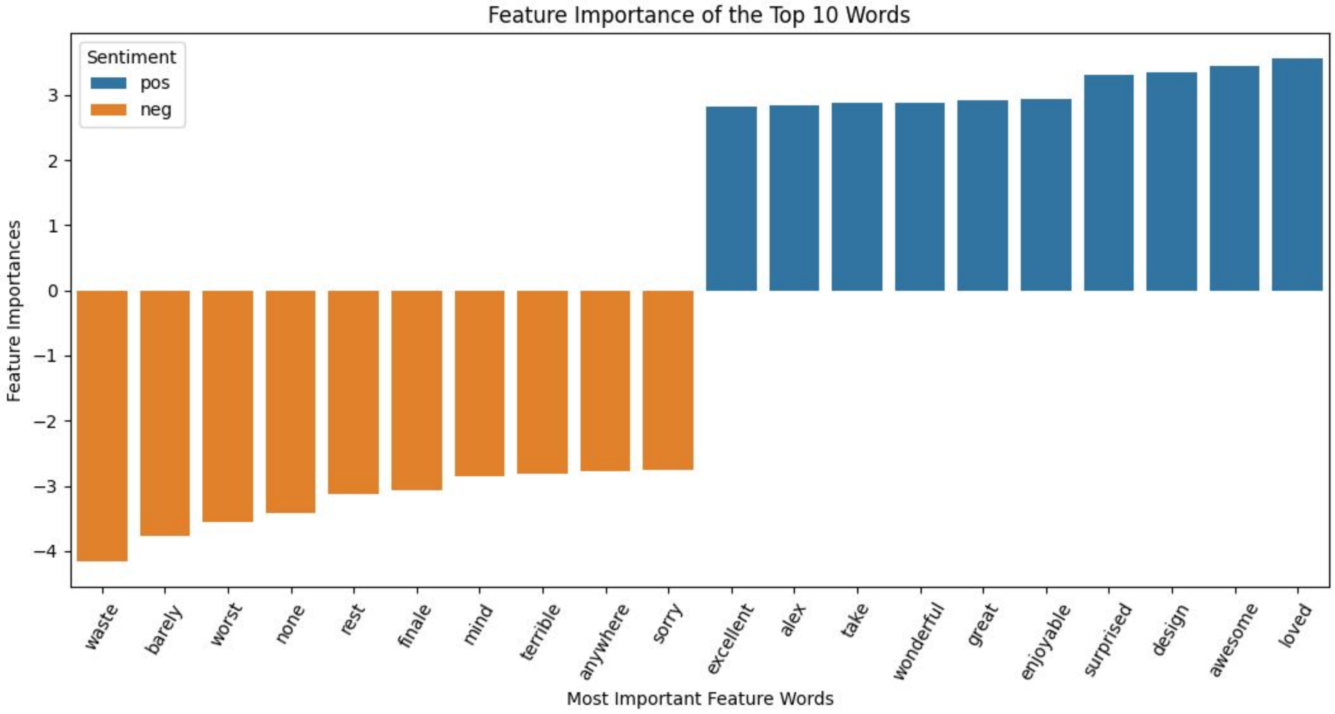
Accuracy = 79% on the testing dataset (1932 examples)



Classification for a dataset of 19317 examples

Accuracy = 78% on the testing dataset (5795 examples)




Feature Importance of the Top 10 Words
Sentiment
pos
neg
Feature Importances
3
2
1
0
−1
−2
−3
−4
waste
barely
worst
none
rest
finale
mind
terrible
anywhere
sorry
excellent
alex
take
wonderful
great
enjoyable
surprised
design
awesome
loved
Most Important Feature Words

Looking ahead

Emotion Prediction

- RNN
- Final dense layer with softmax

One-model, two predictions

- Model that predicts rating and emotions
- Generate learning curves for the model

Generate Reviews

Transformers!!!

Resources:

https://www.kaggle.com/datasets/fahadrehman07/movie-reviews-and-emotion-dataset
https://ai.stanford.edu/~amaas/papers/wvSent_acl2011.pdf
https://www.intofilm.org/resources/1642
https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107
https://realpython.com/python-keras-text-classification/