Progress Report:
Applying a Fuzzy Approach to Emotion Classification of Song Lyric Summarizations
CS 585, UMass Amherst, Fall 2017

Dani Mednikoff & Jen Zhu

1 Outline

In our project, we focus on performing emotion analysis and classification on a corpus of summarized song lyrics. Emotion classification on songs could potentially make it easier for listeners to find songs they are interested on based on their moods. Despite the large quantity of research conducted on emotion analysis and classification, not nearly as much has been done using song lyrics as the sole input for emotion/mood classification. One past technique, for example, was to analyze the lyrics line by line. Our project goal is to also analyze songs via lyrics but in an alternate method. Currently, we analyze lyrics word by word based on a bag of words representation, calculating a song's average valence and arousal, and then map that to an emotion. Moving forwards, we plan to implement fuzzy clustering to generate a better evaluation of the emotions portrayed in a song.

2 Changes from Initial Proposal

After reviewing both feedback from our peers as well as our course evaluator, we decided to make a few changes from our original proposal. In most cases, people agreed using summarization would not increase our accuracy or help us in the process of song classification. Given the loss of detail that occurs with summarization, as emphasized by these comments, we decided to remove the summarization aspect from our project. We still believe the idea of song lyric summarization could yield interesting results, but concluded it would not helpful for our current project. We also realized that neither datasets we initially had proposed would work the way we envisioned them to, due to incomplete and unusable data in both sets. Thankfully, we were able to find a new dataset which aligned with our vision. Finally, after taking into consideration the suggestions regarding emotions themselves, we decided to include more emotions.

Using Russell's circumplex model of emotion as a guide, we have created a dataset of our own that includes 34 different emotions.

3 Dataset Statistics, Annotation Process, Issues Experienced

Our new dataset of lyrics contains the billboard top 100 charts from 1964-2015 https://www.kaggle.com/rakannimer/billboard-lyrics. The dataset includes 5,100 songs with information about each song, including rank, year, title, artist, and lyrics. After further inspection, we came across a few issues that could easily be solved using hand annotation. First, we removed instrumental songs and songs that were not in English. Afterwards, we were left with 5,057 songs. About 200 songs in the remaining dataset had no lyrics. We inserted the lyrics ourselves after removing all punctuation marks and line breaks and making all of the words lowercased. Afterwards, we created a copy of the dataset, and in the copy removed everything except for the song title and the lyric data. For our classification process the other data (i.e. rank, year, artist, etc.) was not necessary. However, we could potentially run some analysis later on to explore alternative patterns in the data.

With our dataset complete and functional we read in the lyrics and analyzed them word by word. During our classification process, we ran into issues we had not noticed before. Due to how both of us, as well as the creators of the dataset, had removed punctuation, some words were now concatenated together. For example, if there had been a word wrapped in parentheses followed by another word with no spaces in between, the words would join together. One example of this can be seen in the song "Halo" where the section "(now)its" became "nowits." This becomes an issue due to how we chose to calculate each song's average valence and arousal. Since our algorithm operates on a word by word basis, there is a chance that the word "nowits" is non-existent. However, the words "now" and "its" are more likely to appear, but we cannot use "nowits" to add to the count of either.

Another issue we came across during classification is how some of the song lyrics provided from dataset are incorrect or have chunks of text that are clearly not part of the lyrics in them. When calculating scores for the song "Happy" we noticed that the lyrics included the line "memory byamandah editor in an indepth interview with oprah pharrell williams famously broke down into happy tears when describing the impact his signature song had on people " in the middle of the lyrics. This was clearly a mistake made by the original creator of the dataset. Seeing how we luckily managed

to stumble upon this error in our dataset, we will have to go through each song in the future to eliminate similar mistakes.

We also utilized another dataset which provided us with the valence and arousal scores of around 14 thousand words http://crr.ugent.be/archives/1003 http://crr.ugent.be/papers/Warriner_et_al_affective_ratings.pdf. The data was gathered using participants from Amazon's Mechanical Turk and words from ANEW. The ANEW words contained a dictionary of words as well as their corresponding valence and arousal scores. The words are given scores in the range of 1 to 9. We used this dataset to assign the words in the lyrics to a corresponding valence and arousal score. If a word in the lyrics could be found in the ANEW dataset, we would add the valence and arousal scores to the running sum for the song. However, it was also possible that the word from the lyrics did not exist, which was frequent due to some of the issues discussed above as well as slang words such as yo and ya, and even musical words like oooh or la. In such cases, we would set their values to 0, essentially ignoring them. Along with the valence and arousal scores, came another score classified as ratings, which were generated by Mechanical Turk participants in the research study. The users scored how each word made them feel in terms of valence and arousal and these were averaged for each word giving them a score from 1-9. We kept these values in case we wanted to do additional analysis in the future.

For our third and final dataset, we worked off of a version of the valence and arousal scores of 34 emotions from http://www.computer.org/csdl/trans/ta/2013/01/tta2013010116-a html. It consisted of the name of the emotion, their valence score (a number between -1 and 1), and their arousal score (again a number between -1 and 1). We later scaled these to our dataset maximum and minimum arousal and valence scores to associate songs to emotions.

4 Classification Methods & Results

Since all of our datasets were stored in CSVs, we read them in and stored the values in dictionaries. In total, we created 3 dictionaries from reading CSVs - the valence and arousal associated to each word, the song and it's lyrics and the scaled emotion valence and arousals. First we created the valence arousal word dictionary, which mapped a word to a list of its valence and arousal. Then, we created another dictionary which mapped a song to an array of its lyrics. Afterwards, we calculated the average valence and arousal per song, and stored that into another dictionary. To calculate the averages, we generated a bag of words dictionary for each song, using each word in the

lyric as a key and the number of time the word appears as a value. Based on the bag of words, we would go through each element in the dictionary and check to see if it was also in the dictionary containing the ANEW words and their associated valence and arousal scores. If the word was not in the dictionary, we would detract the count of the word from the total length of the lyrics and continue onto the next word in the bag of words dictionary.

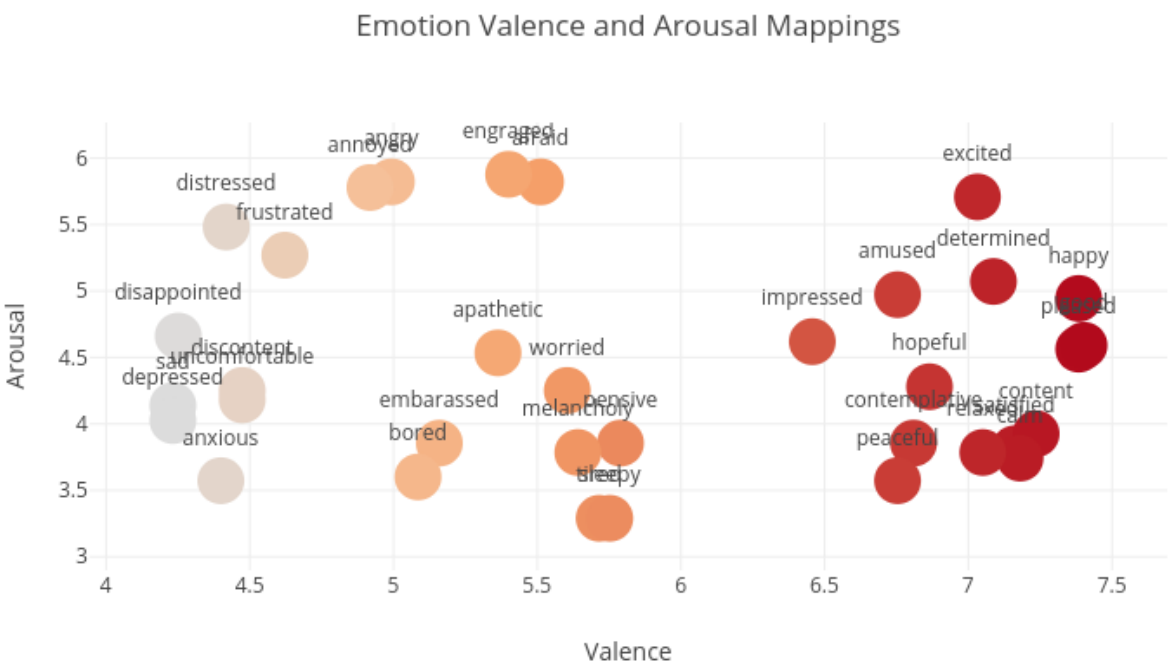
This proves to be a potential problem we plan to address through fuzzy clustering. If a word does not appear in the ANEW dictionary, it is treated as though it is non-existent in the song. Through the application of fuzzy k-nn clustering, we can address such an issue. If we were to average the valence and arousal for nearby words and help determine the word in question's valence and arousal. By doing so, we can even expand and build upon the ANEW dictionary we are currently working with.

If the word is in the ANEW dictionary, we retrieve the associated valence and arousal scores with the word, and add that to a running sum of the lyric's valence and arousal, keeping in mind the count from our bag of words. After we have finished looping through each word in the lyrics, we calculate the average valence and arousal for the song by dividing the sums by number of words in the lyrics (in total). Then, we store this in a dictionary mapping the song with it's average valence and arousal scores.

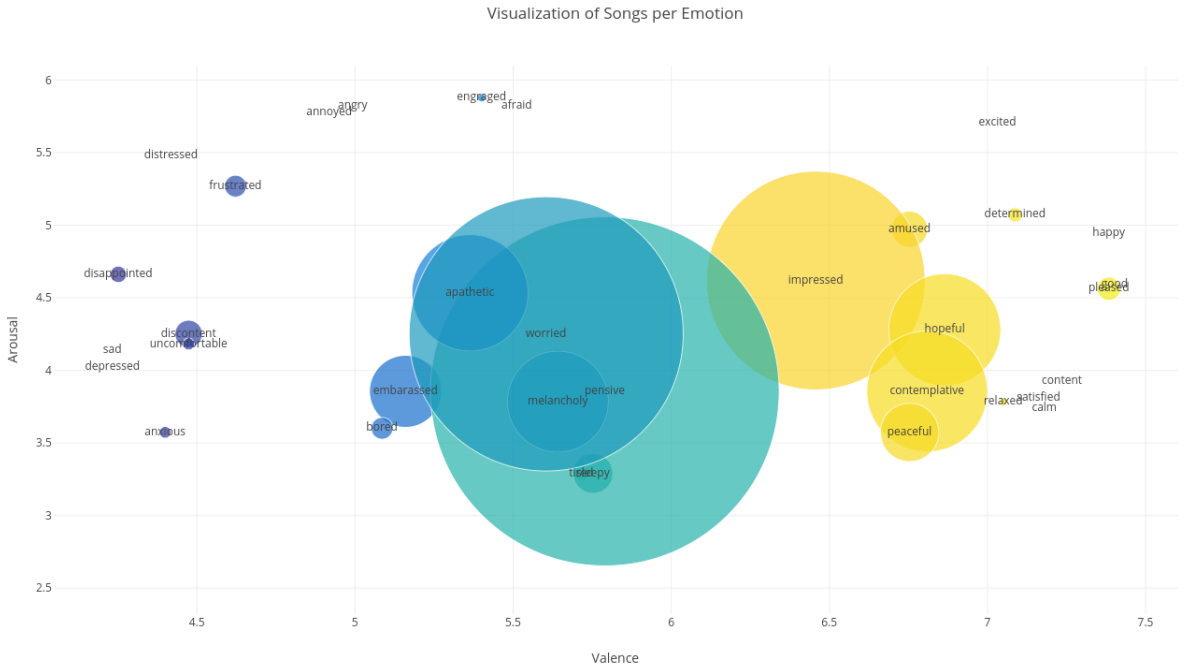
Once we have determined the valence and arousal scores for each song, we then calculate the maximum and minimum scores for both valence and arousal. Using these bounds, we scale the emotion valence and arousal scores to apply to our data set. Afterwards, we find the closest emotion's valence and arousal to the song's averages using the euclidean distance formula.

With our current baseline algorithms, we found that finding the nearest emotion may be an inaccurate representation of songs. For example, if a song had been in between two different emotions, we would not be able to display it as such. Additionally, the problem of words being left out is something we plan to solve by the end of the semester. Both of these issues will be addressed in our fuzzy clustering implementation. Currently, we have no testing or training set, simply because we are not implementing any machine learning tactics. However, when we move to fuzzy clustering, we plan to implement an 80/20 split (i.e. 80% for testing, 20% for training).

5 Graphs



Here we have a graph of the mappings of different emotions along our scaled valence and arousals. The color intensity varies depending on valence. You can view and interact with the data here: https://plot.ly/~jzhu0119/7.



Here is a diagram with the mappings emotions along the valence and arousal, however the size of the circle varies depending on the number of songs that mapped to that emotion in our dataset. You can view and interact with the data here:

https://plot.ly/~jzhu0119/9
To clarify, valence represents a spectrum from positive to negative emotions, and arousal is organizing emotions via low to high energy.

6 Future Goals

Our first goal is to fix the issue of incorrect lyrics. Since we are classifying the songs based on their lyrics it is extremely important that we make sure the songs lyrics are accurate. Our plan is to go through song by song and verify that the lyrics are correct.

Then, given we now have a baseline algorithm and an idea of which emotion a song may be classified as, our goal is to implement a fuzzy clustering technique. This should increase our precision and allow us to map songs to an emotion in a more cohesive way than previously. Instead of the vectors matching exactly or being within range, fuzzy clustering will match a song to its closest emotion as well as to multiple emotions, if applicable.

Finally, in our original dataset of ANEW words and ratings we noticed the ratings are meant to represent on a scale of 1-9 what emotions a word evokes in a reader. Given those ratings, we plan to use a similar algorithm to determine what emotions a song evokes in listeners. To clarify, our current project is classifying songs based on the emotion discussed in the song lyrics. This additional step would be to attempt to classify songs based on the emotions users think of when listening to the song.

Once all of the songs are classified, if time allows, a potential point of interest is to determine patterns in the data based on categories such as artist, year, rank, and ratings. Possible questions we could focus on are: Does an artist usually write happy or sad songs? Are the number 1 songs usually of a positive emotion or of a negative emotion? However, this is an extra feature of our project that will only be implemented if we have completed previous tasks.