IMDB Movie Reviews

Presented by:
Abdullah & Nada

Table Of Contents

01

Introduction

IMDB movies

The IMDb Movie Reviews dataset is a binary sentiment analysis dataset consisting of 253,039 reviews from the Internet Movie Database (IMDb) labeled as positive or negative.

SENTIMENT ANALYSIS

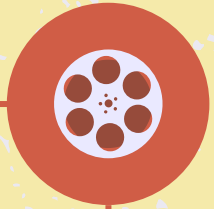
Sentiment analysis is an NLP task that aims to obtain the writer's feelings expressed in positive or negative reviews.

TOOLS

Workflow

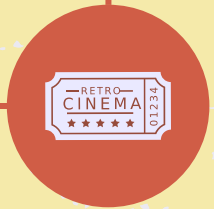
IDEA

EDA



Building Models

Data Cleaning & NLP



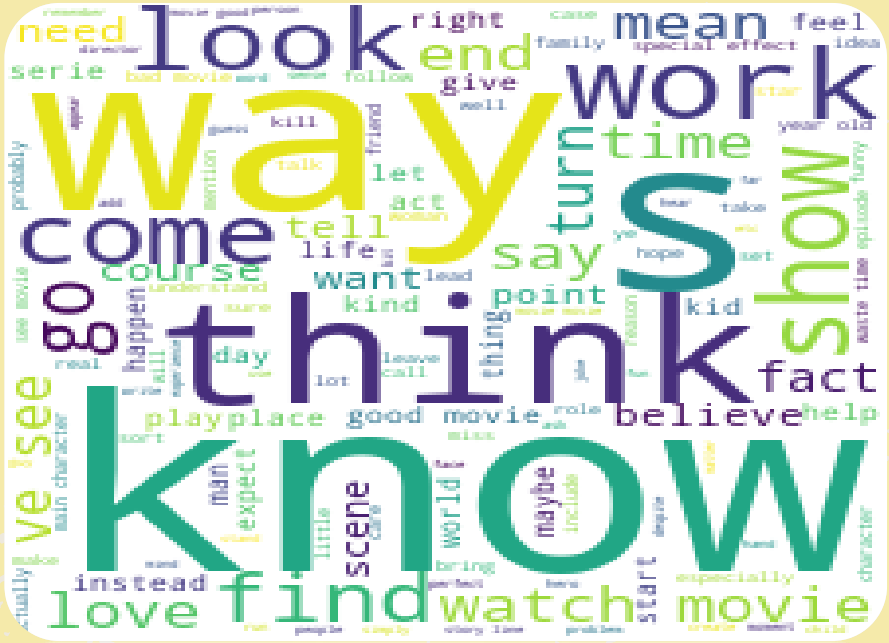
Pre-processing (spaCy)

- Tokenization
- Stop word
- Lemmatization

Data Cleaning

- Replace apostrophe with blank
- Removing Punctuation
- Take only alphanumeric
- Drop duplicates

Data Visualization

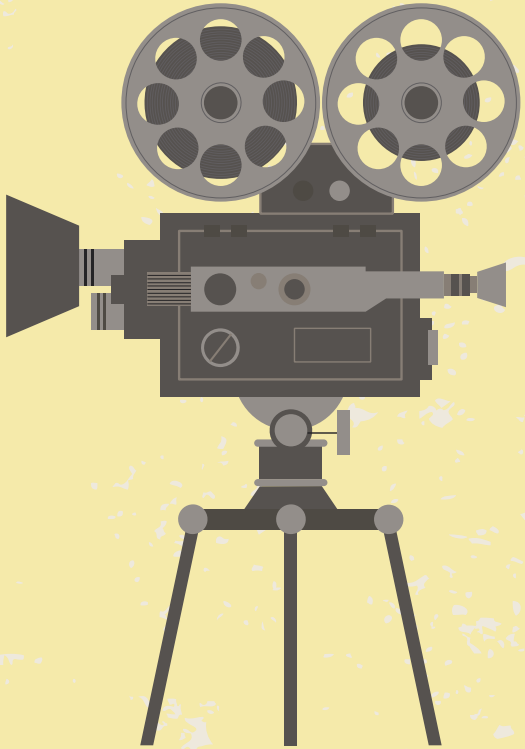


Topic Modelling

NMF

CorEx

LSA



- The best model is CorEx with 7 topics

NMF

Topic 0 : life, way, man, year, people, world, end, real, old, leave

Topic 1: watch, love, people, enjoy, episode, show, feel, season, funny, start

Topic 2: character, plot, series, feel, episode, development, way, season, show, care

LSA

Topic 0: time, character, watch, story, great, way, scene, bad, people, end

Topic 1: play, performance, life, man, role, world, director, young, set, lead

Topic 2: bad, people, plot, kill, minute, end, waste, money, point, happen

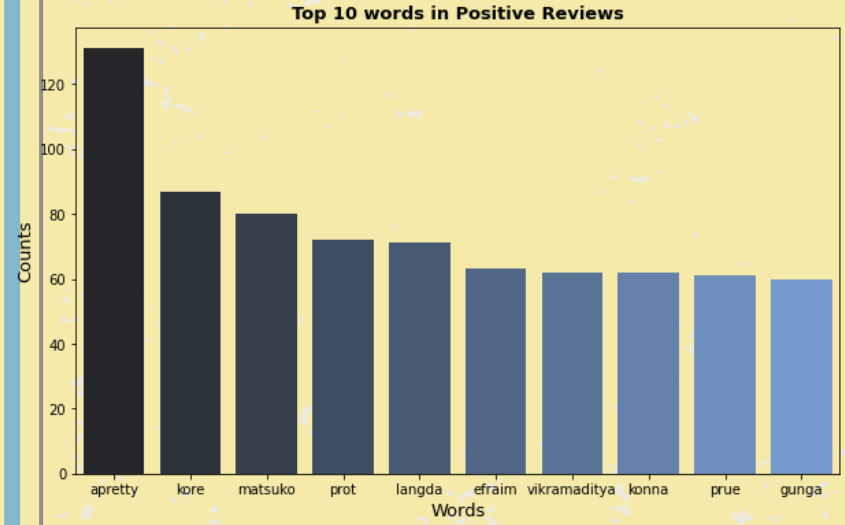
CorEx

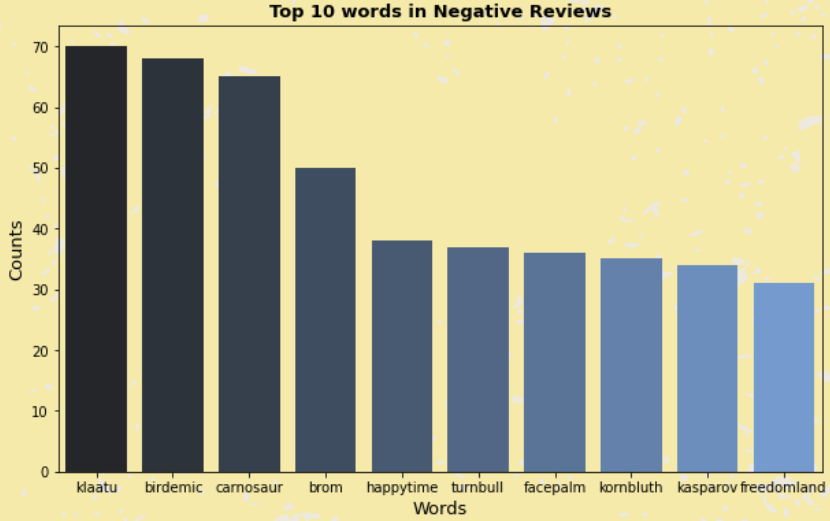
Topic 0: actor, play, director, act, set, audience, lead, screen, moment, human

Topic 1: character, scene, time, action, plot, feel, point, effect, long, sequence

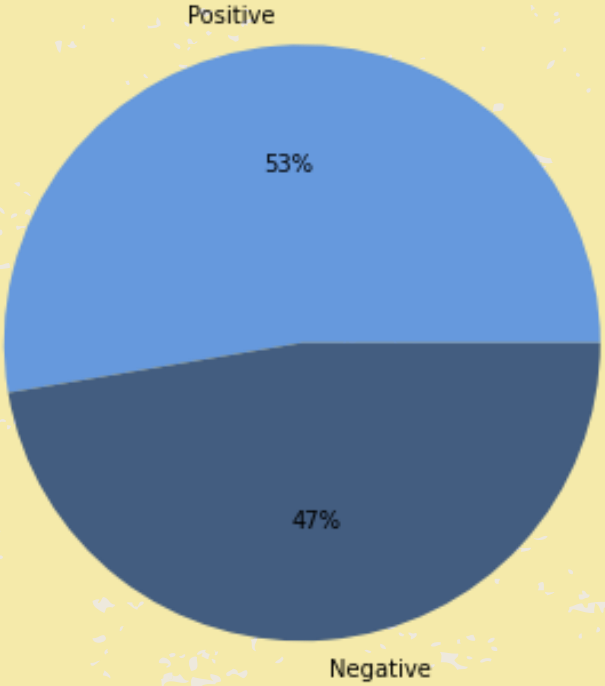
Topic 2: war, fight, star, hero, course, power, battle, return, lose, evil

Top 10 most common words in positive and negative reviews





The balance of target



Models

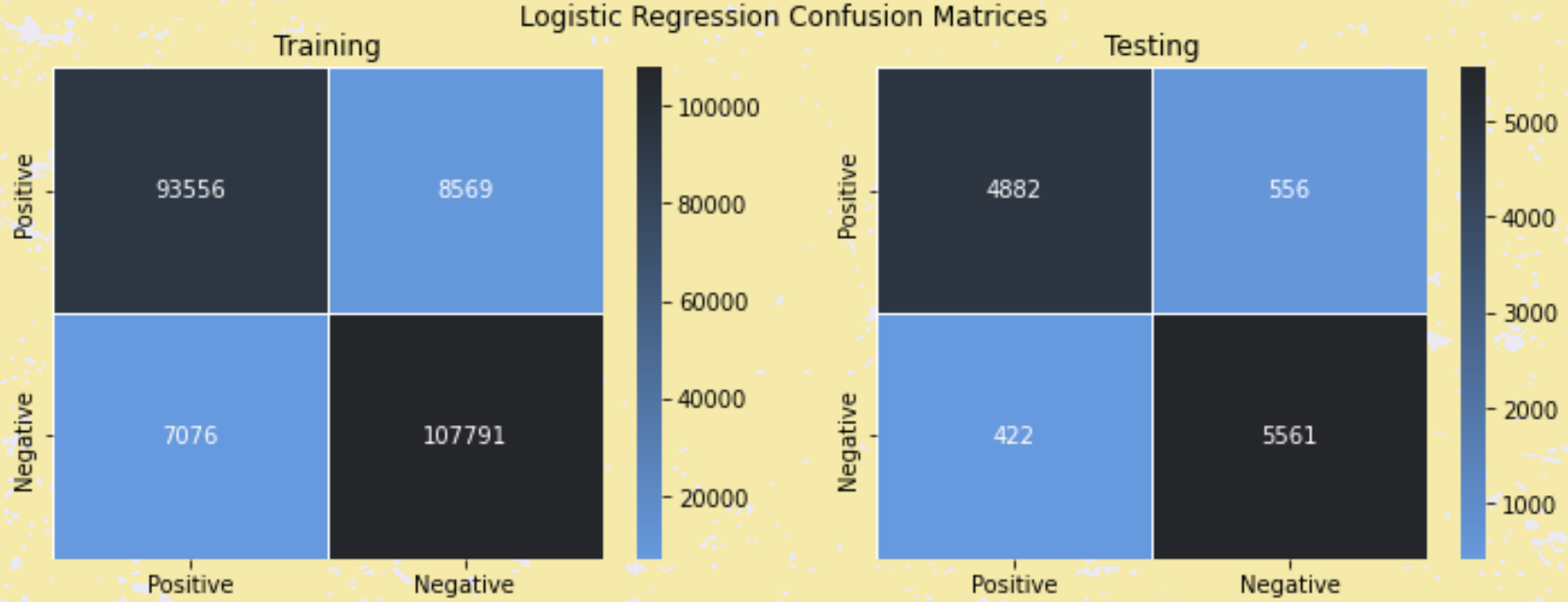
Model	Train	Validation
	Accuracy	Accuracy
K-Nearest Neighbor	82%	72%
Logistic Regression	93%	92%
Decision Trees	100%	76%
Random Forest	100%	88%
Extra Trees Classifier	100%	89%
Bernoulli Naive Bayes	86%	86%
Gaussian Naive Bayes	90%	88%

Final result

Logistic Regression

Training and validation :93%
Test : 91%

Confusion Matrix



Conclusion

After using NLP methods and fitting a classification model were able to predict review sentiment with high accuracy.

Thanks