Sentiment Analysis for Movie Reviews

Harsh Dubey, Vishrut Maheshwari, Chetan Rajput, Rahul Salunke

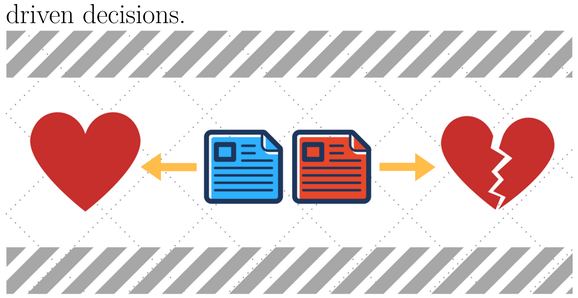
Indian Institute of Technology, Goa

Objectives

- A sentiment analysis system for text analysis combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase.
- Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics
- In this project, our model predicts whether given sentence is either positive or negative using Sentiment Analysis.

Introduction

Sentiment analysis is the automated process that uses AI and ML to identify the sentiment expressed by the text. Sentiment analysis is widely used for getting insights from social media comments, survey responses and product reviews for making data-driven decisions.



- Input: The input to this program is a directory containing training and test sub-directories from which the program extracts reviews and puts them in separate pandas DataFrames.
- Output: The output of this program are the sentiments associated with each review in the test data set. It also outputs the accuracy of the model which is calculated against the actual sentiments associated with the reviews.

The approach

- Data Extraction and Cleaning: Data is present in separate text files which is extracted using the os library. It is cleaned by taking every fifth data value because of the large size of data.
- Data Preprocessing: Stopwords are removed from the data. Vectorization is performed which produces numerical features for the classifier. TF-IDF is used which is a simple vectorization technique that consists of computing word frequencies and downscaling them for common words. Finally, For additional context, we provide the model with N-grams.
- Building the Model: The algorithm used is LinearSVC from scikit-learn. The accuracy of the model is calculated on the test set using accuracy_score from scikit-learn metrics.

Results and Conclusions

- The data set contains 12500 negative and 12500 positive training examples.
- The most common words present in the data set are summarized in the chart below.
- The accuracy obtained after training on the training set against the test set is around 88%.

Graphical work flow

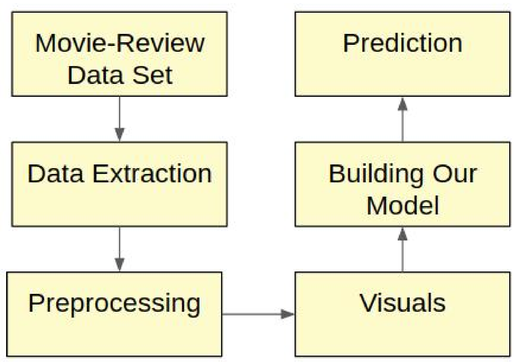


Figure 1:A brief idea about the flow of the program

Inference

The movies were rated from 1-4 and 7-10. This can be inferred from the adjacent figure. The accuracy obtained on the test data set is around 88%.
The age of getting meaningful insights from social media data has now arrived with the advance in technology. Companies have been leveraging the power of data lately, but to get the deepest of the information, one must leverage the power of AI, Deep learning and intelligent classifiers like SVM and Perceptron.
The model can be improved by using advanced NLP algorithms like Convolutional Neural Networks(CNNs), Recurrent Neural Networks(RNNs), Recursive Neural Networks(RNNs), Word Embeddings, Reinforcement Learning or Memory-Augmented Networks.

Experiments and Analysis

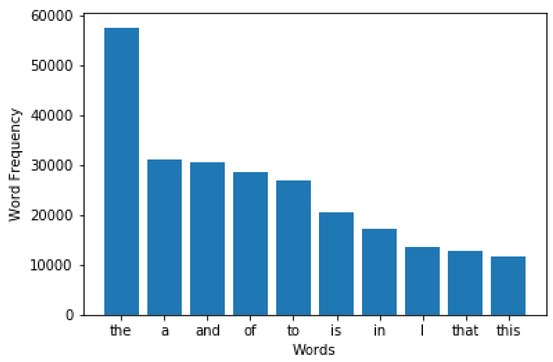
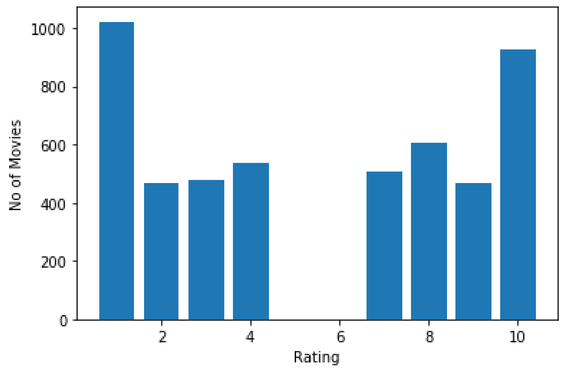


Figure 2:The 10 most common words used in the data set



A bar chart showing the ratings out of 10 vs the number of movies rated

Code Snippet

```
reviews = [
"good movie, must watch!",
"I'd rather eat popcorn than watch this movie.",
"Horrible direction and substandard acting.",
"A classic movie."
]

transformed = vectorizer.transform(reviews)
predictions = model.predict(transformed)

print(predictions)
```

```
['positive', 'negative', 'negative', 'positive']
```

Figure 3:Some exemplar reviews for testing and their outputs

References

- https://medium.com/data-from-the-trenches/text-classification-the-first-step-toward-nlp-mastery-f5f95d525d73
- https://monkeylearn.com/sentiment-analysis/
- https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Acknowledgements

A term project completed under the requirements of course CS 386: Artificial Intelligence Thanks to Dr Clint P. George for assisting and encouraging us to complete this project.