Project Report - CMPE 256

Large Scale Analytics

Determining the type of Happy Moments using Natural Language Processing and Classification techniques

11.20.2019

Submitted to: Prof. Magdalini Eirinaki

Submitted By -

Hansraj Pabbati (SJSU ID : 012540541)

Neha Bindle (SJSU ID: 013763126)

Shrey Patel (SJSU ID: 012430652)

Table of Contents

Chapter 1 - Introduction

Motivation

What are the factors that contribute to the happiness of the people has always been one of the favorite research topics amongst psychologists. In our project, we are trying to leverage Machine learning and Natural language processing techniques to understand different attributes that affect happiness, and how different demographics are related to different categories of happiness. This is a rich dataset and it gave us the opportunity to predict multiple attributes based on different factors like demographics, and description of the moment

Objective

1. To predict the category of happiness based on the description of the moment using NLP text classification techniques.
2. To predict demographic variables such as Age, Country, Marital status, Parenthood, the Reflection period, and Gender depending on the description of a happy moment.

Chapter 2 - System Design and Implementation

Algorithms

For predicting the Happiness Category based on Moment Description

- **Multinomial Naive Bayes** - We used MNB as it is best suited for text classification with discrete features.
- **K Nearest Neighbor (KNN)** - As it works on a similarity measure, so we tried predicting the category based on the similarity of different input descriptions.
- **SGD Classifier** - It works well with data represented as dense or sparse arrays of floating-point values for the features which are basically our input after converting the text in the TFIDF sparse matrix.
- **SVM (Linear Kernel)** - As it is suitable for multiclass classification, and is memory efficient.
- **XGBoost Classifier** - As it is efficient and assumed to gibe better accuracy as compared to other algorithms. It can very well handle parallel processing, missing values, regularization, and Cross-Validation.
- Logistic Regression -
- **RNN for Text Classification** - Regular ML classifiers were not giving good accuracy so we tried with several deep learning algorithms.
- **LSTM** - This performs better than regular deep learning models due to its capability of selectively remembering patterns for long durations of time, thus giving better predictions.

For predicting the Demographic attributes based on Moment Description

- **Multinomial Naive Bayes** - We used MNB as it is best suited for text classification with discrete features.

For predicting the Age based on multiple attributes

- **Decision Tree Regressor** - We used a decision tree regressor based on multiple parameters to predict the age. This model is best suited for situations where prediction depends on the presence of an attribute or not. We converted demographics attributes into binary and the decision tree works really well.
- **Random Forest -** We used this one to predict the happiness category as well as to find an age range depending on the different parameters combined from two datasets.

Technologies and Tools

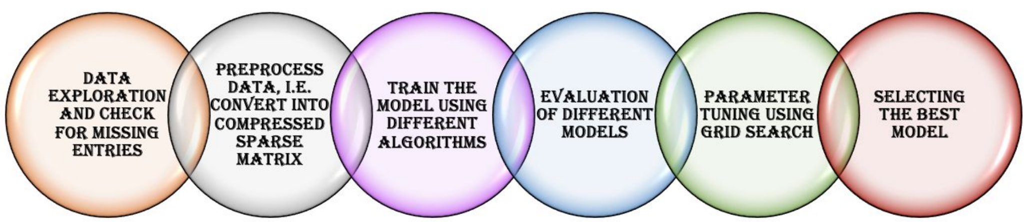
a. Scikit Learn, Keras, Tensorflow, Spacy, NLTK and NLP, Matplotlib, Wordcloud, Pandas, Numpy, Google Colab, Seaborn, HPC

System Design and Implementation

The most challenging part of this project was to extract meaningful words from the description of a happy moment. For this task, we worked on the following text-centric features-:

- **Structural features:** Review length, frequency of most common words.
- **Semantic features:** We used a pre-trained model Textblob to find out the polarity and sentiment for the description, and how polarity is related to different categories of happiness.
- **n-gram features:** We have extracted unigram, Bigram, Trigram features from the description of happy Moment
- Converting Categorical attributes into Binary Attributes.
- Label encoding for predicted category

System Flow for predicting the Happiness Category based on Moment Description



System Flow for predicting the Demographic attributes based on Moment Description

Chapter 3 - Experiments / Proof of Concepts and Evaluations

Dataset Description - HappyDB is a dataset of 100535 instances for happy moments crowd-sourced via Amazon's Mechanical Turk. It has two distinct datasets first one containing the description of happy moments and category describing the type of happy moments for 10,843 distinct workers. The second dataset contains the demographic details of those workers who described their happy moments. For our project, we have merged the dataset on the basis of worker ID.

Data Preprocessing

As this dataset contained both numerical and categorical attributes, so different techniques for preprocessing was required.

Preprocessing steps for Text Data (Description of Happy Moments)

- Removal of regex, stopwords and converting all the text into lower case
- Dropping the instances with null values
- Removal of Noise and less meaningful words to get a better analysis
- Tokenization of sentences using Spacy
- Stemming and Lemmatization
- Using Countvectorizer() to convert into a sparse matrix
- Using TF IDF() to convert the input into CSR Matrix.

Preprocessing steps for Demographic attributes

For 6 attributes named Age, Gender, Reflection_Period, Country, Parenthood and Marital status we converted each one into binary attributes as described below -

- Age less than 25 as 0, and greater or equal to 25 as 1. For age, there were several records that did not contain the age in the proper format or was nan. Those instances were dropped from the data frame.
- Parenthood no as 0 and yes as 1.
- If Single, divorced, Widower or separated as 0, Married as 1
- For Reflection Period 24 hours as 0, and 3 months as 1
- For Gender Male as 0 and Female as 1
- For Country, all the instances from the USA were labeled as 1 and rest everything as 0
- For predicted category numerical labels were assigned. Exercise - 1, Enjoy_the_moment - 2, achievement - 3, nature - 4, bonding - 5, affection - 6, and leisure - 7.
- If any instance contained age greater than 125 was dropped.
- Records containing null values were also dropped.

Methodology - For regular machine learning classification models following are the parameters used -

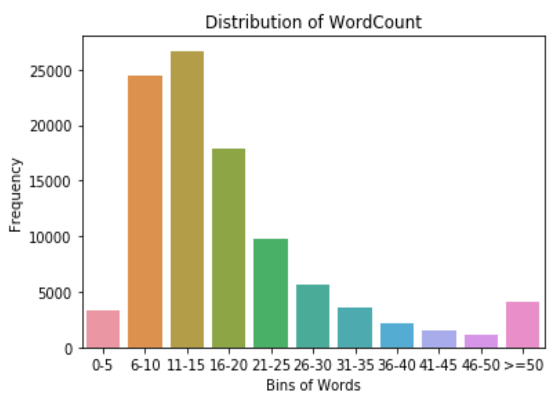
- For TFIDF - TfidfVectorizer(use_idf=True, min_df=3, max_df=0.5, ngram_range=(1,2) sublinear_tf= True,max_features=5000). These were decided on the basis of Grid Search.

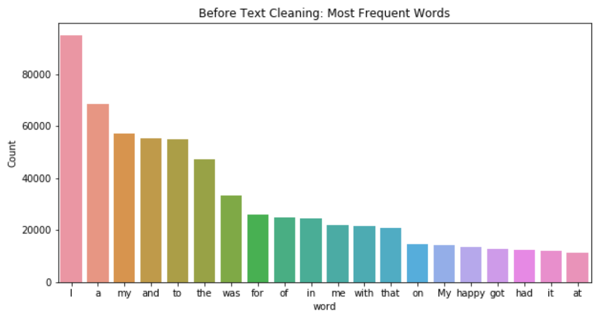
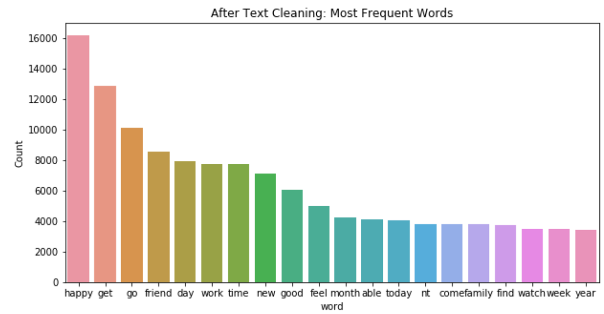
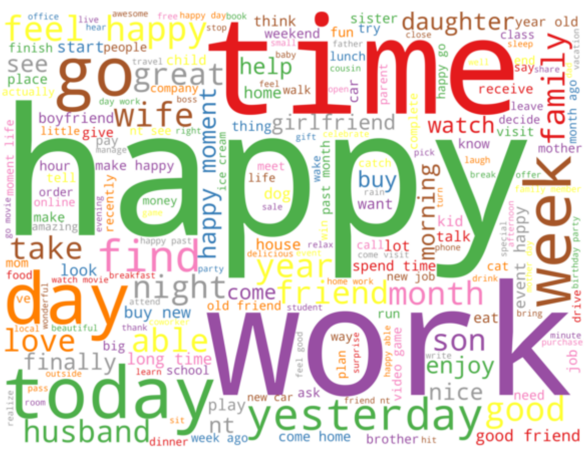
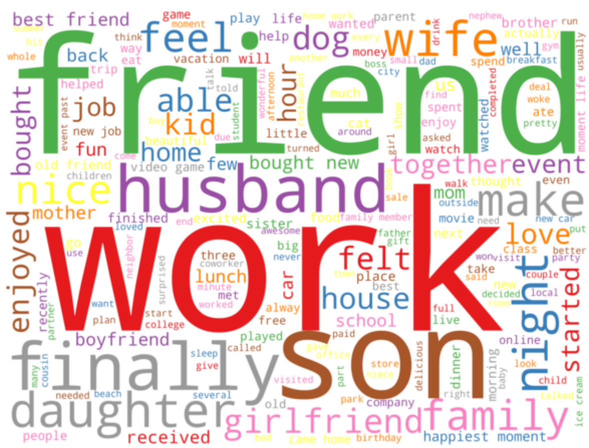
- KNN - 3 nearest neighbor was used to predict the category.
- SGD Classifier - Pipeline was made with the following parameters, loss='hinge', penalty='l2',alpha=1e-3, random_state=42, max_iter=5, tol=None
- Logistic Regression was performed on the default parameters.
- RNN used the maximum length of vector as 1000, and the maximum number of words to be considered as 10,000 which was decided on the basis of the total number of words in the dataset. Below mentioned are other parameters used

```
Model: "sequential_8"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_6 (Embedding)      (None, 1000, 64)          640000
_________________________________________________________________
lstm_5 (LSTM)                (None, 64)                33024
_________________________________________________________________
dense_7 (Dense)              (None, 7)                 455
=================================================================
Total params: 673,479
Trainable params: 673,479
Non-trainable params: 0
```

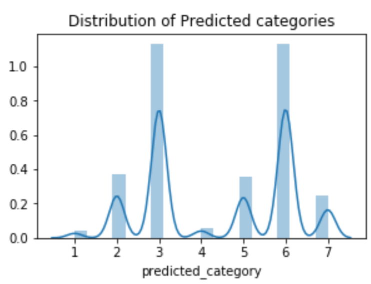
- LSTM - Parameters used were : MAX_NB_WORDS = 50000, MAX_SEQUENCE_LENGTH = 250 EMBEDDING_DIM = 100
- Train-test split was 80:20.
- 5 fold cross-validation used in Deep learning models.

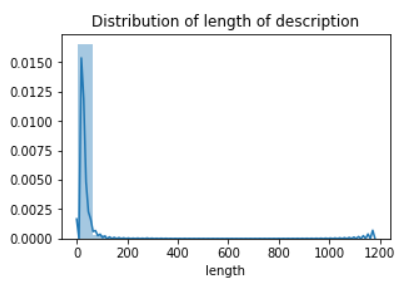
Graphs

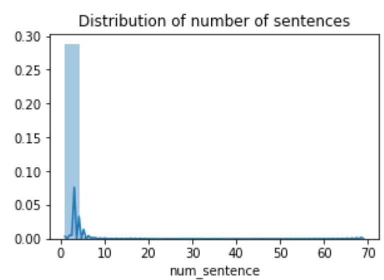


Before Text Cleaning	After Text Cleaning
	
	

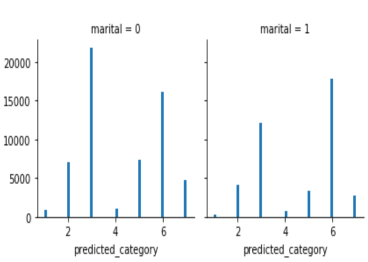
MetaData Analysis

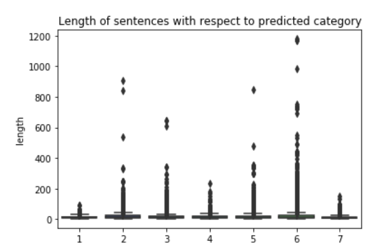






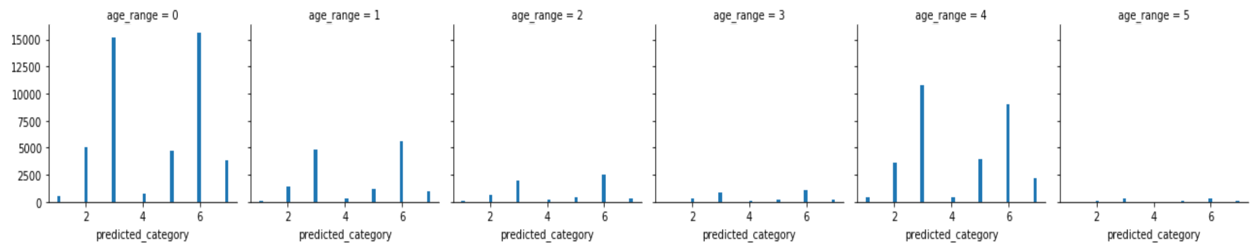
Descriptive analysis of Description of Happy Moment





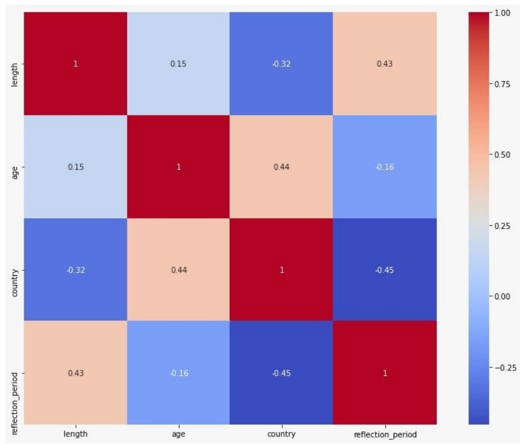
Bivariate analysis of Marital Status

Boxplot for the length of a sentence with predicted category



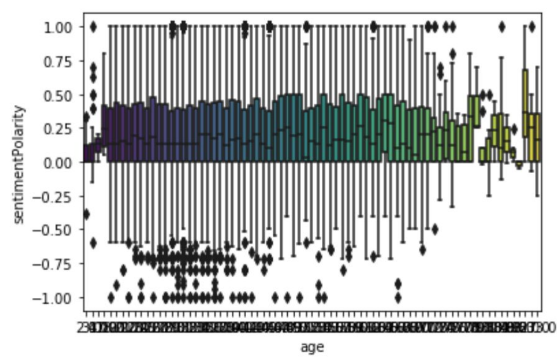
Bivariate analysis of age-ranges and predicted category.

HeatMap to see the correlation between different attributes



	importance
sentimentPolarity	0.273675
sentimentSubjectivity	0.272600
length	0.269831
predicted_category	0.068483
marital	0.054031
parenthood	0.047000
reflection_period	0.014380

Feature importance for predicting age_range as given by RandomForest Classifier.



Box plot for 'sentiment polarity' vs 'age'

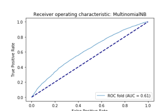
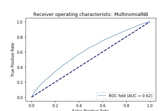
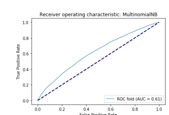
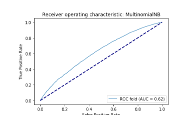
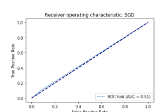
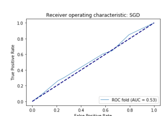
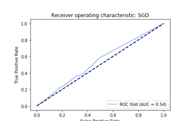
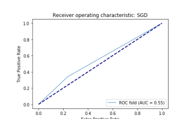
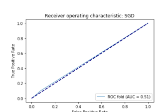
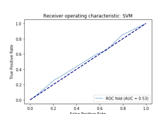
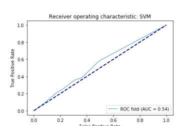
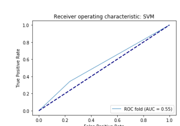
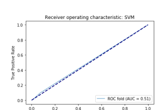
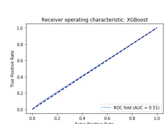
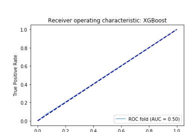
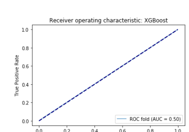
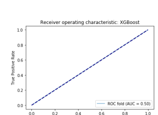
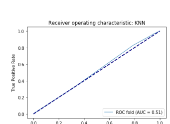
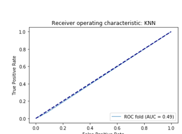
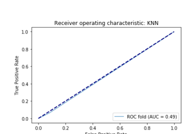
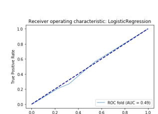
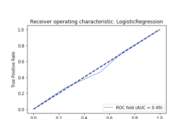
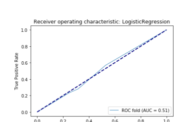
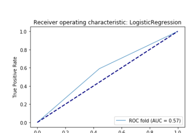
	hmid	wid	num_sentence	predicted_category	length	age	country	reflection_period
count	100535.000000	100535.000000	100535.000000	100535.000000	100535.000000	100442.000000	100332.000000	100535.000000
mean	78213.756722	2746.619028	1.340767	4.413289	18.298503	0.818622	0.788014	0.504342
std	29178.959001	3535.010347	1.297159	1.688824	21.474146	0.385333	0.408717	0.499984
min	27673.000000	1.000000	1.000000	1.000000	2.000000	0.000000	0.000000	0.000000
25%	52941.500000	410.000000	1.000000	3.000000	9.000000	1.000000	1.000000	0.000000
50%	78204.000000	1125.000000	1.000000	5.000000	14.000000	1.000000	1.000000	1.000000
75%	103490.500000	3507.000000	1.000000	6.000000	21.000000	1.000000	1.000000	1.000000
max	128766.000000	13839.000000	69.000000	7.000000	1179.000000	1.000000	1.000000	1.000000

The statistical distribution of values for all the numeric attributes

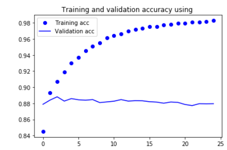
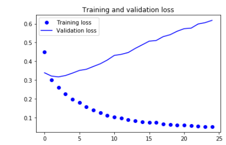
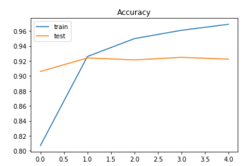
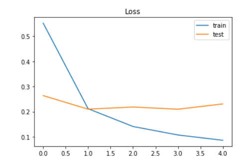
Analysis of results

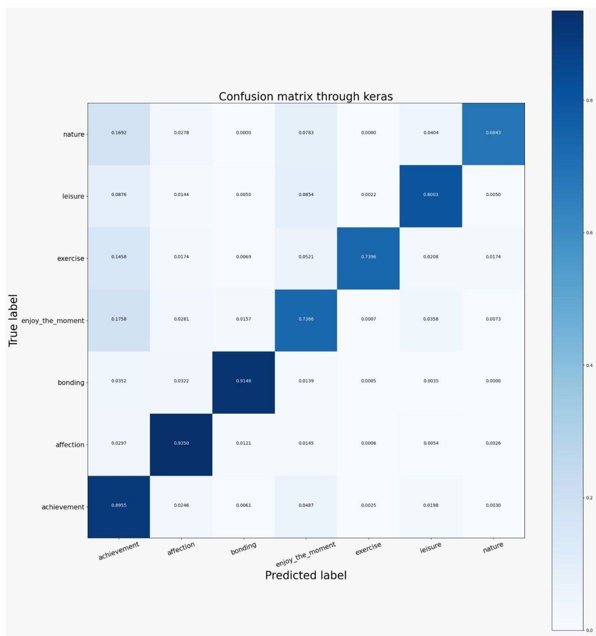
Machine learning models for predicting the Happiness Category based on Moment Description

Model Used	F1 score Regular	F1 Score with SMOTE	F1 Score with ROS	F1 score with Binary Conversion
Multinomial Naive Bayes	0.39	0.36	0.36	0.60
SGD	0.35	0.24	0.30	0.52
SVM	0.31	0.22	0.30	0.48
XGBoost	0.34	0.33	0.32	0.43
KNN	0.34	0.13	0.33	0.43
Logistic Regression	0.32	0.30	0.32	0.58

Model Used	ROC Curve Regular	ROC with SMOTE	ROC with ROS	ROC Binary Conversion
Multinomial Naive Bayes				
SGD				
SVM				
XGBoost				
KNN				
Logistic Regression				

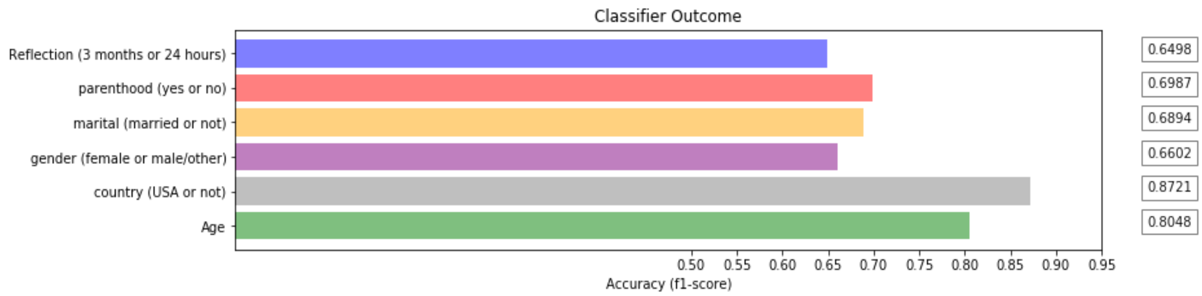
Deep learning models for predicting the Happiness Category based on Moment Description

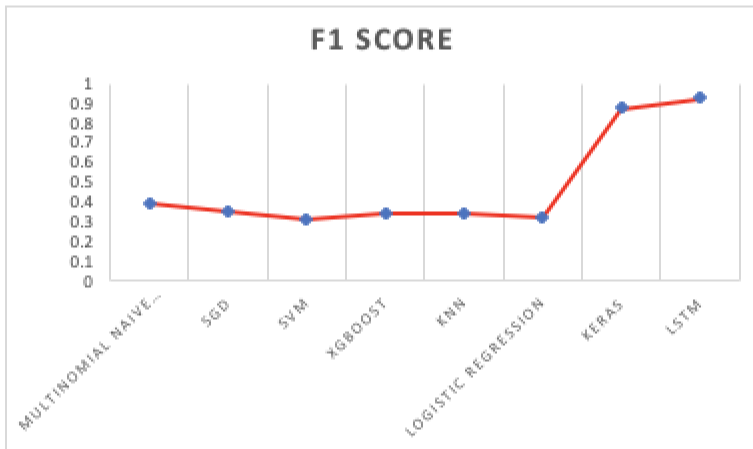
Model Used	Test Accuracy	Training and validation Accuracy	Training and validation Loss
RNN	0.87		
LSTM	0.92		



Multinomial Naive Bayes Model for predicting the Demographic attributes based on Moment Description

	Age	Country	Gender	Marital	Parenthood	Reflection_period
Accuracy	0.80	0.87	0.66	0.68	0.69	0.64





Chapter 4 - Discussion & Conclusion

Decisions Made

- We analyzed the demographic attributes and plotted the correlation heatmaps to check about which attributes have more importance.
- For predicting the positive emotion, we decided to use the LSTM model, based on evaluation methods.
- For predicting the demographics, most features like gender, marital status, etc by converting them to binary data.
- Having obtained the poor accuracy (RMSE score) with predicting age, we decided to transform the age to the categorical attribute of age ranges and utilize a Random forest classifier.

Difficulties faced

- On extracting new features, (ex. Length of a happy moment, age ranges) the new features definitely propagate the errors from the original features.
- The presence of irrelevant words and stop words occupied the majority of word cloud plot (ex. “Today”, “ going”, etc.). For this, we had to remove the stop words and the top irrelevant words.
- Deciding feature importances, feature extraction techniques, and hyperparameter tuning.

Things worked

- Deep learning models such as LSTM, gave a high accuracy of 0.9293 with a loss of 0.2 when trained with 5 iterations.
- Converting the demographic attributes to binary helped in some of the cases, as it made the attribute more balanced. Age_ranges turned out to be a stronger target variable than the age variable.

Things that didn't work

- Due to the highly imbalanced nature of the dataset, simple models such as SVM, XGBoost and Naive Bayes didn't give acceptable F1-score.
- Even with random over-sampling, these models gave only moderate accuracy.
- It turned out that for several demographic related target variables such as “age”, none of the other attributes were strong predictors.

Conclusion & Future Scope

- Considering the variety and volume of the HappyDB dataset, we discovered some very interesting patterns of this data. For instance, in the bivariate analysis of age_range with predicted_emotion_category, we found that age_bin=2 have the opposite relationship with age_bin=6, on “affection” and “bonding” emotions.
- We successfully developed models to predict the emotion category based on the text data. While analyzing and developing models for predicting other demographics of users such as gender, marital status, parenthood, etc we achieved moderate results, while certain attributes such as “age” had no strong predictors.

Chapter 5 - Project Plan and Task Distribution

Task	Contributor
Data Cleaning and Preprocessing	Hansraj, Neha, Shrey
Model Applications MNB SGD KNN Logistic Regression SVM XGBoost LSTM RNN	 Hansraj Neha Neha Hansraj Shrey Shrey Shrey and Neha Hansraj and Neha

Random Oversampling with SMOTE MNB SGD KNN Logistic Regression SVM XGBoost	 Hansraj Neha Neha Hansraj Shrey Shrey
Random Oversampling with ROS MNB SGD KNN Logistic Regression SVM XGBoost	 Hansraj Neha Neha Hansraj Shrey Shrey
Model Application with Binary Conversion MNB SGD KNN Logistic Regression SVM XGBoost	 Hansraj Neha Neha Hansraj Shrey Shrey
Evaluation of Results	Hansraj, Neha, Shrey
Report and Presentation **Presentation**	Hansraj, Neha, Shrey Hansraj, Neha, Shrey

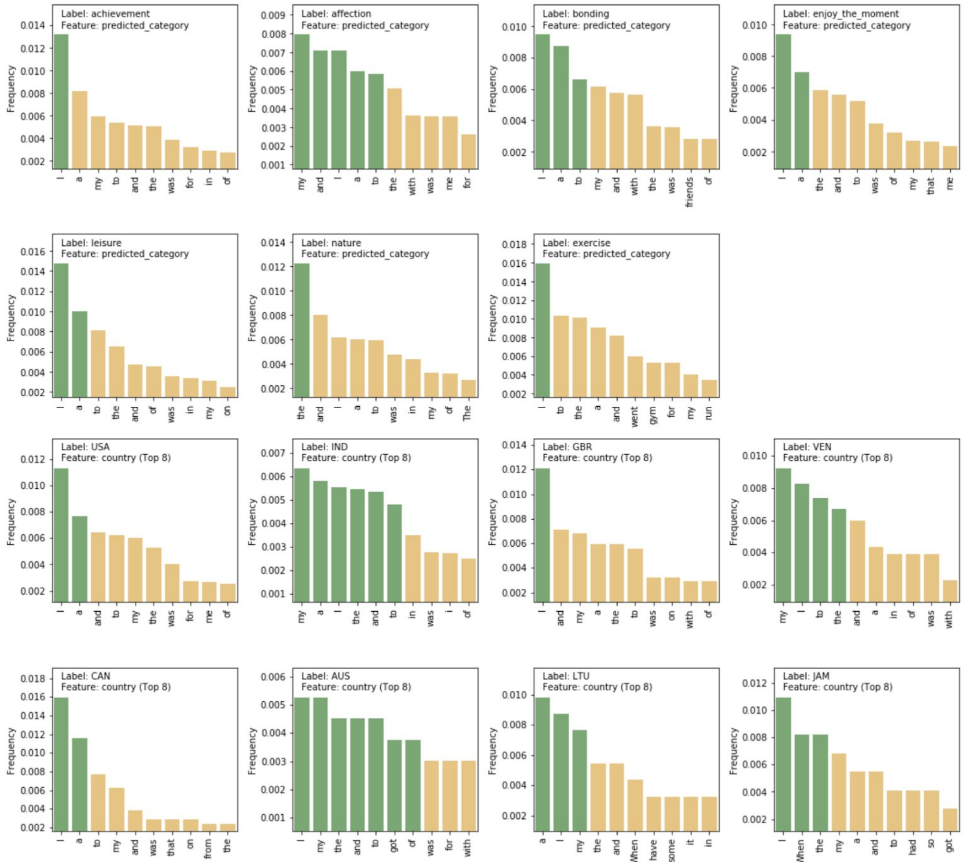
Our team made sure that everyone contributes in every step. We divided our work equally and ended up working as a team where everyone gave their best inputs.

As we have implemented multiple models, each one of us ran a few models to combine and evaluate the final result.

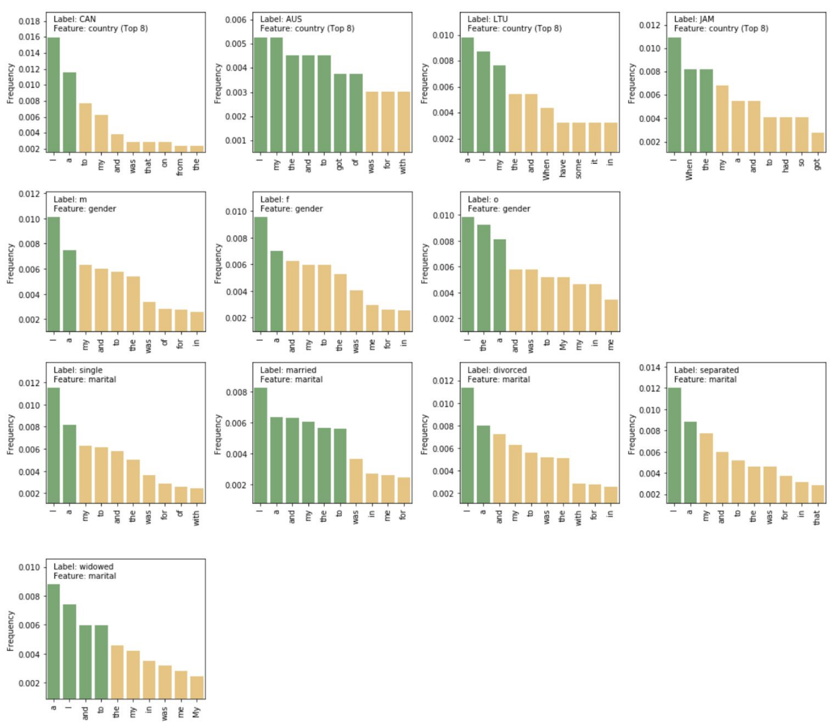
References:

[1] https://www.kaggle.com/ydalat/happydb-what-100-000-happy-moments-are-telling-us
[2] https://www.kaggle.com/powderist/happydb-analysis
[3] https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568
[4] https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/
[5] https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html
[6]https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/

Appendix



Label: achievement
Feature: predicted_category
Label: affection
Feature: predicted_category
Label: bonding
Feature: predicted_category
Label: enjoy_the_moment
Feature: predicted_category
Label: leisure
Feature: predicted_category
Label: nature
Feature: predicted_category
Label: exercise
Feature: predicted_category
Label: USA
Feature: country (Top 8)
Label: IND
Feature: country (Top 8)
Label: GBR
Feature: country (Top 8)
Label: VEN
Feature: country (Top 8)
Label: CAN
Feature: country (Top 8)
Label: AUS
Feature: country (Top 8)
Label: LTU
Feature: country (Top 8)
Label: JAM
Feature: country (Top 8)
Frequency



Label: CAN
Feature: country (Top 8)
Label: AUS
Feature: country (Top 8)
Label: LTU
Feature: country (Top 8)
Label: JAM
Feature: country (Top 8)
Label: m
Feature: gender
Label: f
Feature: gender
Label: o
Feature: gender
Label: single
Feature: marital
Label: married
Feature: marital
Label: divorced
Feature: marital
Label: separated
Feature: marital
Label: widowed
Feature: marital
Frequency

predicted_category	hmid
achievement	33993
affection	34168
bonding	10727
enjoy_the_moment	11144
exercise	1202
leisure	7458
nature	1843

Count of each predicted category

country	hmid
1	79063
0	18236
PHL	279
MEX	150
VNM	126
BRA	123
AUS	117
MKD	104

Count of the country before binary conversion

country	hmid
1.0	79063
0.0	21269

Count for the country variable after binary conversion

parenthood	hmid
n	60937
y	39520

Count of Parenthood

marital	hmid
0	59035
1	41343

Count of marital status

Classification Report - MNB

	precision	recall	f1-score	support
0.0	0.70	0.83	0.76	17623
1.0	0.67	0.49	0.57	12491
accuracy			0.69	30114
macro avg	0.68	0.66	0.66	30114
weighted avg	0.69	0.69	0.68	30114

Confusion Matrix - MNB

	Predicted: 0	Predicted: 1
Actual: 0	14673	2950
Actual: 1	6403	6088

Classification report for Multinomial Naive Bayes for Demographic variables

	precision	recall	f1-score	support
achievement	0.47	0.45	0.46	6905
affection	0.43	0.50	0.46	6768
bonding	0.44	0.28	0.34	2236
enjoy_the_moment	0.16	0.18	0.17	2165
exercise	0.01	0.01	0.01	225
leisure	0.23	0.17	0.19	1461
nature	0.02	0.01	0.01	347
accuracy			0.39	20107
macro avg	0.25	0.23	0.24	20107
weighted avg	0.39	0.39	0.38	20107

Classification report for Multinomial Naive Bayes for predicting the category

	precision	recall	f1-score	support
achievement	0.34	0.92	0.50	6905
affection	0.42	0.09	0.15	6768
bonding	0.09	0.00	0.01	2236
enjoy_the_moment	0.09	0.00	0.00	2165
exercise	0.00	0.00	0.00	225
leisure	0.33	0.01	0.02	1461
nature	0.14	0.01	0.03	347
accuracy			0.35	20107
macro avg	0.20	0.15	0.10	20107
weighted avg	0.30	0.35	0.22	20107

Classification report for Stochastic Gradient Descent for predicting the category

	precision	recall	f1-score	support
achievement	0.36	0.73	0.48	6905
affection	0.39	0.10	0.17	6768
bonding	0.24	0.03	0.05	2236
enjoy_the_moment	0.14	0.16	0.15	2165
exercise	0.00	0.00	0.00	225
leisure	0.09	0.08	0.08	1461
nature	0.04	0.02	0.03	347
accuracy			0.31	20107
macro avg	0.18	0.16	0.14	20107
weighted avg	0.30	0.31	0.25	20107

Classification report for Support Vector Machine for predicting the category

	precision	recall	f1-score	support
achievement	0.34	0.96	0.51	6905
affection	0.38	0.03	0.05	6768
bonding	0.09	0.00	0.01	2236
enjoy_the_moment	0.00	0.00	0.00	2165
exercise	0.00	0.00	0.00	225
leisure	0.24	0.01	0.02	1461
nature	0.13	0.02	0.04	347
accuracy			0.34	20107
macro avg	0.17	0.15	0.09	20107
weighted avg	0.27	0.34	0.19	20107

Classification report for XGBoost

	precision	recall	f1-score	support
achievement	0.34	0.95	0.51	6905
affection	0.36	0.02	0.04	6768
bonding	0.23	0.01	0.01	2236
enjoy_the_moment	0.11	0.02	0.03	2165
exercise	0.00	0.00	0.00	225
leisure	0.30	0.04	0.07	1461
nature	0.00	0.00	0.00	347
accuracy			0.34	20107
macro avg	0.19	0.15	0.09	20107
weighted avg	0.30	0.34	0.20	20107

Classification report for KNN

```
accuracy 0.3176505694534242
                 precision    recall  f1-score   support

     achievement       0.38      0.50      0.43      6905
       affection       0.40      0.29      0.34      6768
         bonding       0.25      0.19      0.22      2236
enjoy_the_moment       0.13      0.17      0.15      2165
        exercise       0.02      0.00      0.01       225
         leisure       0.13      0.09      0.11      1461
          nature       0.02      0.02      0.02       347

        accuracy                           0.32     20107
       macro avg       0.19      0.18      0.18     20107
    weighted avg       0.32      0.32      0.31     20107
```

Classification report for Logistic Regression

```
I made the most delicious meal for my significant  ...
Actual label:affection
Predicted label: affection

I spent time with colleagues at a work conference  ...
Actual label:bonding
Predicted label: bonding

I donated a bunch of old books I had to the local  ...
Actual label:enjoy_the_moment
Predicted label: enjoy_the_moment

I attended the wedding of my cousin. ...
Actual label:affection
Predicted label: affection

I found $50 in my winter jacket ...
Actual label:achievement
Predicted label: achievement

I Went to the Dollar Store earlier and was able to ...
Actual label:achievement
Predicted label: achievement

I got a big lead at work and was recognized for it ...
Actual label:achievement
Predicted label: achievement

Watch a movie of terror at home, eating cotufas an ...
Actual label:leisure
Predicted label: leisure

I got a really nice desert last night and a fidget ...
Actual label:achievement
Predicted label: achievement

last month i went a tour to banglore, and i enjoye ...
Actual label:enjoy_the_moment
Predicted label: enjoy_the_moment
```

Sample for an actual and predicted label for predicted_category from RNN

Future Scope

- Future work may include the usage of this model to other ML pipelines, for better recommendations of items or ideas, based on the provided user inputs and demographics. This can help users to get new ideas/suggestions to spread more happiness!