CF-AFF Shared Task – In Pursuit of Happiness

(Group -17)

Abstract:

In Pursuit of happiness is an interesting field of research where the problem statement is to understand, define & quantify “Happiness” in short, it is an attempt to answer the question, “What makes one happy?” Answering the above question, has several practical and business advantages. For e.g. *“What are the popular sports/movies/books/purchased products/tourist destinations that make people happy?”*. Hence, there has been various attempts in the state of the art to define happiness. The state of the art has both Machine Learning & Deep learning approaches. One of the challenging aspects of modeling happiness is to define the features that help define a happy moment accurately. Is it just words used in the moment? The emotion behind the moment? The intent behind the moment? etc., There are several feature engineering ideas in the state of the art, such as to use emolex, pos tagging etc., Our primary focus in this work is to identify the feature set that will help predict the agency & social aspects of the happy moments with high accuracy. Our secondary yet crucial focus is to come-up with a model to predict social & agency aspects of a happy moment that performs equal or better than the state of the art approaches.

Introduction:

Motivation:

The ***pursuit of happiness*** *is defined as a fundamental right mentioned in the Declaration of Independence to freely pursue joy and live life in a way that makes you happy, as long as you don't* ***do*** *anything illegal or violate the rights of others*. – [Ref: YourDictionary]

Fundamentally, it is important for one to be happy independent of age, gender, profession etc., But the interesting and challenging aspect here are “**What are the causes of Happiness?”, “How do we quantify Happiness?”** etc.,

Let us assume, there is some black box which helps to detect the causes of happiness !! Yet another interesting aspect is **“How can we use this information?”.** Some of the questions that can be answered with the help of the above information is,

1. What are the popular sports/movies/books/purchased products/tourist destinations that make people happy?

2. Can we predict gender/marriage status/parenthood/age groups based on happy moment texts?
3. How many indoor and outdoor activities are in the corpus respectively?
4. Can we find interesting ways of clustering happy moments?

[Ref: https://www.kaggle.com/ritresearch/happydb]

The above questions are some generic questions, however predicting the causes of happiness is also beneficial for quite many fields of business.

For e.g.

1. What makes my employees happy during their working hours in office?
    a. These are the technologies that are making them happy ?
    b. New challenging learning is making them happy?
    c. The subsidized coffee is making them happy?
    d. What is the productivity of those employees that are happy?
2. For e-commerce/travel sites?
    a. Which products/travel locations that are making the customers happy?

To conclude, it is beneficial to predict/detect the causes of happiness!!

Data:

To predict/detect the causes of happiness the first step would be to curate the happy moments along with meta-data/features that help analyze the happy moments. This task has already be done by Megagonlabs and a public HappyDB Corpus has been published for researches to contribute to the field. HappyDB is a corpus of 100,000 crowd-sourced happy moments.

Collection of Happy Moments:

- Each worker has been asked to do the below task:

  > What made you happy today? Reflect on the past 24 hours, and recall
  > three actual events that happened to you that made you happy. Write
  > down your happy moment in a complete sentence.
  > (Write three such moments.)

- Along with each happy moment, the demographic information of the worker who provided the moment is also collected.

Annotating the HappyDB:

Annotators were required to annotate manually each moment along two binary dimensions: Agency and Sociality.

Agency:

Agency means their focus of control, or the degree to which an author in control of their surroundings. It means ‘is the author in control?’

Examples of sentences where the author is in control (Answer is YES):

“*I ran on the treadmill for 20 minutes straight when I could barely do 5 minutes 3 months ago.”*

Examples of sentences where the author is not in control (Answer is NO):

“*A small business deal change over for small profit.”*

Sociality:

Sociality conceptualizes interpersonal engagement, evinced in writing as the description of any activity performed with or in the company of others. It means ‘does this moment involve other people other than the author?’

Examples of sentences which involve other people (Answer is YES):

“*My youngest daughter got accepted to many prestigious universities and accepted an offer to attend college in San Diego.”*

Examples of sentences which are not social (Answer is NO):

“*The bus came on time, so I reached work early.”*

Topic Labeling:

For every happy moment, tags were also annotated which specify potential topics like Family, Food, Entertainment etc., Each moment could have a maximum of four tags if at least two annotators agreed on them.

Problem Statement:

There are two major tasks defined in this work.

Task 1: **WHAT ARE THE INGREDIENTS FOR HAPPINESS?** Predict the agency, social labels for labelled happy moments.

Task 2: **HOW CAN WE MODEL HAPPINESS?** Propose new characterizations and insights for happy moments.

In this project we focused on Task 1.

Data Analysis:

Below are some of the basic analysis we performed on the data.

Distribution of Yes/No labels for Social and Agency

```
------------------ Social -------------------
                 YES         NO       SUM
Agency YES       3554        4242     7796
       NO        2071        693      2764
       SUM       5625        4935     10560
```

Statistics of Length of Happy Moments

```
Unique words in Vocabulary :  7604
----------------- SENTENCE STATISTICS OF MOMENTS ----------------
Minimum Length :  2
Maximum Length :  70
Average Length :  13
Median Length  :  12
```

Statistics of Length of Happy Moments based on Label Yes/NO

Avg length of moment for Agency-YES : 13.0538

Avg length of moment for Agency-NO : 14.5358

Avg length of moment for Social-YES : 14.6746

Avg length of moment for Social-NO : 12.0364

Correlation among Features:

To understand the relation among features we chose the features which are numeric or that can have meaning when label encoded.

The features used and their encodings are:

agency - [0:no, 1:yes]

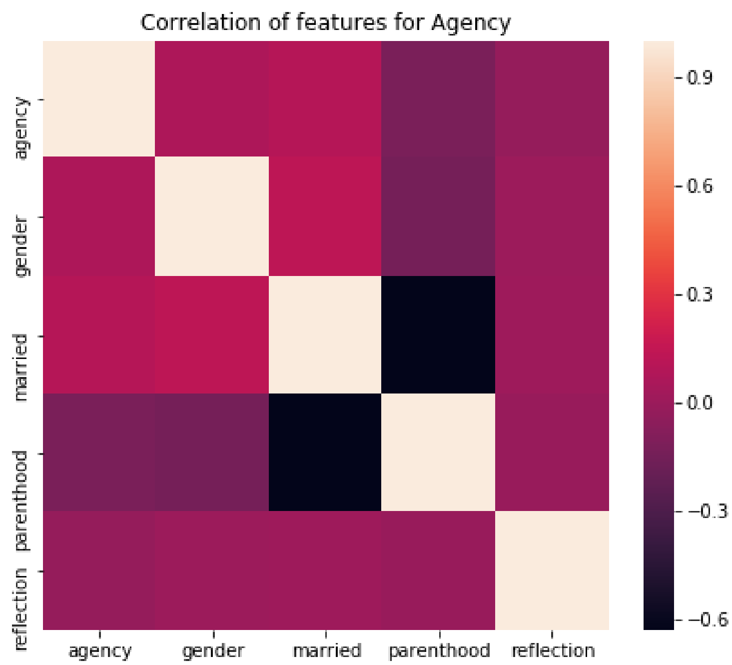
social - [0:no, 1:yes]

gender - [0:f, 1:m, 2:nan, 3:o]

married - [0:divorced, 1:married, 2:nan, 3:separated, 4:single, 5: widowed]

parenthood - [0:n, 1:nan, 2:y]

reflection - [0:24h, 1:3m]



There is no correlation between the features and hence we can drop all the other features and focus only on target labels.

Literature Survey:

CL-Aff Shared Task is a part of the AffCon Workshop @ AAAI 2019 for Modeling Affect-in-Action. 11 teams participated in the task. We discuss the some of the interesting works from the 11 teams, as part of our literature study (for task one alone). We discuss the state of the art as two broad tasks of research, “Feature Engineering” & “Modelling”.

Arizona State University (ASU) proposed a Word Pair Convolutional Model. As per ASU, feature engineering for is motivated by observing the patterns as below in majority of the sentences. For e.g. “I Word Pair Convolutional Model for Happy Moment Classification 3 walked” in “I walked my dog to the park” is critical for understanding the agency of the speaker. Hence the word-pair based feature engineering. Although these can be provided to the NN they preferred the NN to learn the word pairs.

University of California Santa Cruz(UCSC) proposed an approach with focus inclined towards feature engineering. They explored the use of syntactic, emotional & survey features for feature engineering. They used XGBoost Forest & CNN models. Syntactic features correspond to count of occurrence of nouns, verbs, adjectives & adverbs, tense & aspect information. Emotional features correspond to the count of emotions such as anger, negative, sadness, positive, surprise trust etc.,

International Institute of Information Technology Hyderabad, proposed an approach to use inductive transfer learning. They build a pre-trained language model using AWD-LSTM on WikiText. The later step is to use the pre-trained language model to fine tune the model for the classification task.

Gyrfalcon, proposed an interesting approach which is to map the English words into squared glyphs images. These sequence of glyphs act as input to a 2D-CNN model.

A*STAR: proposed an approach which primarily focusses on emotional aspects of the happy moments. The emotional features(namely intensity of emotion) with word embeddings form their feature set. Logistic regression is used for modelling.

University of British Columbia proposed an approach focusing on various word embeddings such as CoVe, ELMo etc., for feature engineering and a BiLSTM both with & without attention.

There are many other works, Fraunhofer try to analyze the demographic aspect. ESPOL proposed semi-supervised adaption of KMeans using NN.

Our Approach:

Baseline Model:

The approach is to predict 'yes' for both social & agency for all the moments independent of its features. For which the accuracy is as below:

Preprocessing	NONE
Features	NONE
Classifier	RULE BASED, PREDICT ALL AS YES
Social Accuracy	53.26%
Agency Accuracy	73.82%

Conclusion & Insights:

Although this is a naïve approach we get a good accuracy for agency. This acts as our baseline accuracy and all the models that we try here forward need to out-beat the accuracies produced by this model.

Machine Learning Models:

Model 1:

We started using the Profile features [Meta-data of a happy moment such as age, category, gender etc., etc.,] without considering the actual moment. This model is an attempt to understand if everything conveyed in the moment has been captured and would it suffice to predict the accuracy of social & agency. If the profile features are sufficient to predict social & agencies then we may get rid of the complexities involved in text understanding & rather focus on generating/capturing the profile features.

Preprocessing	LABEL ENCODING
Features	Profile Features
Classifier	XGBoost
Social Accuracy	61.2%
Agency Accuracy	72.83%

Conclusion & Insights:

Profile features alone are not sufficient to build the model & we need to consider the moments data. However, profile features may act well to support the model built on moments data. We may use, model built on profile features as a complementary model i.e. an ensemble model can be built with profile features model as one of them.

Model 2:

Text (moment) based feature engineering. As an exploratory analysis we initially adopted one of the approach suggested by one of the state of the art papers i.e. to use 4grams to build features. We do not have any intuition behind the same, however did this as an exploratory learning experiment. Below are the results wrt the same.

Model	Preprocessing	Features	Feature Representation	Social	Agency
Naive Bayes	Stopword+ Stem	4-Grams	TFIDF	50.19%	72.54%
Logistic Regression	Stopword+ Stem	4-Grams	TFIDF	50%	72.92%
XGBoost Forest	Stopword+ Stem	4-Grams	TFIDF	58%	74.59%
Neural Network with 2 hidden layers	Stopword+ Stem	4-Grams	TFIDF	49.19%	74.61%

Conclusion & Insights:

We could not deduce anything definite from the experiment. However, we doubt if 4 grams is a good representation of a happy moment as with varied classifiers there is not much difference in the accuracy of the models. Also, they are either less or closer to the baseline accuracies.

Model 3:

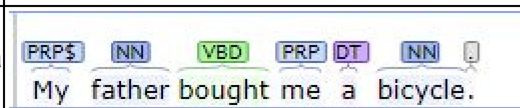
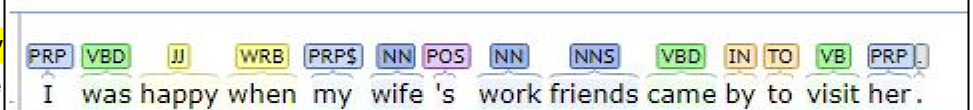
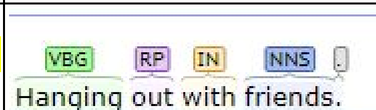
In this attempt we opted for a bottom up approach i.e. to study the data and come up with an approach to define the right set of features.

Social:

Social as defined is, if a happy moment involves more than one person then it is social happy moment. So if there is a way to identify if there is more than one person involved in the happy moment ?

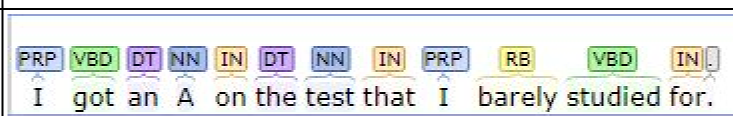
Some patterns that we observed are as below:

SOCIAL = YES

Moment data	POS tag
I was happy when my son got 90% marks in his examination	
My father bought me a bicycle.	
I was happy when my wife's work friends came by to visit her.	
A good dinner in a nice restaurant with coworkers.	
Hanging out with friends.	

Presence of Pronoun followed by JJ*NN or presence of IN followed by NNS has high probability of the sentence being a Social Happy Moment.

SOCIAL = NO:

Moment Data	POS tag
I cooked a new recipe for dinner and it turned out great	
I got an A on the test that I barely studied for.	

Typically, all the non-social moments only involve I or just a statement. With the above observations, we concluded the below:

1. Do not remove stopwords. Or do not include pronouns in the stopword list.
2. Do not stem the words. Friends & friend may impact the prediction.
3. Try building Bag of words with the relevant pos tags.

Agency:

Agency is a very tricky problem to solve. Even manual examination of the data to predict the agency is difficult. However, some of the intuitions that we explored are:

Hypothesis 1:

If agent is not in focus of control, he/she might end up using either too little words to express the moment or too many words to express the moment.

Experiment:

Average length of the moments for both agency yes & no are similar.13.05(Yes) & 14.53(No)

Conclusion: The hypothesis fails.

Hypothesis 2:

If agent is not in focus of control, he/she might end up using either neutral or abnormal sentiments while describing the moment.

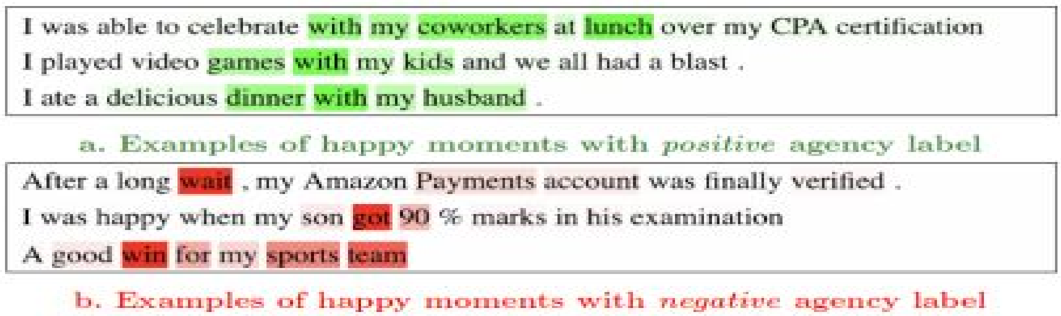
Experiment:

We manually examined a sample of both the classes of agency with Semantria Lexalytics tool (Demo is used) to calculate the sentiment of given sentences.

Agency	Sentence	Sentiment
YES	Went to movies with my friends it was fun	0.550
YES	The day I got my degree in industrial engineering	0.0
NO	My father brought me a bicycle	0.0
NO	My girlfriend told me she is really happy with me and that she was happy she met me	0.68

Conclusion: Hypothesis fails. The distribution of sentiment score for both the classes did not follow any pattern. We can safely conclude that emotion features are not good indicators of agency

Inspired by the self-attention heat maps provided by UBC paper, we observed that verbs play a crucial role in Agency prediction.



a. Examples of happy moments with *positive* agency label

b. Examples of happy moments with *negative* agency label

Model	Preprocessing	Features	Feature Representation	Social	Agency
XGBoost Forest	POS Tagging	Noun Phrases	TFIDF	84%	76.39%
XGBoost Forest	POS Tagging	All except verbs, punct, adjectives & adverbs	Word2Vec	88.9%	NA
XGBoost Forest	POS Tagging	All except adjectives & adverbs	Word2Vec	NA	82,53%

Deep Learning Models:

This classification model works on deep learning classifier, Convolutional Neural Network. The model implementation details are as follows:

Data Preprocessing:

1. Split the sentences into word lists and omit all punctuation marks.Sentences are processed one by one into arrays of words. All punctuations, including comma, period, exclamation mark, question mark and so on, are discarded. A special case is that all abbreviations, like I'm, and we're, remain unchanged. Stop words are also removed.
2. Transformed the sentences into sequences and pad them so that they are of the same length.
3. The agency and social labels were converted to binary vectors for training and validation purpose.

Split of the data set into a 3:1 ratio for training and testing has been done.

The Model has the following layers:

1. *Embedding Layer* : The embedding layer is fed with 1-D moment description, which are then em-bedded into 2-D matrices. The size of the second dimension is 100, which means every word will be transformed into a 100-dimensional vector. For example, for a sentence with a length of 13, we first add 7 zeros in the front to reach the length 20. After passing the embedding layer, the size of the output matrix will be(20, 100). Currently, the values in the embedding layer are randomly initialized. Using pre-trained embedding like GloVe is our next step.
2. *Convolutional Layer* : There is a 1D convolutional layer with kernels that produce 1-D vector after sliding over the 2-D matrix so formed from the embedding layer.
3. *Dropout Layer* with a value 0.2
4. *Max Pooling Layer*

During training mini-batch gradient descent with batch size 32 and Adam optimizer is used with a learning rate of 0.1. The loss function used is binary cross entropy.

The results are as follows:

(Average Accuracy over 10 epochs)

We have tried with different word embeddings for CNN. Pre-trained glove provided better results than the glove trained on the corpus as the corpus is small sized.Below this the comparision of emebeddings:

Model	Embeddings	Social	Agency
CNN	Randomly initialized	88.46%	82.177%
CNN	Pre-Trained GloVe	89%	83.08%
CNN	GloVe trained on Corpus	80.6%	81.065%

GITHUB Repository Link:

https://github.com/DeepthiKarnam00/InPursuitOfHappiness

**Please refer to the individual branches for code.

Conclusions & Future Work:

Feature Engineering:

Summary:

We tried various feature engineering approaches such as:

- N-grams
- Bow
- POS - Noun Chunks
- POS with customized feature engineering
- Customized BOW (as in CNN Model 5)
- Sentiment score as feature.

Conclusions:

- N-grams may not be the right feature engineering for this task.(Our customized features with ngrams did not enhance the performance of the models)
- Emotional features may not add any value towards the accuracy of both the classifications.
- POS tagging with customized selection of words using tags worked the best for us. (Inclusion of ‘verbs’ for Agency classification increased the accuracy by 0.03%)

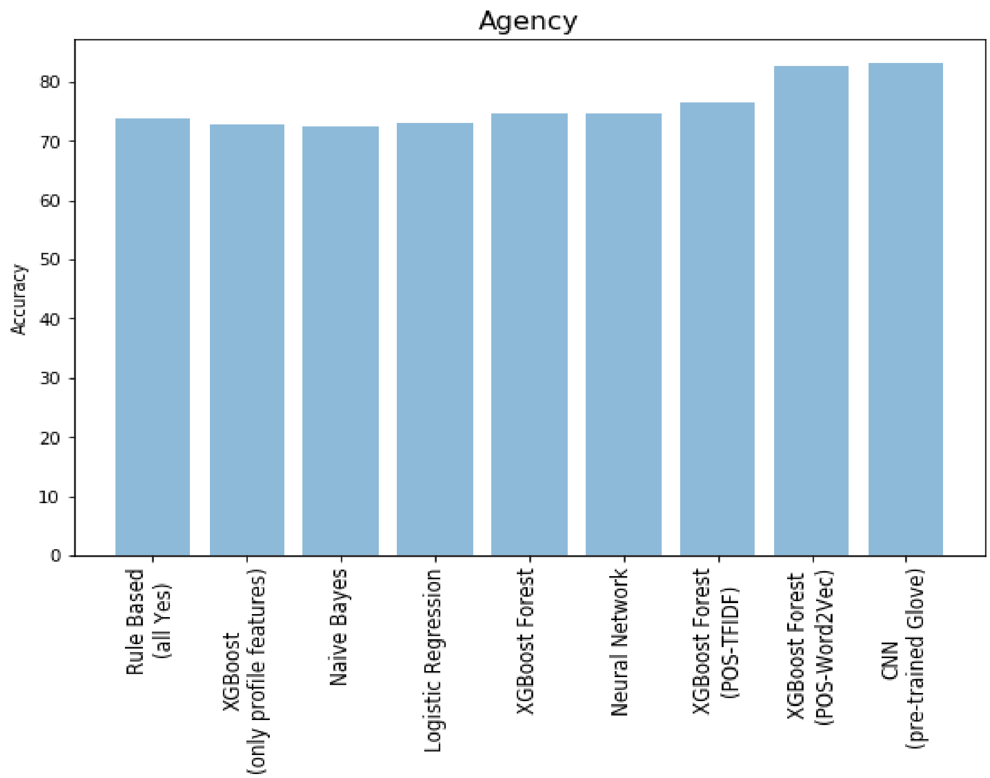
Modeling:

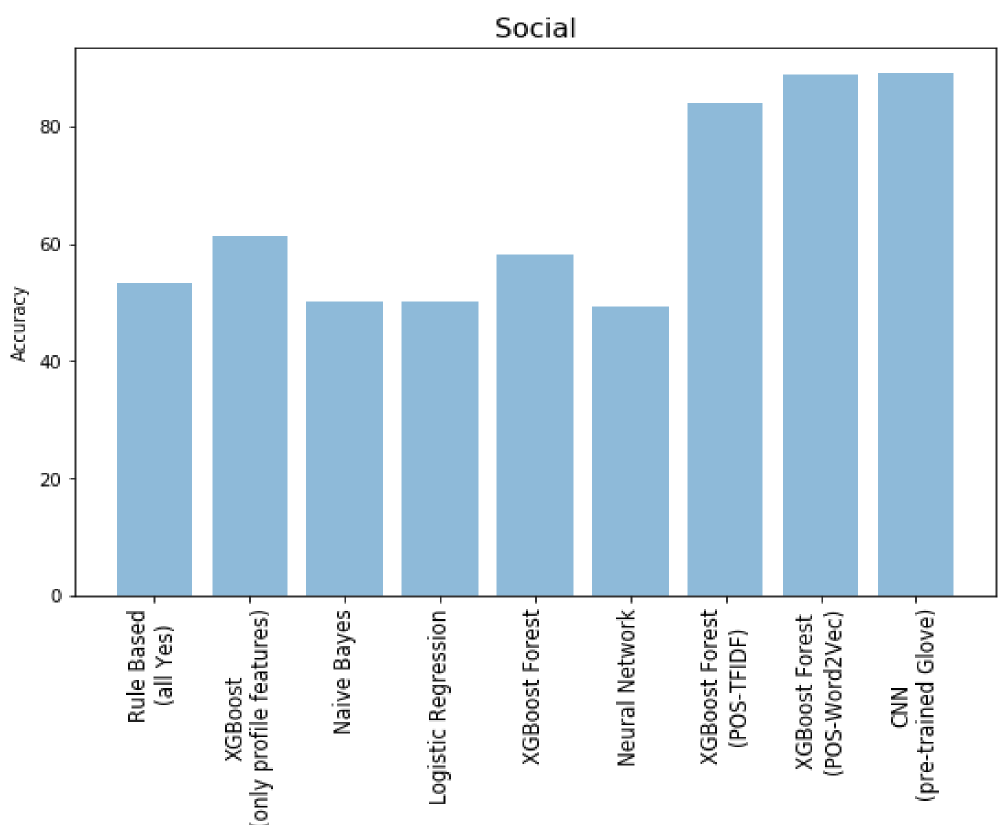
Summary:

We tried various classification models such as:

- Naive Bayes
- Logistic Regression
- XGBoost forest
- CNN

The following diagrams shows the comparison among all the models with different features


Agency
Accuracy
0
10
20
30
40
50
60
70
80
Rule Based (all Yes)
XGBoost (only profile features)
Naive Bayes
Logistic Regression
XGBoost Forest
Neural Network
XGBoost Forest (POS-TFIDF)
XGBoost Forest (POS-Word2Vec)
CNN (pre-trained Glove)


Social
Accuracy
0
20
40
60
80
Rule Based (all Yes)
XGBoost (only profile features)
Naive Bayes
Logistic Regression
XGBoost Forest
Neural Network
XGBoost Forest (POS-TFIDF)
XGBoost Forest (POS-Word2Vec)
CNN (pre-trained Glove)

Conclusions:

- An interesting and obvious observation is using word embeddings for our Machine Learning approach increased the accuracies of the model. Hence, transfer based learning approaches show promise.
- To classify agency, self-attention based models show high promise. [Deduction from the UBC paper]
- Our current best accuracy is for CNN model with pre-trained glove
  - Agency: 83.08%
  - Social: 89%

References:

- The CL-Aff Happiness Shared Task: Results andKey Insights paper by Kokil Jaidka et al.
- CruzAffect at AffCon 2019 Shared Task:A feature-rich approach to characterize happiness paper by Jiaqi Wu et al.
- [CL-Aff Shared Task] Happiness Ingredients Detection using Multi-Task Deep Learning paper by Weizhao Xin and Diana Inkpen.