SENTIMENT ANALYSIS

November 11, 2018

OVERVIEW

1. Project Background and Description

The domain of this project is Natural Language Processing and sub-domain is Sentiment Analysis (Opinion mining).

Opinion mining (sometimes known as sentiment analysis or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.[Source: Wikipedia]

Lot of research work has been done on this topic as this poses certain challenges like sarcasm which highly misleads the conventional models. Below are few works carried out which inspired me.

1. *Zhang, Lei, Shuai Wang, and Bing Liu. "Deep Learning for Sentiment Analysis: A Survey." arXiv preprint arXiv:1801.07883 (2018).*
2. *Sosa, Pedro M. "Twitter Sentiment Analysis Using Combined LSTM-CNN Models". Konukoii.Com, 2018, http://konukoii.com/blog/2018/02/19/twitter-sentiment-analysis-using-combined-lstm-cnn-models/. Accessed 22 Feb 2018.*

2. Problem Statement

Given a sentence, the model identifies whether the sentiment of the sentence is either positive or negative based on the information learned using Supervised Learning technique.

3. Requirements

Tensorflow Hub, Tensorflow, Keras, Gensim, NLTK, Numpy, tqdm

4. Dataset

Twitter Sentiment analysis corpus(Sentiment140)

Link: http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip

By looking at the description of the dataset from the link, the information on each field can be found.

0—the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
1—the id of the tweet (1997)
2—the date of the tweet (Fri Nov 15 23:58:44 IST 2018)
3—the query (iPhoneX). If there is no query, then this value is NO_QUERY.
4—the user that tweeted (suryavamsi)
5—the text of the tweet (iPhoneX is cool)

Dataset has 1.6million entries, with no null entries, and importantly for the "sentiment" column, even though the dataset description mentioned neutral class, the training set has no neutral class. There are 800,000 positive tweets and 800,000 negative tweets. So, the dataset is a balanced one.

I have split the dataset into positive tweets file and negative tweets file and shuffled them to maintain random distribution. After that these tweets are tokenized and then lemmatized to store only the root words. The limit on number of tokens per tweet is kept as 15.

Dataset is split as 90% for training and 10% for testing.

5. Solution Statement

In this we are going to use a CNN-LSTM DeepNet with 3 different word embedding algorithms Word2Vec, FastText and Universal Sentence Encoder and compare the results using these algorithms.

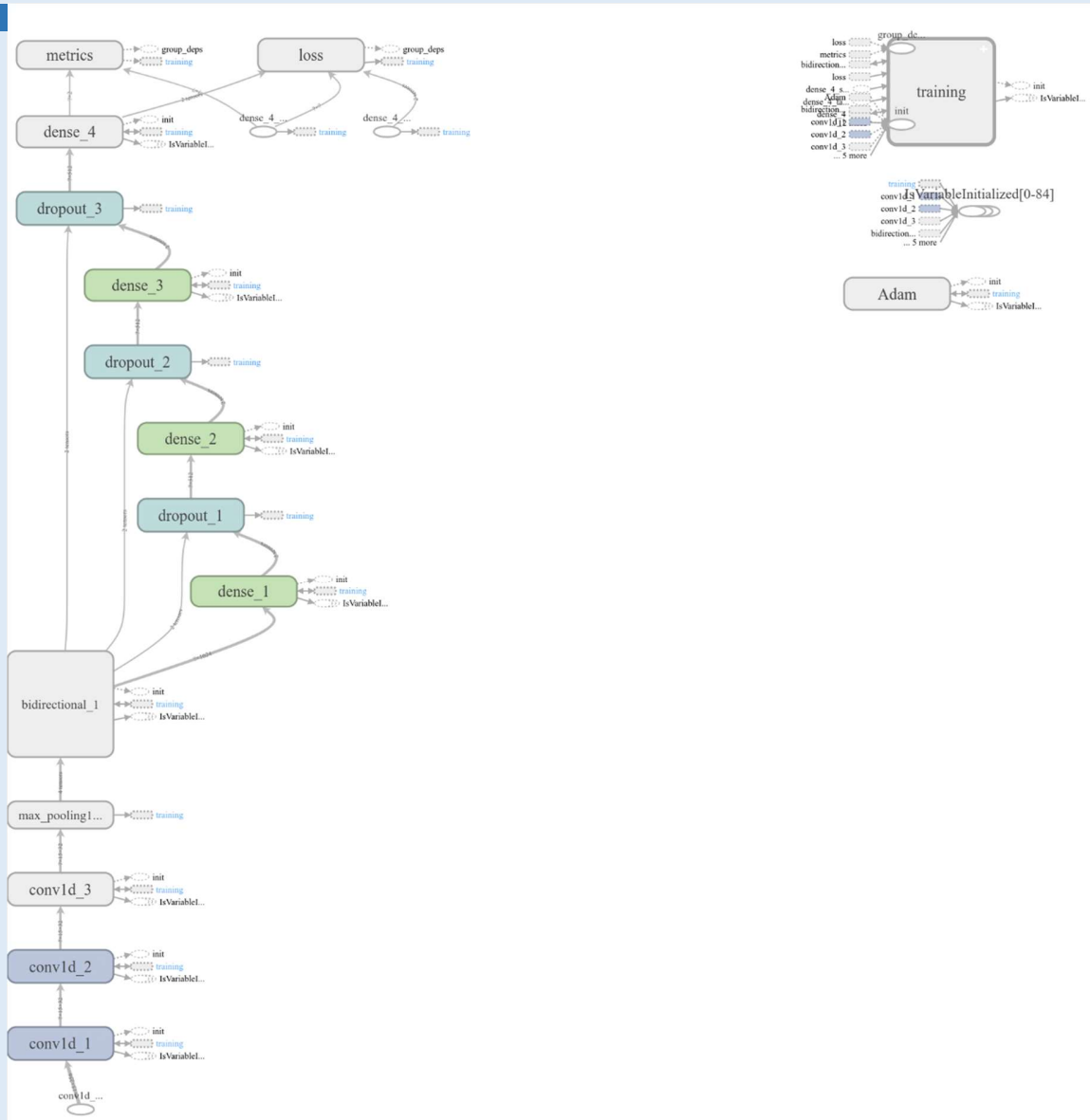
6. Benchmark model

Sentiment140, a standalone tool on Tech test bed achieved 67.82% accuracy and on Telco test bed achieved 71.79% accuracy on the Sentiment140 dataset.

7. Evaluation metrics

The evaluation metrics of the model are 'validation loss' and 'accuracy'.

8. Project design



The above image is the architecture of model generated using TensorBoard.
Conv1D -> Conv1D -> Conv1D -> Max Pooling1D -> Bidirectional LSTM -> Dense -> Dropout -> Dense -> Dropout -> Dense -> Dropout -> Output

This is the architecture of the network. The reason to use CNNs is that though they are widely used in image-related tasks, they are able to detect patterns in the data. So, a 1D convolutional filter

would help to identify the patterns which help us to distinguish the negative tweets from the positive ones more easily.

Long-Term Short Term Memory (LSTMs) are a type of network that has a memory that "remembers" previous data from the input and makes decisions based on that knowledge. These networks are more directly suited for written data inputs, since each word in a sentence has meaning based on the surrounding words (previous and upcoming words). In our particular case, it is possible that an LSTM could allow us to capture changing sentiment in a tweet.