Final Project Report

Analysis of Human Speech and Text Emotion Using CNN

Team Members

Aman Singh- A20491333
Radhika Malhotra- A20491601
Kandalam Sai Sree- A20496672

Professor

Dr. Yan Yan
Department of Computer Science
Illinois Institute of Technology

1 Introduction

The way humans express their opinion is truly based on the emotions they carry in day-to-day activities. Human emotions play a vital role in everyone's life. The communication among different aspects depends on their emotions and the way they are conveying them. These emotions in any sort of communication can be detected using several factors such as sound pressure level(SPL), pitch, tone, timbre and time gap between every word. The main aspect that can be used to detect emotion utilises a goldstandard of list of features that can be defined from human experts who are consistent in their speech. Another methodology which is used till today that got adopted during the construction of Valence Aware Dictionary for Sentimental Reasoning (VADER) is widely used sentimental lexicon.

Analysing the emotion has different applications. The applications can be improvised by considering modes of data which is used in emotional analysis should not have any restrictions to detection either speech or text.

In general, text provides semantics which are in relation to the context. Additionally, speech is one of the important factor as it provides a large number of characteristics that can be robust and efficient to the model. Moreover, motion detection is also important that can consider facial expressions, head movement, hands movement that can be used in detecting emotions.

The above three modes i.e., speech, text and motion can be obtained by considering the IEMOCAP dataset. The dataset was widely used for emotion detection. This dataset was collected at University of South California(USC) by the team Signal Interpretation and Analysis Laboratory(SAIL). The IEMOCAP database contains audiovisual data which includes speech, text, video and faces of actors that are recorded in motion making it a multi-model database.

1.1 Summary of the problem

We, as humans, have different ways of expressing things. These feelings usually vary according to the circumstances. The different types of feelings usually expressed stages from anger, disgust, worry, happiness, unhappiness, and wonder. We can detect all these feelings and emotions from the sentences they use while they speak, i.e., from phrases or text. These emotions play a major role in day-to-day life because it depends on play human behavior and consumer interactions. We have the ability to investigate textual content and speech to understand feelings can result in advanced programs. This undertaking awareness on emotion popularity within the context of the text in addition to speech and any viable integration to make a better model for detection. We are taking into consideration the previous activities or researches that took place so as to create a model with better detection activities.

We are planning to use a module technique that investigates speech and textual content, i.e., sentences one after the other, and detect emotional levels. To implement the stated module from input with special processing techniques for speech and textual information, a one-directional convolutional neural network(CNN) is used. We use datasets called IEMOCAP and REVDES to train and improve the quality of detection.

This task presently aims at distinguishing among seven exclusive forms of feelings which can be correctly carried out for hospital purposes in addition to assistive generation consisting of chatbots to offer a far extra natural and customized approach to a response.

1.2 Previous works, Methods and Results

There were several researches that were made on Emotions detection. Few of them are:

1.2.1 Detection and Analysis of Human Emotions through Voice and Speech Pattern Processing

In this research, voice and speech analysis can be used to detect human emotions. Here algorithmic approach is used in analysing various voice attributes during speech.

Human speech has different characteristics. While having conversation, for each type of speech there is change in emotions which is determined with voice attributes. Taking all the attributes into picture, the attributes are used to measure and differentiate different emotion states.

Researchers have taken three test cases such as normal, angry and panic state. The normal states stands as a base for other two states. Two speech samples are taken so as to measure pitch, SPL and time gaps. The analysis provided orations in panicked and angry state.

1.2.2 VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text

VADER is most popular model which can be used in detection of simple emotional analysis. VADER stands for Valence Aware Dictionary for Sentiment Reasoning. In this research VADER is been taken and compared it with eleven state-of-the-art benchmarks. These include
1. LIWC
2. The General Inquirer
3. SentiWordNet
4. ANEW and
5. Techniques based on Maximum Entropy, Naive Bayes, and Support Vector Machine (SVM) algorithms.
Here three things have been researched on:
1. **Constructing and Validating a Valence-Aware Sentiment Lexicon: A Human-Centered Approach**
Using word banks about nine thousand feature candidates were obtained. For this lexical features such as emoticons, acronyms and slangs are added to portray emotions. Each of these features were rated from -4 to 4 where 0

detected neutral expression, -4 detected negative and +4 represented highest positive expression. This was done by using Wisdom-of-the-Crowd (WotC) approach for sentiment expressions. The result of all these features can be used as standard features for VADER.
2. **Identifying Generalizable Heuristics Humans Use to Assess Sentiment Intensity in Text**
Few tweets were picked from the twitter. From those tweets 400 positive and 400 negative snippets were extracted. Analysis were made on these snippets. The extracted snippets were then rated by human on the same scale. This analysis was made to compare the accuracy and following results were obtained.
1. Punctuations and Capitalization can increase the intensity of text.
2. Degree modifiers can increase or decrease the intensity of any given text.
3. The conjunction such as "But" changes the emotion after the latter making the emotion to be dominating or dictating.
3. **Controlled Experiments to Evaluate Impact of Grammatical and Syntactical Heuristics**
In this heuristics were identified before few tweets were removed and variations are made by controlling grammatical and syntactical meanings. The VADER is been imposed. Further analysis of VADER against ML models prove that it performs as well as or better than the machine learning models in their respective trained domains.

1.2.3 Emotional Analysis of Bogus Statistics in Social Media

This is one of the unique study where emotions are analysed based on posts in Social Media platform. Currently, there are many fake news spreading across the platform. This research detects emotion based on these posts so as to detect fake as well as truthful information. This can be fulfilled by using recurrent neural network and comparing it with exiting neural network techniques.

RNN is used to evaluate sequences with hidden layers which helps in understanding and learning data from past layers. It takes the whole data, trains and predicts results. This can also help in passing dynamic inputs of various lengths without affecting size of model.

Backpropogation algorithm is used in training the model. This helps in measuring error functions and ignores time function. Chain rule can be used to define error. The output which is obtained results in positive and negative outputs.

1.2.4 Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)

This research is based on rule-based approach defining a set of rules and inputs. The supervised classification techniques has been explored in this study. The approach states how dataset can work when it is divided into testing and training sets based on 80:20 rule. Here 80Here two accuracies have been obtained
1. Accuracy for Naive Bayes, SKLearnBernouli and Sklearn SVC() as 70.4%, 70.1% and 75.3%.
2. Accuracy for Decision Tree, Random Forest and KNN has been represented as approximately 52%, 80% and 71%.

Metrics for the given approaches gives us a series of probable algorithms that could be implemented for emotion analysis to achieve results.

2 Problem Description

Artificial intelligence and Deep learning has been current technologies which are being used in many sectors such as industrial, medical and many more. Among the applications under these technologies, Emotion recognition is one of the major sector which uses Artificial intelligence and their different models to bring the accuracy in the result.

We are focusing on the same Emotion detection. Here two attributes are used text and speech for detecting emotion. The main motto is to establish a model which differentiates among several emotions that a person usually expresses in his/her day-to-day life in a normal conversation. Here along with Artificial intelligence we use Machine Learning algorithms to our model. So our model will be able to detect emotion(i.e., happy, sad, angry and many more) of a person who is having conversation.

Here many low-level and high-level features are taken into consideration so as to get good results in the output. We consider real-time inputs and pre-loaded inputs so as to check with the detections.

Here the dataset IEMOCAP is used to carry out several experiments. Here other features such as anxiety as well as depression can be used in our research so as to get details of a patient related to mental health, which can be useful for several treatment purposes.

There were many approaches that were done in the field of detecting the feelings using the IEMOCAP dataset. This dataset contains twelve audiovisuals that contain different actions like text, audio and many more. The previous models lack in considering few lower level features, which prevents the model from giving the proper accuracy. Lower level features play a major role in detecting features of speech.

Our proposed model tries to differentiate between emotions while having a conversation. So in our model will consider the lower level features as well as high-level features. By applying algorithms from ML as well as AI, our model would be able to predict accurate emotions.

2.1 Description of the Model (Theory and Applications)

Firstly, we take speech transcriptions along with the speech features in it. The speech features are Spectrogram and MFCC. Both the features together provide deep neural network with semantic relationships and low-level features to differentiate among different emotions. This can provide results with good accuracy score.

2.1.1 Model 1: CNN Model based on Speech Features

The recent researches in several technologies has developed many successful applications. The research conducted on speech processing and concepts such as CNN and long short term memory(LSTM) and deep learning methods has made it successful. CNNs are mostly used in many applications for extracting information from raw signals. Some of the applications where CNNs are used are speech recognition, image recognition and many more.

In our research, we use Spectograms and MFCCs as they are commonly used to represent speech features along with CNNs for emotion detection.

2.1.2 Model 2: CNN Model with MFCC input

MFCC stands for Mel Frequency Cepstral Coefficents. It is a representation of Short-Term Power Spectrum of sound. This is based on linear cosine transform of log power spectrum on non-linear Mel-scale of frequency. The MFCC is very popular. So, in our model we use the same for our research on emotion detection.

The hyper-parameters and the python package (librosa) used for MFCC generation are similar to the ones described for Spectrogram generation. The only difference is that 40 MFCCs per window are generated compared to the earlier mentioned 128 Spectrogram coefficients per window.

2.1.3 Model 3: Text Model based on TextBlob

For creating the text model we use TextBlob. TextBlob is a python (2 and 3) library for processing textual data. It gives the polarity of the entered text, which indicates it's emotion. A negative polarity would mean a negative emotion such as anger, disgust, sadness, etc. A positive polarity would mean any positive emotion such as happiness, surprise etc. Any value close to zero, would mean a neutral expression. We use speech transcriptions in various sentimental analysis applications. In emotion detection, it is necessary for our model to know the context of utterance so as to predict intent in a proper way. For example: if we consider a word "good". In general, we know the word good relates to something positive. But, in some conversations good can be used in sarcastic way.

2.2 Architecture of Model

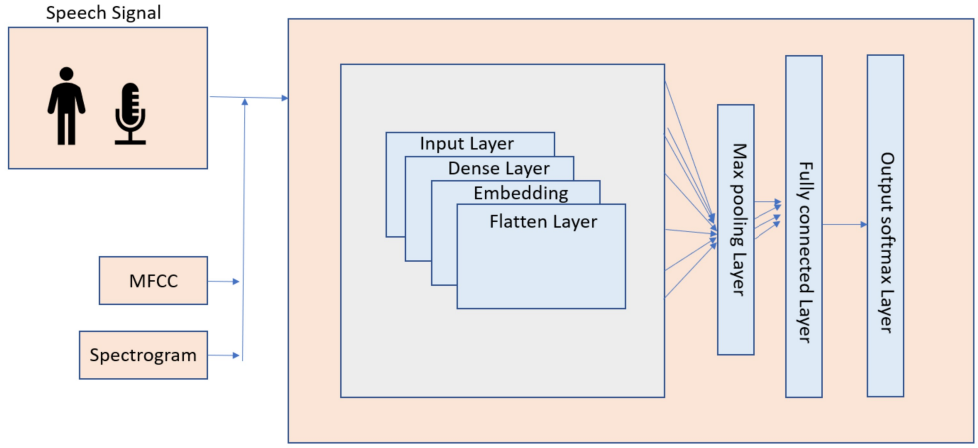

Figure 1: Speech Emotion Analysis Proposed Architecture.

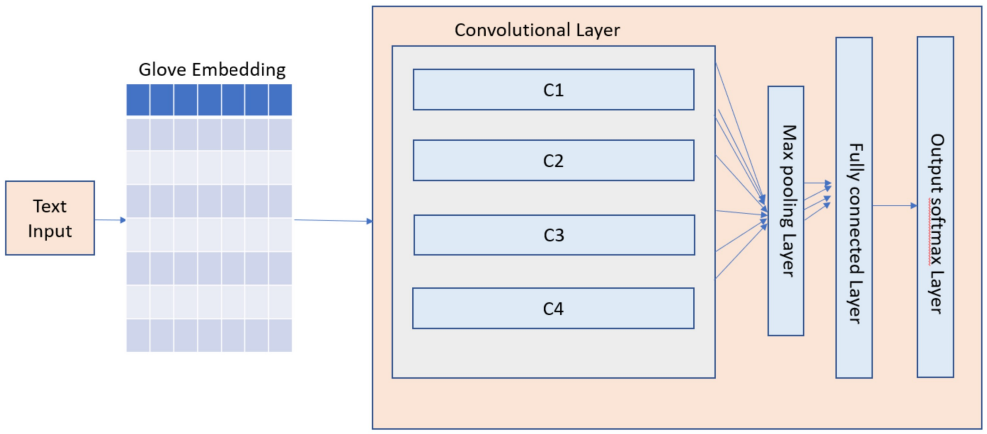

Figure 2: Text Emotion Analysis Proposed Architecture.

Results

We make confusion matrix in order to check the predicted values whether they are correct or incorrect. Based on the results we can state that there are two emotions.
1. Fear
2. Surprised
These two emotions are more commonly misclassified audio and text files.

The emotion recognition model through speech and text input proved to be quite efficient and novel. With the inputs from datasets we were able to train and create models which provided greater accuracy than some of the pre-existing models.
The main findings from the previous model with the current model are:
Previous Models
1. Emotion detection on multi-models on IEMOCAP dataset using deep learning technology.
2. Here the accuracy obtained for Test model was 54% and for speech model it was 64.78%.
3. Speech emotion recognition using heterogeneous feature unification in deep learning gives an accuracy of 64%.
Our Proposed Model
1. Human, speech and text recognition using the methodology of TextBlob and Convolutional Neural Network(CNN).
2. Here the accuracy obtained is 95.05%.

Conclusion and Future Work

All the current technologies have become a major part in our life. Every thing is dependent on many technologies. Machine Learning and Artificial Intelligence play a major role in our life cycle. Many of the lives are also saved because of these technologies. In this project we tried to work on the major topic currently trending everywhere. We were able to successfully create two deep learning model which when provided with suitable inputs i.e. .wave extension files and .csv files can classify and analyses between 7 different emotions viz. calm, angry, surprised, sad, neutral, disgust, happy. Currently these models are capable of predicting the emotion of voice inputs as well as text inputs from a user with an efficient accuracy but a little human effort is required for the preprocessing. We believe that with a mode powerful system and a varied dataset the results of these modules can be improved to an extent that they can be implemented in real world problems and evolve into a larger AI. In future, we will try to work on incorporating CNN in Text Sentiment Analysis and try to increase our model accuracy greater than 95.05%.

1 Output

1.1 Accuracy

```
Loaded model from disk
accuracy: 95.05%
```

Figure 1: Model Accuracy.

1.2 Speech and Text Output

```
if '_' in livepredictions[0]:
    print('Emotion:',livepredictions[0].split('_')[1])
else:
    print('Emotion:',livepredictions[0])
```

```
Emotion: sad
```

Figure 2: Output of Speech Analysis.

```
df
```

	Sr No.	Utterance	Speaker	Emotion	Sentiment	Dialogue_ID	Utterance_ID	Season	Episode	StartTime	EndTime
0	1	Oh my God, he's lost it. He's totally lost it.	Phoebe	sadness	negative	0	0	4	7	00:20:57,256	00:21:00,049
1	2	What?	Monica	surprise	negative	0	1	4	7	00:21:01,927	00:21:03,261
2	3	Or! Or, we could go to the bank, close our acc...	Ross	neutral	neutral	1	0	4	4	00:12:24,660	00:12:30,915
3	4	You're a genius!	Chandler	joy	positive	1	1	4	4	00:12:32,334	00:12:33,960
4	5	Aww, man, now we won't be bank buddies!	Joey	sadness	negative	1	2	4	4	00:12:34,211	00:12:37,505
...	...	...	...	...	...	...	...	...	...	...	...
1104	1174	No.	Monica	sadness	negative	113	9	6	2	00:19:28,792	00:19:29,876
1105	1175	What? Oh my God! I'm gonna miss you so much!	Rachel	sadness	negative	113	10	6	2	00:19:33,213	00:19:35,965
1106	1176	I'm gonna miss you!	Monica	sadness	negative	113	11	6	2	00:19:36,175	00:19:37,967
1107	1177	I mean it's the end of an era!	Rachel	sadness	negative	113	12	6	2	00:19:39,094	00:19:40,928
1108	1178	I know!	Monica	sadness	negative	113	13	6	2	00:19:41,138	00:19:42,638

1109 rows × 11 columns

```
[ ]  score = TextBlob("this is horrible")
     score.sentiment
```

```
Sentiment(polarity=-1.0, subjectivity=1.0)
```

Figure 3: Output of Text Analysis.