Report

October 6, 2019

1 Machine Learning Engineer Nanodegree

1.1 Sentiment Analysis on IMDB Movie Review Dataset

Piyush Moolchandani
October 04, 2019

1.2 Definition

1.2.1 Project Overview

This project deals with the problem of sentiment analysis. IMDB's movie review dataset has been used to train the model in this project.

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. Sentiment analysis basically provides service provider a idea of polarity of user's views on the product by doing some processing on the text review/feedback given by the user. In today's competitive market, it is very important to analyse and act upon the reviews but with the ever growing data, it is impossible to get reviewed each and every review by an human, hence sentiment analysis algorithm becomes an crucial advancement in this field.

1.2.2 Problem Statement

The objective of this project is to create a model to predict whether a given review is of positive and negative view, and for training and testing this model, IMDB dataset from this ACL 2011 paper by Learning Word Vectors for Sentiment Analysis (http://www.aclweb.org/anthology/P11-1015) is to be used.

A LSTM based RNN Network is used to solve this problem. Initially data is preprocessed to make it suitable to be used with the lstm network and then a lstm network with embedding layer is defined. This network is then trained using training dataset of IMDB dataset and after suitable training, it is tested on other half (testing dataset) of the data and we have a model with trained parameters for sentiment analysis.

1.2.3 Metrics

Accuracy metric is used in this project for performance evaluation.

it is a common metric for binary classifiers. It takes into account both true positives and true negatives with equal weight.

accuracy = (True Positives + True Negatives) / Size of Dataset

Accuracy is a correct metric to evaluate the performance on this project because the most important thing to understand while evaluating the performance of a sentiment analysis algorithm is to know what fraction of reviews are correctly classified. Accuracy metric does exactly this as is visible in formula mentioned above.

1.3 Analysis

1.3.1 Data Exploration

The core dataset contains 50,000 reviews split evenly into 25k train and 25k test sets. The overall distribution of labels is balanced (25k pos and 25k neg). It also include an additional 50,000 unlabeled documents for unsupervised learning.

In the entire collection, no more than 30 reviews are allowed for any given movie because reviews for the same movie tend to have correlated ratings. Further, the train and test sets contain a disjoint set of movies, so no significant performance is obtained by memorizing movie-unique terms and their associated with observed labels. In the labeled train/test sets, a negative review has a score <= 4 out of 10, and a positive review has a score >= 7 out of 10. Thus reviews with more neutral ratings are not included in the train/test sets. In the unsupervised set, reviews of any rating are included and there are an even number of reviews > 5 and <= 5.

Only the train and test sets are used in this project.

Examples from dataset > 1. **From pos directory :** Bromwell High is nothing short of brilliant. Expertly scripted and perfectly delivered, this searing parody of a students and teachers at a South London Public School leaves you literally rolling with laughter. It's vulgar, provocative, witty and sharp. The characters are a superbly caricatured cross section of British society (or to be more accurate, of any society). Following the escapades of Keisha, Latrina and Natella, our three "protagonists" for want of a better term, the show doesn't shy away from parodying every imaginable subject. Political correctness flies out the window in every episode. If you enjoy shows that aren't afraid to poke fun of every taboo subject imaginable, then Bromwell High will not disappoint!

2. **From neg directory :** From the beginning of the movie, it gives the feeling the director is trying to portray something, what I mean to say that instead of the story dictating the style in which the movie should be made, he has gone in the opposite way, he had a type of move that he wanted to make, and wrote a story to suite it. And he has failed in it very badly. I guess he was trying to make a stylish movie. Any way I think this movie is a total waste of time and effort. In the credit of the director, he knows the media that he is working with, what I am trying to say is I have seen worst movies than this. Here at least the director knows to maintain the continuity in the movie. And the actors also have given a decent performance.

This dataset needs to be preprocessed before giving input to the network. Distribution of datasets: 1. Test data 2. Train data: - 80 % : actual train data - 20 % : validation data