I'm Positive/Negative/Neutral

Homayoun Banazadeh*
Karly Kussainova*
Pranjal Keshari*
hbanazad@sfu.ca
karly_kussainova@sfu.ca
pkeshari@sfu.ca
Simon Fraser University
Burnaby, British Columbia, Canada

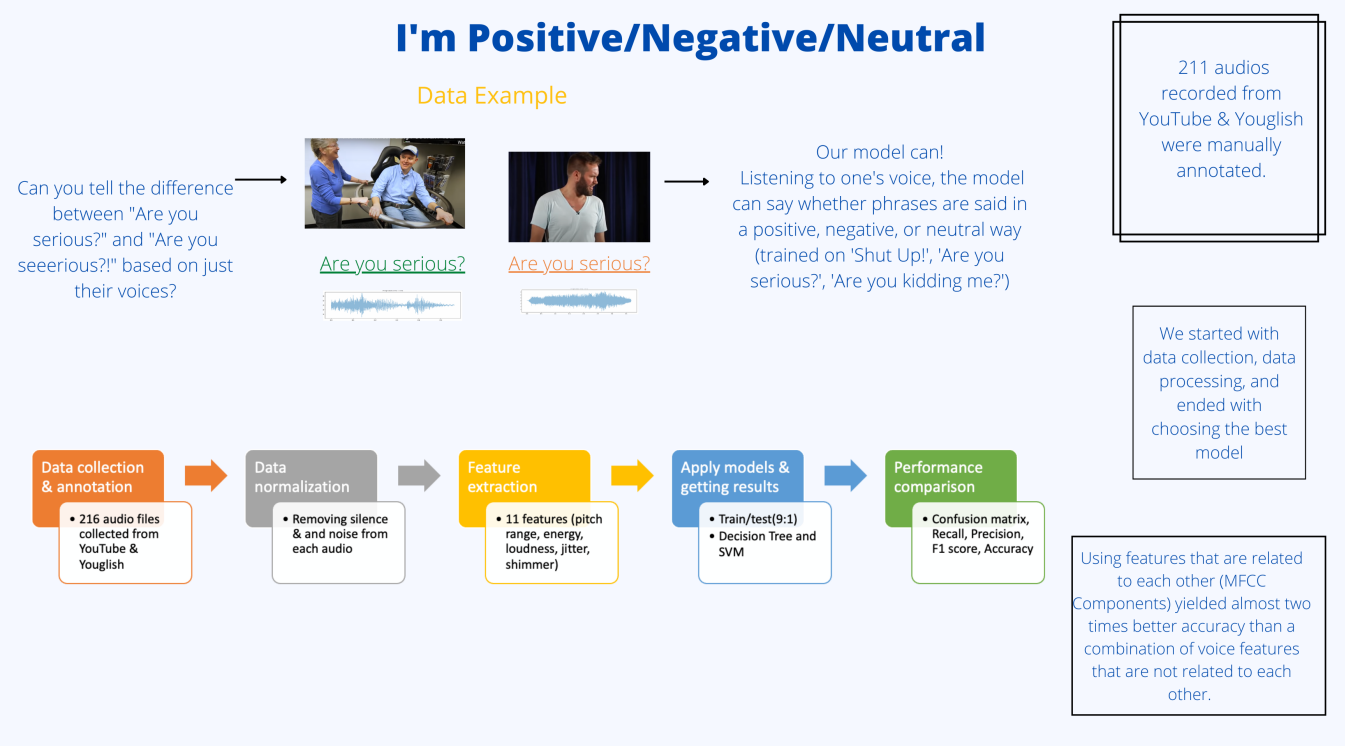


Figure 1: Poster Described the approach and data collection method

ABSTRACT

Using voice/audio data for social signal recognition is an integral part of affective computing and natural language processing (NLP). However, classifying the given data into particular social stands (emotions) is a challenging task. A variety of approaches to cluster the audio data exist, but only a few of them are effective. In this paper, we propose an approach to classify the audio data into positive, negative, and neutral using different feature extraction techniques. We first collect and annotate a dataset of 211 audio files with 3 phrases, normalize them, and finally apply different machine learning (ML) algorithms to classify each audio. The main advantage is that we use a variety of feature extraction techniques and compare them to extract the variability invoice data. The most effective model identifies the intonation of the voice with 74% accuracy compared to the other 2 models. Our work provides an annotated dataset on 3 phrases and an effective model, which extends and justifies existing methods in social signal recognition using voice/audio data.

KEYWORDS

social signal, SVM, Decision Tree, classification, performance

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
I'm Positive,Negative,Neutral, Burnaby, BC,
© 2022 Association for Computing Machinery.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Homayoun Banazadeh, Karly Kussainova, and Pranjal Keshari. 2022. I'm Positive/Negative/Neutral. In *Proceedings of January 31st 2022 (I'm Positive,Negative,Neutral).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Can you ask the person to shut up, but mean absolutely the opposite of what you said? The answer is yes! The nature of human emotions is a complicated phenomenon questioned by the greatest minds throughout history. Although understanding the meaning of 'shut up' (or any other phrase) is comprehensible given the context, facial expression, and intonation, what are the criteria that differentiate the direct meaning of 'shut up' from its positive (excited) meaning?

All of these are the questions that are asked when we talk about Human-Robot interaction (HRI). An interactive robot Pepper, some voice interfaces such as Amazon Alexa, OK Google, Siri, and other interactive systems for customer service are successful examples of artificial intelligence (AI). Although some of these systems demonstrate impressive results, most of them fail to detect the social signals, especially when using voice data. Thus, scientists propose different methods to handle this issue. Some researchers use various modalities combining different types of data, while others only focus on audio data.

Various studies are conducted to identify an emotion given a dataset of audios and using different modalities. One of the researches focuses on both facial expressions (video) and emotional speech (audio), where authors combine two modalities to classify 6 basic emotions (anger, dislike, fear, happiness, sadness, and surprise) based on 2 human subjects [1]. Similarly, the other research focuses on detecting the emotion of the song based on its lyrical and audio features [2]. In this paper [2], the authors combine 2 modalities, where the lyrical features are generated by segmentation of lyrics during the process of data extraction, and features like energy, tempo, and danceability are extracted, to compute Valence and Arousal values.

At the same time, other studies for social signal or 'emotion' recognition are only based on one modality: the audio itself. For example, the research on the evaluation of musical features for emotion classification uses a ground truth data set of 2904 songs that have been tagged with one of the four words "happy", "sad", "angry" and "relaxed", on the Last.FM website [3]. They then extract 55 features from the audios using the MIR toolbox and use k-nearest neighbor (KNN) and support vector machines (SVM) to classify the songs. Another research on audio signal processing for speech emotion recognition uses Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) datasets to classify the audios using a SoftMax classifier [4].

To train the model that would be able to differentiate emotions for this paper, we will focus on audio data only. To narrow down and systemize the study, we chose 3 phrases: 'Shut Up', 'Are you serious', and 'Are you kidding me'. All these phrases can have different moods. For example, the first phrase can be used in its actual meaning angrily (negative), it can be used when the person is surprised and excited (positive way), or it can be said without involving any emotions (neutral). The audio samples are collected using the Internet sources and annotated into positive, negative, or neutral manually. Given the prior work done and the secondary research we have conducted, the features of the audios such as loudness, energy, jitter, etc. help cluster different types of emotion. As a result of extracting some of these features, several models could be trained to detect whether the emotion of the phrase was positive, negative, or neutral.

Our project stands out amongst the related work because firstly, we collected our data and secondly, we compared a combination of feature selections and machine learning models to grasp a feeling of what the best model can look like. We based the philosophy of our paper on the fact that the right set of features can yield a much better result than just using a variety of models on a set of features that may not harmonize together. We tested this by creating 2 feature sets and using 2 models.

2 APPROACH

The general approach to building such a system had 4 steps:

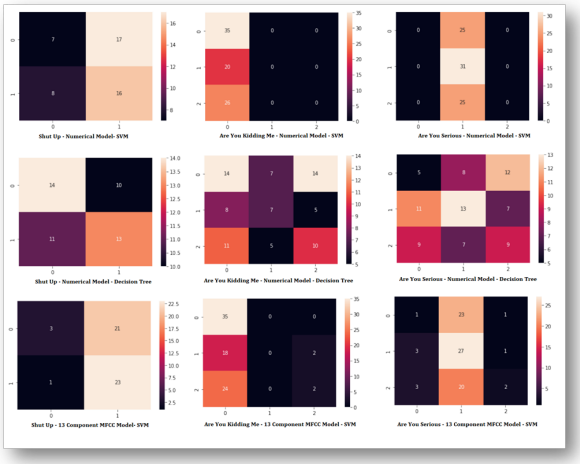
(1) Collect and manually annotate data. Going through Youtube [5] and Younglish [6] websites, recording the audios of different people pronouncing 3 phrases, and manually annotating/naming each audio file into positive, negative, neutral. After converting all audios into .wav format using online converters [7], they were placed under 3 folders for each phrase.
(2) Preprocessing and normalizing the dataset. Removing noise and unnecessary silence from the audios, making sure the length of each audio does not exceed 3 seconds. Some of the work was done manually by cutting the files, while another part was done using numpy and librosa.
(3) Feature extraction and training model. Extracting necessary features (numeric/graphic) from the audios and training Decision Tree and SVM classifiers on the features. Choosing the best model based on their performance.
(4) Cross-Validation to get the scores. Calculating Accuracy, Weighted Precision, Weighted Recall, and Weighted F1-Score based on 5-fold cross-validation and shuffling of the data.

However, there are different methods that were used to extract features and train the models. The first method is based on numeric values of the audio, while the second one uses graphical features for training.

For the numeric method, we extracted 11 features (rmse, chroma_stft, spec_cent, spec_bw, rolloff, zcr, mfcc, intensity, sample_width, frame_rate, frame_width) from each audio returning them first as an array, and then, as a mean. The features were extracted using librosa, NumPy, and pydub (AudioSegment) libraries, and the dataset was created using the Pandas library. After creating a labeled and processed dataset of 211 audios as a Pandas dataframe, the dataset was divided into testing and training sets (ratio 8:2). After training on the Decision Tree Classifier (trained on 11 features) and SVM (trained on the first 10 features), the SVM outperformed its competitor with 41.6% of accuracy after applying 5-fold cross-validation. Sklearn is the main library that was used for training and performance evaluation.

As for the graphical method, we used the means of 13 Mel-frequency cepstral coefficients (MFCC) components to get a graphical representation of the audio and fed that graphical data into a polynomial SVM classifier. After getting the features we realized 13 features were translated to 159 data points for each row. Some of these features were Null for most of the rows, and they were replaced with 0. We experimented with different number of MFCCs including 3,5,10,13,15 and 30. While the results were varying each time we ran the experiment, 13 MFCC always had one of the best performances. We also experimented with different SVM classifiers including linear, polynomial and, RBF. Polynomial SVM would often yield better results.

Model name	Accuracy	Precision	Recall	F1-score
Shut Up				
Karly's SVM	0.47	0.47	0.60	0.37
Karly's Decision Tree	0.55	0.55	0.61	0.52
Homa's SVM	0.54	0.54	0.71	0.42
Are You Kidding Me				
Karly's SVM	0.43	0.43	0.75	0.26
Karly's Decision Tree	0.38	0.38	0.41	0.38
Homa's SVM	0.45	0.45	0.67	0.31
Are You Serious				
Karly's SVM	0.38	0.38	0.76	0.21
Karly's Decision Tree	0.33	0.33	0.34	0.32
Homa's SVM	0.37	0.37	0.51	0.25



3 EXPERIMENTS AND RESULTS

First, to explain our evaluation method, we used shuffling and 5-fold cross-validation to report the average scores for Accuracy, Weighted Precision, Weighted Recall and Weighted F1 score. Weighted gives the mean of parameter with weights equal to class probability and thus, it is something that can be used for multiclass classification. Other methods are "Macro" and "Micro". We are not certain how or why the Accuracy score ended up being equal to the Weighted Precision score in all the experiments, but that is something to further explore.

An interesting hypothesis to explore is whether the 13 MFCC components yield better results than the 11 numeric components. Based on initial evaluations, the MFCC components were producing accuracy results as high as 74 percent, however, these numbers were reduced to 50-60 percent in later trials. It is not evident to us why these scores keep changing so often, but the only explanation we could think of could be not having enough training data.

Therefore, based on the variability of the results, it remains unanswered questions for us how to choose the best set of features that captures all the variability in the tonality of the phrase, and even if found, how to find the best Machine Learning algorithm for that those set of features. We firmly believe that finding answers to those questions is what shapes the art of Data Science, and we think that at this point we are just in an exploratory phase of learning and not at a level of understanding to provide sophisticated answers to those questions.

Second, the results are summarized below, with the confusion matrices for all of the models.

Third, to compare the performance of our features and machine learning algorithms, we can say Decision Tree with numerical features and SVM with 13 MFCC features produced better scores than the third model. In terms of phrases, the scores for "Shut Up" phrase were generally higher than other phrases. Perhaps the fact that no positive label was associated with this phrase was an influential factor, or maybe just because this phrase is shorter than the others it got better results.

In terms of comparing the results before and after removing the leading silence, we can say that both cases yielded similar results. We experimented with the scores in both situations, and they were relatively close. In other words, the variation in the running model multiple times was more significant than before or after removing the leading silence.

The recall scores were generally higher than other ones and we suspect the models produce better recall scores than other explanations which is this is just happening by accident.

Fourth, the dataset was collected using a manual collection of data from Youglish. We attempted several times to use an already available dataset used in similar research, but each time we would

encounter either the dataset was not what we wanted, or it was just not publicly available and we would need to submit a request.

Fifth, examples of the voice data we created can be found in the upload folder along with our code.

4 DISCUSSION

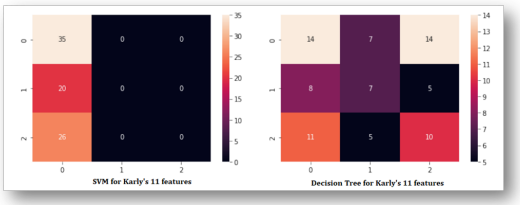
First, let's discuss what pattern we found interesting. The biggest pattern that our team found was that in previous experiments, 13MFCC features would yield a much better result when it came to "Shut Up" phrase, whereas in other phrases all the models would perform relatively the same. MFCC features for "Shut Up" would produce a result of a decent of 74 percent. However, when we integrated all the codes and ran the experiments again, the accuracy was reduced to 54 percent. We can have a look at the latest results in the previous section. One of the significant observations is that a recall score is always relatively higher than any other score. Could this be a result of our models producing better recall scores or is just happening by nature would be a question to explore.

Second, we want to discuss the limitations of our project. While collecting the dataset which contains overall 211 instances of 'Shut Up', 'Are you serious', and 'Are you kidding me', we manually need to record the audio file from the websites like YouGlish which help us to find the videos on Youtube speaking these phrases. While annotating these phrases some chances include human factors. The human factor could have affected our judgment on whether the emotion was positive, negative, or neutral. Since we are collecting the dataset manually, this restricted us to collect more data due to restrictions on time, we initially tried to automatically gather the data, however, API provided on YouGlish[6] is written in javascript, and had restrictions in downloading, cutting the required part, and downloading it in WAV format. This is the reason only gathered two hundred sixteen data points while applying our model. Automatic data collection would have provided more accuracy for our model and provided more accuracy since it would have more data for training and cross-validity.

Another limitation is the fact that Accuracy and Weighted Precision would always produce the same results. Our team is not certain why this might be happening but according to our scoring, these two parameters are always the same value.

Third, we would like to mention the future work of our project. We could involve making the data collection process automatic, training the data on more data, or possibly oversampling the existing dataset which is creating new samples that are changed a small amount from the existing sample. Using 211 samples might not be enough for the model to train properly. This will help us to train our data for the specific situation such as we could collect data from customer care representatives, after taking the data we could train and test them to understand sarcasm in customers. Another future work could be understanding the voice features, and what MFCC means in our model as this will help us use MFCC features to make our model more accurate.

In the confusion matrix for SVM features, we observed that the classifier tended to classify the phrases in only one category. Perhaps this could be more explored to find out why this phenomenon is happening. Decision Trees would often yield a more balanced result.



5 CONCLUSION

Overall, we can say that our results are much better than random chance which is 50 percent for "Shut Up" phrase due to having 2 outcomes, and 33 percent for "Are You Serious?" and "Are You Kidding Me?" phrases with 3 outcomes. Decision Trees had a more robust outcome whereas, for SVM, the model would prioritize predicting one label only. 13 MFCC components and 11 numeric components both resulted in similar caliber results, and perhaps a more detailed analysis could be performed on selecting the most ideal number of components for MFCC and removing noise, causing components from numeric features. To conclude, we can say that we have to build a machine that can identify when a person says "Shut Up" when they are angry or when they say it naturally with decent accuracy. The next step is to gather more data and manipulate the features to get better results.

REFERENCES

[1] L. De Silva and C. N. Pei, "Bimodal emotion recognition," *IEEE*, Aug. 2002. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/840655/authors#authors
[2] A. Jamdar, J. Abraham, K. Khanna, and R. Dubey, "Emotion analysis of songs based on lyrical and audio features," *Artificial Intelligence Applications (IJAIA)*, Jun. 2015. [Online]. Available: https://arxiv.org/abs/1506.05012
[3] Y. Song, S. Dixon, and M. Pearce, "Evaluation of musical features for emotion classification," *ISMIR*, 2012. [Online]. Available: https://www.researchgate.net/profile/Yading-Song/publication/277715954_Evaluation_of_Musical_Features_for_Emotion_Classification/links/557187c408ae7467f72ca201/Evaluation-of-Musical-Features-for-Emotion-Classification.pdf
[4] Mustaqeem and S. Kwon, "A cnn-assisted enhanced audio signal processing for speech emotion recognition," *Interaction Technology Laboratory*,, Dec. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/20/1/183
[5] "Youtube," http://youtube.com.
[6] "Youglish," https://youglish.com.
[7] "Online converter," https://cloudconvert.com.
[8] "Javascript api," https://youglish.com/api/doc/js-api.

6 APPENDIX

3.3 Collection Process

How was the data associated with each instance acquired?

The data is in WAV format and collected by manually recording each data point and converting it from mp3 to WAV format.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?

First, we tried using YouGlish API [8] and collecting the data automatically but did not work out, so we manually recorded 211 data points to train and test our model.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?

First, we searched the phrase on YouGlish, open it on youtube, record it, and convert it to WAV from MP3 using websites such as cloudconvert.com.

Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?

Since we collected our data points from Youtube, no one was involved in the data collection process.

Over what timeframe was the data collected?

The dataset is created between 5th Feb 2022 to 15 April 2022.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

Since the data was collected online, and it is acted data, there will be no impact on the subject. We also made the data more secure by hiding the name of the video from we were recording the audio and making our data set.

3.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

For both sets of features, we removed the leading, ending silence from the voice data and then normalized it. When it comes to 13 MFCC features, we replaced NA values in some columns with 0.

Was the "raw" data saved in addition to the preprocessed/ cleaned/ labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

We saved our raw in google drive, this is the link - Link for Google Drive (Please click Here)

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

We performed preprocess/clean/label using the python in the .ipynb file instead of using the software.

Contributions

Task	Responsible Person
Project proposal	Karly, Pranjal, Homayoun
Data collection	Karly, Pranjal, Homayoun
Code	Karly, Pranjal, Homayoun
Report writing	Karly, Pranjal, Homayoun
Presentation	Karly, Pranjal, Homayoun

Core Certificate Of the Team


PANEL ON RESEARCH ETHICS
Navigating the ethics of human research

TCPS 2: CORE 2022

Certificate of Completion

This document certifies that

Pranjal Keshari

successfully completed the Course on Research Ethics based on the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2: CORE 2022)

Certificate # 0000800525

11 February, 2022