Sentiment Analysis of Movie Reviews

1 Group members

Name	Roll number	email-id
Harsimran Bedi	1611CS03	harsimran.mtcs16@iitp.ac.in
Nikhil Cheke	1611CS02	cheke.mtcs16@iitp.ac.in

2 Abstract of the project

Our project topic is Sentiment Analysis of Movie reviews. In this project we will do automatic classification of subjectivity of IMDB movie reviews. Opinion mining or Sentiment Analysis will be done using neural network. The neural network will be trained and then tested to compute the accuracy of our model. We will take two classes of sentiment positive and negative.

3 Introduction

Sentiment analysis has become very popular over the years. Every manufacturer, service provider wants to know how much a customer likes their product or service. With the increasing number of review blogs and forums, a vast resource of data has been created which can be mined for sentiment on a particular entity. For example, the entire cast and crew of a movie would want to know the public opinion of their movie.
State of the art opinion miner has been developed using supervised classification methods. The fundamental working principle of such techniques is feature extraction. In any machine learning approach feature extraction and selection has to be done manually, thus posing a challenge. Nowadays, deep learning approaches have started to be employed for many tasks. The main advantage is that we do not need to extract features anymore. Neural Network takes care of that part.
This project aims at applying neural network approach to the popular problem of sentiment analysis of movie reviews. We will use Keras deep learning library for this task. We will build a simple multi-layer perceptron model and one dimensional convolutional neural network model on imdb movie review dataset.

4 Literature survey

Work on sentiment analysis of reviews and opinions about products is plentiful, and specially in the area of movies.Some researchers in this area has used combination of machine learning and some NLP methods, they trained supervised classifier SVM and achieved an accuracy of 85%[1]. The work in twitter sentiment analysis with recursive neural networks obtained the average F1 score as 0.512 for one hidden layer and 0.517 for two hidden layer[2].Another proposed network named character to sentence convulational neural network(CharSCNN) uses two convolutional layers to extract relevant features from word and sentences of any size.This model obtains an accuracy of 81.9% with random word embeddings[3]. Also some researchers has extensively analyzed and classified sentiments in online opinions.Few of them has also examined economic impact of the reviews. With this work some researchers has figured out effect of reviews on the sales.

The highest accuracy achieved on IMDB dataset is 88.89% [4]. However there was a Kaggle competition named Bag of Words Meets Bags of Popcorn. The winner achieves area under ROC curve as 0.9925[5].

5 Resources

Keras built-in IMDB movie review dataset

6 Work Done

- Description of the data:
  The dataset we are going to use for our project is "Large Movie Review Dataset" often referred to as IMDB dataset. This dataset contains movie reviews along with their associated positive and negative sentiment labels. it is intended to serve as a benchmark for sentiment classification. The core dataset contains 50,000 reviews divided equally 25,000 train and 25,000 test sets. The overall distribution of labels is balanced(25k positive and 25k negative). In the entire collection no more than 30 reviews are allowed for any given movie. The train and test sets contain a disjoint set of movies. In label train and test sets a negative review has a score $<= 4$ out of 10, and positive review has a score $>= 7$ out of 10. Thus reviews with more neutral ratings are not included in the train/test sets.

- Exploration of different NN
  Initially we explored a multi layer perceptron model for this task.The embedding layer is be used as input layer and the output of this layer is flattened to one dimension,then we add one dense hidden layer of 250 units with rectifier activation function,the output layer has one neuron and we use sigmoid activation function to output value of 0 and 1 as predictions. We have tried different number of hidden layers and epochs. The highest accuracy achieved with this Neural Network Model is 87.37.
  Next we explored 1-D Convolutional network. The embedding layer acts as the input layer followed by the 1-D CNN layer. Then a hidden layer is added with 250 units followed by the output layer. In this model also we experimented with different number of hidden layers, epochs and the different parameters like dropout, kernel size, strides. The best accuracy achieved is 89.07%. We also explored LSTM model for our classification problem. We used 100 memory units and dropout value of 0.2. It resulted in 85.56% accuracy. We tried combining 1-D CNN and LSTM by adding a conv1D layer after the embedding layer followed by the LSTM layer. This resulted in 86.15% accuracy. Since the results obtained in 1-D CNN were better, we didn't experiment further.

- Error plot on Validation set
  The epoch vs loss plot on validation set is given in Figure 1.

- Final Architecture
  1-D CNN:
  Input layer is the embedding layer with 32 length word vectors followed by 1-D convolutional layer. The number of filters is 32, strides are 1, kernel size is 3 with relu as the activation function. The max pool size is 2. Then there is a hidden layer of 250 units with rectifier activation function followed by an output layer of one unit with sigmoid activation function. The number of epochs is 2. This is our final architecture which achieves the best accuracy which is 89.16%.

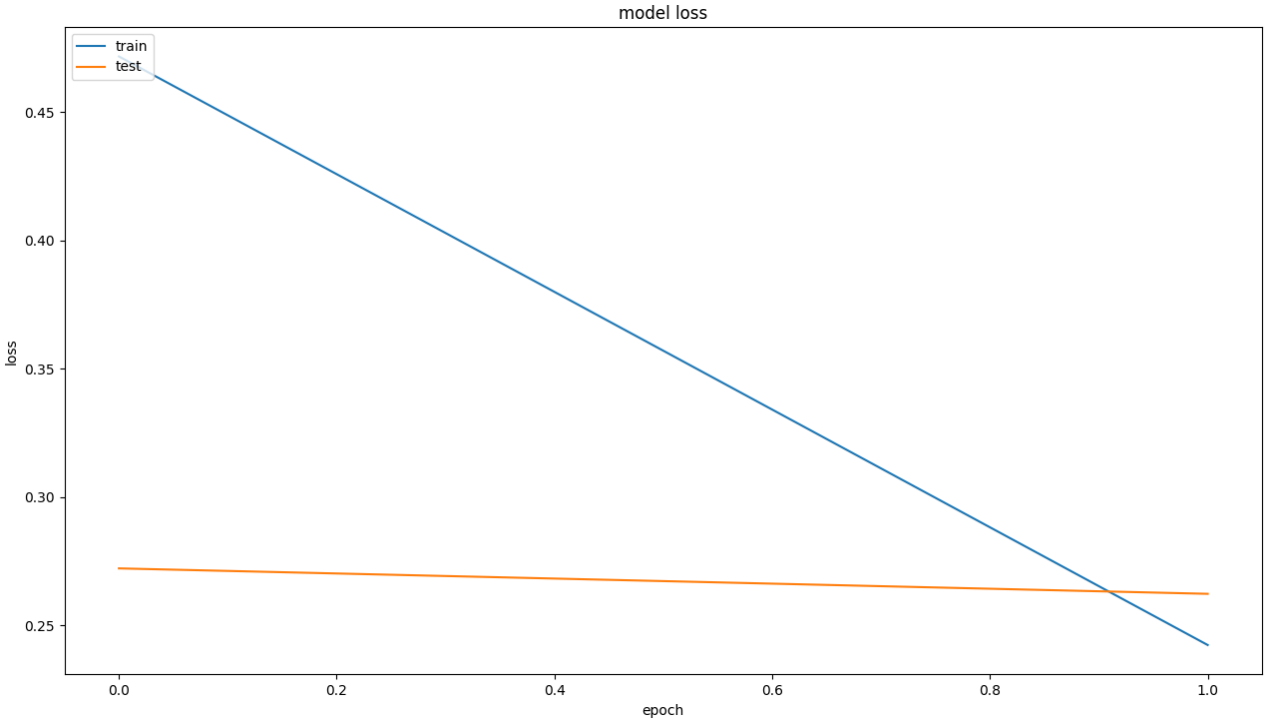

Figure 1: Epoch Vs Loss

- Result of different optimization techniques

Sr. No	Neural Network	Parameters	Accuracy(%)
1.	Multi layer Perceptron	hidden layer=1 epochs=2	87.37
2.	LSTM	memory units=100 epochs=3 dropout=0.2	85.56
3.	LSTM and CNN	memory units=100 epochs=3	86.15
4.	1-D CNN	hidden layer=1 epochs=2 dropout=0.0 strides=2	87.53
5.	1-D CNN	hidden layer=1 epochs=2 dropout=0.2 strides=2	88.70
6.	1-D CNN	hidden layer=1 epochs=2 dropout=0.4 strides=2	88.92
7.	1-D CNN	hidden layer=1 epochs=2 dropout=0.4 strides=1	**89.16**

- GitHub location of project
https://github.com/nikhilcheke/cs551

7 Future work

- Sarcasm detection with our model
- Humor detection with our model

References

[1] Dave, S. Lawrence, D. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings WWW 2003, 2003.

[2] Ye Yuan, You Zhou.Twitter Sentiment Analysis with Recursive Neural Networks.

[3] Cicero Nogueira dos Santos,Maira gatti.Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts

[4] Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher. Learning Word Vectors for Sentiment Analysis. Association for Computational Linguistics 2011.

[5] Kaggle Competition leader board https://www.kaggle.com/c/word2vec-nlp-tutorial/leaderboard