Analyzing Happiness: Investigation on Happy Moments using a Bag-of-words Approach and Related Ethical Discussions

Riddhiman Adib, **Eyad Aldawod, Nathan Lang, Nina Lasswell, Shion Guha**
Department of Mathematics, Statistics and Computer SCience,
Marquette University, Milwaukee, WI

*Presented at **IEEE COMPSAC 2019***

1

Introduction

- Analysis of moments and activities making people happy, based on a collection of “happy moments”
  - *“I watched a great TV show while petting my cat today”*
  - *“I woke up today”*
- A collection of text responses shared through Amazon Mechanical Turk (MTurk)
  - Crowd-sourcing platform
- Possibility of predicting people's happiness
  - by gathering data from personal experiences and the correlation of the emotion felt
- Through this data analysis and investigation procedure, we have quantitatively and qualitatively studied reasons that make certain group of people happy.
  - Impactful in helping the broader society
  - Impactful for companies

2

Related Works

- Process: Started with previous studies conducted on crowd-sourcing, the behavior of online decision making, and the study of human factors of happiness

- Motivation: Growing issue and importance of mental health [1]

- Platform: Crowd-sourcing, one of the most efficient methods for analyzing online decision making [2]

- Literature: Life events affect people's happiness levels [3]
  - broad spectrum of events, ranging from health, relationships, employment, money and other

- Approaches:
  - Sensor-based [4]
  - Text-based

3

Research Questions

- RQ1: Can we predict the reason of happiness from a person's happy moment?
  - Based on this dataset, are we able to accurately predict happiness category in the past 24 hours or 3 months?
  - What was misclassified? Why?

- RQ2: By looking at specific subgroups *(certain country, age, gender, etc.)*, what are the reasons of happiness in that group?

4

Dataset: Description

- HappyDB [5]:
  - a collection of crowd-sourced happy moments
  - 100,922 happy moments, 10,843 distinct participant
- Tables: `cleaned_hm`, `demographic`
- Cleaned_hm: 100,535 observations and 9 variables
  - Reflection period (24 hrs or 3 months)
  - Happy moment text
  - Ground truth category (Achievement, Affection, Bonding, Enjoying the moment, Exercise, Leisure, Nature)
  - Number of sentences
- Demographic: 10,844 observations and 6 variables
  - Age, country, gender, marital status, parenthood
- Cleaning and preprocessing:
  - Removal of *null*, invalid age, misspellings and wrong texts
  - Equal width binning: Age (17-20, 21-30, 31-40, 41-50, 51-60, 60+)

5

Dataset: Exploratory

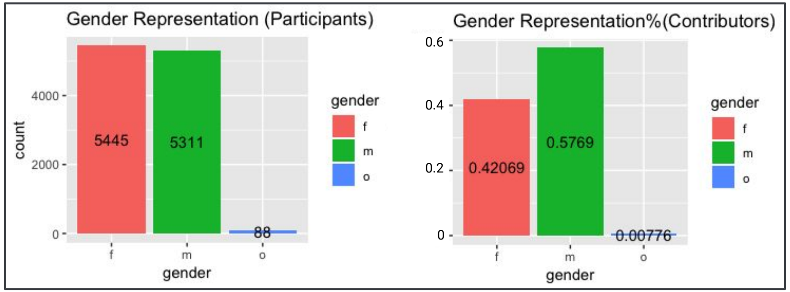


Fig: Gender distribution in participants vs contributors

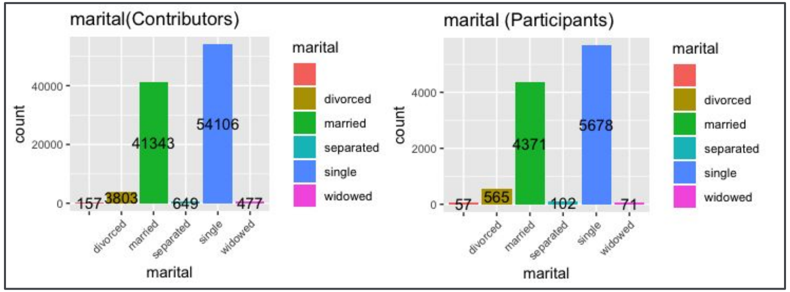


Fig: Marital status distribution in participants vs contributors

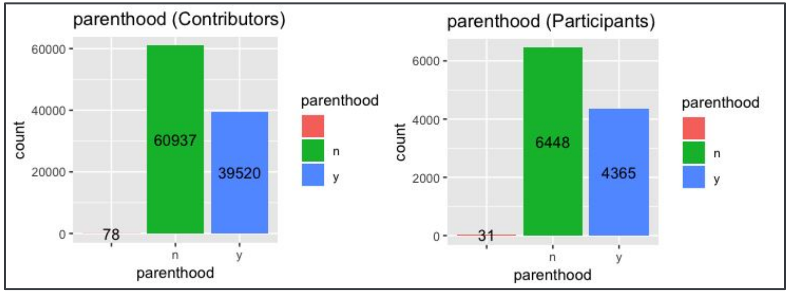


Fig: Parenthood distribution in participants vs contributors

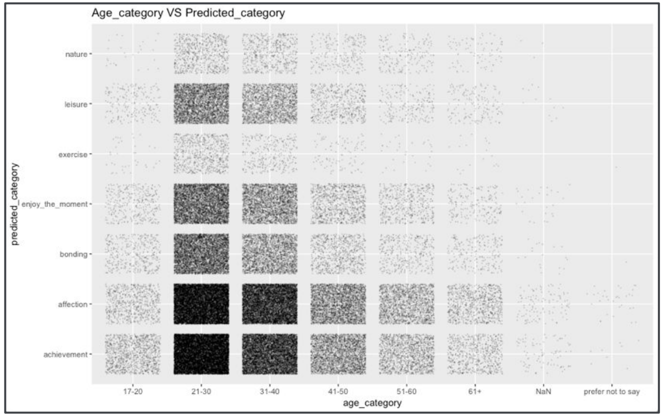


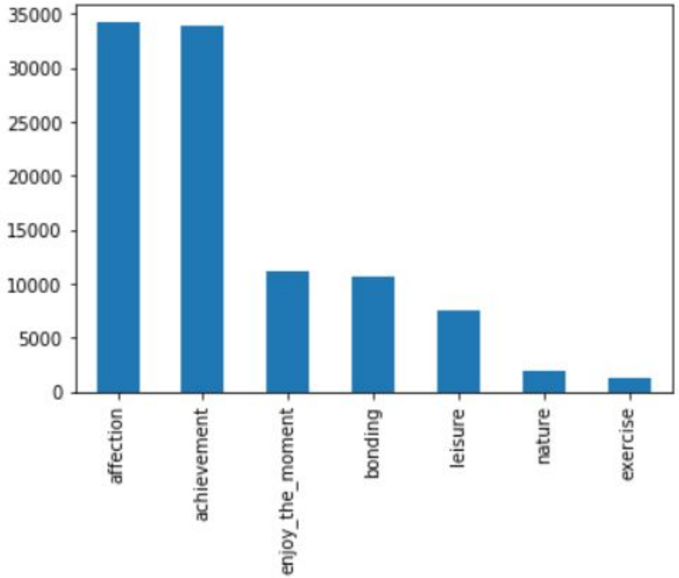
Fig: Age group vs happiness category

6

Data Methods

- About two-thirds of our categories being made up by affection and achievement
- Bag-of-words approach
  - 25,400 columns
  - Count vectorizer → TF-IDF
- Linear Support Vector Classifier
  - 75:25 train-test split
  - Accuracy plateau at around 5,000 features
- We used 6,700 features for our model to obtain best accuracy (~94% on average)

Fig: Count vs ground truth category of happiness



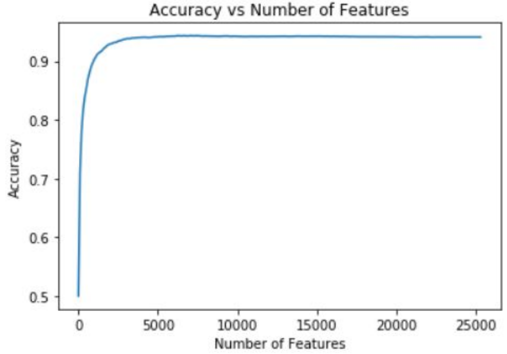


Fig: Accuracy vs number of features

7

Results

Metrics:

Category	Precision	Recall
affection	0.94	0.95
enjoy the moment	0.96	0.98
achievement	0.97	0.95
bonding	0.88	0.86
leisure	0.91	0.87
nature	0.91	0.87
exercise	0.92	0.89

Fig: Precision-recall for each category

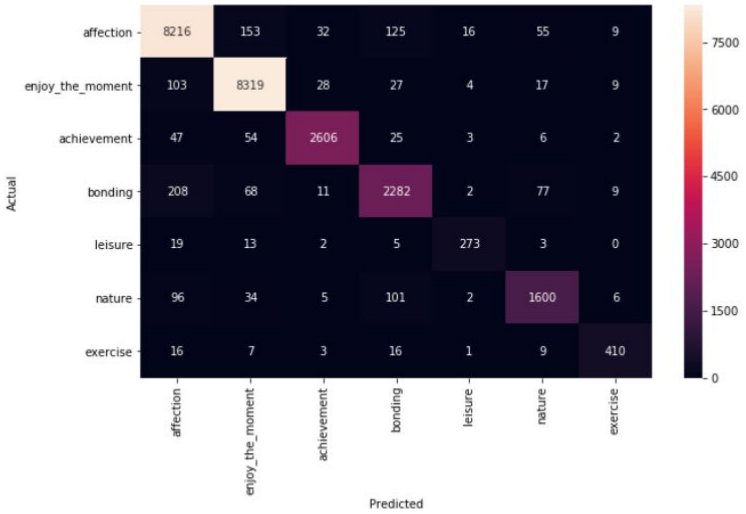


Fig: Confusion matrix - Predicted vs actual categories

Misclassifications and Context:

- *"My students gave me a card"*
  - Predicted: Achievement, Ground truth: Bonding
- *"Ran my fastest 5K ever!"*
  - Predicted: Exercise, Ground truth: Achievement
- Reason 1: Shorter sentences and less meaningful (and common) words used
- Reason 2: Listing multiple moments in the same response

8

Dissecting Happiness: Case Study 1

Married vs Single:

- Group: People who are from *'USA'* and talked about happiness in the category *'enjoy the moment'* in the last *'24 hours'*
- Subgroup: Within this group, people married vs. people single

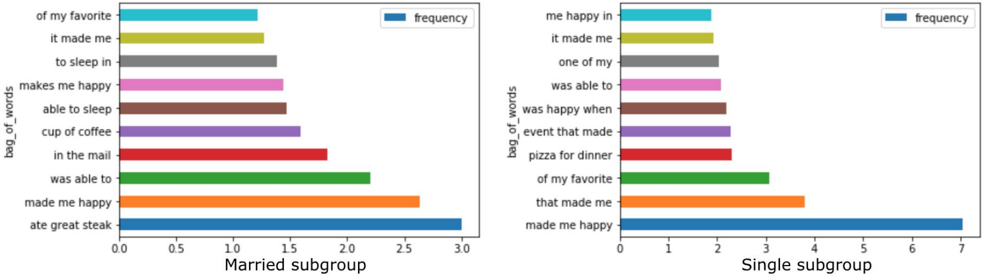


Fig: Top 10 trigrams from each group

Group	Count
Married	220
Single	402

Table: subgroup count

Findings:

- Food: *'ate great steak'* vs *'pizza for dinner'*
- Social contexts: Food priorities, Financial Conditions, Social status
- Example: *'The delicious steak that I had for dinner tonight made me very happy.' vs. 'I ordered two of my favorite pizzas from Pizza Hut and it was cooked just right.'*

9

Dissecting Happiness: Case Study 2

Parents vs. Non-parents:

- Group: People who are from *'USA'* and talked about happiness in the category *'enjoy the moment'* in the last *'24 hours'*
- Subgroup: Within this group, parents vs. non-parents

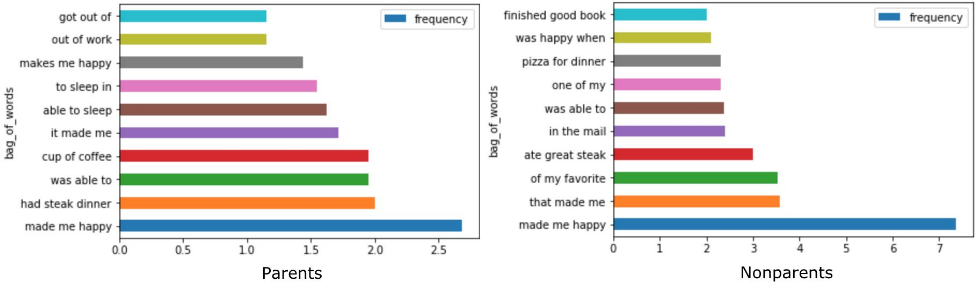


Fig: Top 10 trigrams from each group

Group	Count
Parents	210
Non-parents	447

Table: subgroup count

Findings:

- Sleep: *'able to sleep' , 'to sleep in'*
- Example: *'I got a full night of sleep. That does not often happen with a 3 month-old in the house.'*

10

Ethical Perspective

- Data Collection
  - Informed consent, Collection bias, Limit PII exposure
- Data storage
  - Data security, Right-to-be-forgotten, Data retention plan
- Analysis
  - Missing perspective, Honest representation, Privacy in Analysis, Explainability, Auditability, Fairness across groups
- Deployment
  - Concept drift, Unintended use

Inspired by Deon's Ethics Checklist [6]

11

Summary

- Happiness through ‘happy moments’
- Detection of ‘category’ of happiness
- Reasons for happiness
- Social and ethical implications
- Future work
  - Bigger dataset, more emotions (disgust, anger, sadness)
  - Applied work: sentiment analysis tool
  - Better classification algorithm

12

References

1. Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language processing to mental health. Transactions of the Association for Computational Linguistics, 4, 463-476.

2. Hossain, M. (2012, May). Users' motivation to participate in online crowdsourcing platforms. In 2012 International Conference on Innovation Management and Technology Research (pp. 310-315). IEEE.

3. Clark, A. E., & Oswald, A. J. (2002). A simple statistical method for measuring how life events affect happiness. international Journal of Epidemiology, 31(6), 1139-1144.

4. Budner, P., Eirich, J., & Gloor, P. A. (2017). " Making you happy makes me happy"-Measuring Individual Mood with Smartwatches. arXiv preprint arXiv:1711.06134.

5. Asai, A., Evensen, S., Golshan, B., Halevy, A., Li, V., Lopatenko, A., ... & Xu, Y. (2018). Happydb: A corpus of 100,000 crowdsourced happy moments. arXiv preprint arXiv:1801.07746.

6. An ethics checklist for data scientists. (n.d.). Retrieved from http://deon.drivendata.org/

13

Acknowledgements
QA

- Fellow classmates of Ethics of Data Science Fall ‘ 18 course, Marquette University
- Dr. Shion Guha

THANKS!

Any questions?

You can find me at: **riddhiman.adib@marquette.edu**