Machine Learning Engineer Nanodegree

Capstone Proposal

Sujay Bhowmick

August 8th 2018

Proposal

Domain Background

In NLP sentiment analysis is the most common problems through which we try to infer the sentiment of sentence or a paragraph.

My motivation for this project has come from the fact that I have been working in a company which is working on Sentiment Analysis of News, Blogs and Tweets. My doing this project I am looking to work with the data science team to understand and deploy the machine learning models at scale.

Problem Statement

The goal of this Machine Learning Nanodegree Capstone project is to analyse the sentiment of various twitter tweets which is publicly available.

The tweets are related to financial news which have been labelled by a human for training and testing purpose of the Machine Learning model. There are approximately 8000 tweets which have been labelled with labels **Positive** and **Negative** for tweets indicating a **Positive** sentiment and **Negative** sentiment respectively.

Data Sets and Inputs

I have prepared a list of approximately 8000 Twitter tweets and labelled them accordingly with **Positive** and **Negative** labels. The input data feature in this case is the Twitter Tweet content and I will try to predict the Sentiment of the Tweet content.

Note: Neutral sentiments are not included in the labelled data.

Solution Statement

I have decided to use Convolutional Neural Network (CNN) classifier to predict the sentiment (positive or negative) of a tweet

Benchmark Model

I am using the Afinn word list based approach of Twitter sentiment analysis which is currently in use in my company as my benchmark to compare my CNN based model.

Evaluation Metrics

I will use the accuracy score and F1-Score of my model for evaluation.

Project Design

Data Preprocessing

First step is to collect the data i.e. Twitter Tweets and label them with positive and negative labels based on the sentiment of the tweet. Once I have the labeled data set, I will encode the labels to 0 for negative and 1 for positive sentiment respectively. Next step is to preprocess the tweet content and remove punctuations, links (with tag), tweet handles(with), hash tag (with) and certain special characters.

Splitting data

I will split the tweet dataset into training and test datasets using sklearn's split function. I going to do 80 to 20 split for training and test data set respectively

Model

I am choosing a 1-D Convolutional Neural Network (CNN) using Keras and Tensorflow to build train the model

References

Afinn - https://github.com/fnielsen/afinn

Another Twitter sentiment analysis with Python - https://towardsdatascience.com/another-twitter-sentiment-analysis-with-python-part-11-cnn-word2vec-41f5e28eda74

Sentiment Analysis Using Convolutional Neural Network - https://ieeexplore.ieee.org/document/7363395/