Tom Booth

TBooth3@UCLAN.ac.UK Tom.Booth@BAESystems.com

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

Artificial Intelligence Report

IMDB Movie Dataset – Sentiment Analysis: Review Categorisation

# Table of Contents

[1. Table of Contents 1](#_Toc125910079)

[2. Introduction: 1](#_Toc125910080)

[3. Background Reading: 1](#_Toc125910081)

[4. Data: 2](#_Toc125910082)

[5. Model Development: 2](#_Toc125910083)

[6. Model Evaluation: 2](#_Toc125910084)

[7. Conclusion: 2](#_Toc125910085)

[8. Supplementary Material: 2](#_Toc125910086)

# Introduction:

This report aims to explain the process the author has taken to analyse the IMDB Movie Dataset, **Link to Dataset**, and the approach they have taken to develop and utilise a sentiment analysis to categorise movie reviews as either positive or negative, utilising the dataset split between a training and testing divide, allowing for a high accuracy of prediction. 90% Accuracy. Aiming to provide a clear indication as to the review’s direction.

**Add more.**

# Background Reading:

This report has utilised the available recourses online to investigate sentiment analysis and its uses within model development and creation. When determining which analysis types and algorithms to utilise for the model’s development, the author utilised published articles and material to create an estimation of approaches and algorithms to test.

(Medhat et al., 2014) discusses how the Naïve Bayes Classifier is one of the simplest and most utilised classifiers, “computes the posterior probability of a class based on the distribution of words in the document” This accurately describes the approach to this dataset, there is several reviews to be analysed that contain various keywords that can be used to determine and signify either a positive or negative review. This approach can be used to identify the number of either portion of words. Attaching a determination to the review.

(Rish, 2001) mentions that the Naïve Bayes classifier is ‘surprisingly effective’, as its classification decisions may often be correct even if its probability estimates are ‘inaccurate’, this report therefore suggests the introduction of a second method of analysis, introducing a different approach potentially yielding better and more accurate results.

(Medhat et al) discusses the various methods of analysis and algorithms and approaches, mentioning that Linear Classifiers, specifically Support Vector Machine classifiers are good ways to determine ‘linear separators’ in the search space. For the dataset this report is analysis, there is a clear difference between positive and negative reviews. They go onto discuss how “text data is ideally suited for SVM” (Support Vector Machines).

# Data:

* Describe the data you are working with for your project. What type of data is it? Where did

it come from? How much data are you working with? Did you have to do any preprocessing, filtering, or other special treatment to use this data in your project? If you are collecting new data, how will you do it and incorporate into your model?

The dataset utilised within this model was retrieved from Kaggle, a subsidiary of Google, an online community of Data Scientists allowing for the sharing of datasets and learning. The dataset was retrieved from Kaggle directly, from the following link: <https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews>

User Lakshmipathi N has uploaded the dataset for utilisation within the data science and artificial intelligence field. The dataset is free to use and available to utilise without any additional licence.

# Model Development:

# Model Evaluation:

# Conclusion:

# Supplementary Material: