Marquette University

e-Publications@Marquette

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

Riddhiman Adib
Marquette University

Eyad Aldawood
Marquette University

Nathan Lang
Marquette University

Nina Lasswell
Marquette University

Shion Guha
Marquette University, shion.guha@marquette.edu

Computer Science Faculty Research and Publications/College of Arts and Sciences

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

Riddhiman Adib
Department of Mathematics, Marquette University, Milwaukee, WI

Eyad Aldawod
Department of Mathematics, Marquette University, Milwaukee, WI

Nathan Lang
Department of Mathematics, Marquette University, Milwaukee, WI

Nina Lasswell
Department of Mathematics, Marquette University, Milwaukee, WI

Shion Guha
Department of Mathematics, Marquette University, Milwaukee, WI

Abstract:

In this research paper, we analyzed what moments and activities make people happy, based on a collection of happy moments. We are focusing on specific happy moments from a collection of text responses that people

have shared through the crowd-sourcing platform: Amazon Mechanical Turk (MTurk). Using crowd-sourcing to collect our data allows us to advance our understanding of the cause of happiness, by focusing on words and real human experiences. Workers of MTurk were asked to reflect on what makes them happy in a given period and share three specific moments in complete sentences. Through text-based analysis, we will look to see what other components have a role in making a specific event happy and further analyze how we can classify such words. Also, we dive deeper into specific subcategories of classifiers in an attempt to form insights about their happiness level based on specific factors. With the goal to extract features from the text in HappyDB, in this study we used the bag of words approach. Through doing so, our results were successful at predicting the happiness category, concerning both accuracy and context. Our models were able to accomplish the goal of understanding a happy moment and fit such a moment into one of the seven ground truth happiness categories we set at the beginning of this study. We finished the article with the ethical perspective of such research works and related social implications.

SECTION I. Introduction

With the increased popularity and efficiency that crowd-sourcing platforms provide, online markets can gain the necessary data needed to increase their competitive advantage in the market [1]. This, coupled with the ever-pressing issue of mental health, it is crucial to have the data and facts on what makes people happy in order to best handle such crises. According to a report, in the United States alone, 18.1% of the population experience a mental illness in a year [2], thus we must understand this aspect of human psychology. By gathering data from personal experiences and the correlation of the emotion felt, it is possible to predict people's happiness based on specific events that have made a user happy in a given reflection period. Using such results allows one to reaffirm the environments that make people happy and drive such events.

Through our research, we are using text-based analysis to predict someone's happiness, based on life events and categories of such. With the higher demand for finding technological solutions to understand human behavior, we have an opportunity to analyze human-generated data in an attempt to find what categories and activities are driving happiness among people the most. Such categories depend on aspects such as personal relationships, work location, family life, health status, and other life events. Also, through a study, it was shown that not only does the life status of a particular person influence their happiness level but also that people who smile more are more likely to be described as happy by others [3].

In our paper, we analyze the HappyDB dataset [4]. This dataset is a collection of over 10,000 happy moments from the crowdsourced platform Amazon Mechanical Turk (MTurk). When this study is conducted, a user is asked the following: What made you happy today? Note, there is a particular reflection period, which would be specified as either 24 hours or 3 months. It is requested that the user completes the task by focusing on happy moments that are not in the distant past, sharing three such moments, and by writing such a happy moment in complete sentences.

Through this data collection, we have text-based data and look to discover if there is a set category of words that are classified as happy, the importance of words provided as responses, and the qualitative study. This analysis is of high importance, since studying what makes on happy not only affects those you interact with daily, but the insights help the broader society and companies, who can use the insight-driven data to meet the needs of consumers by promoting things that directly reflect happiness for the given user. Also, using this relatively new dataset, we have an opportunity to set the bar for this area of study, since not many other studies have been done on this dataset before. With researchers looking for what makes people happy and understanding this space more for many years, we will attempt to provide an answer to this question through data-driven analysis.

Given people's sentiment of happiness from the past three months or 24 hours as prompted we predicted which category of happiness the moment fell under. Using a Linear Support Vector Classifier, we were able to predict the category with a 94% cross-validated accuracy. We used a bag of words method to create our features, meaning each word in the dataset was treated as an individual feature. There were roughly 25,400 words in the dataset, and after optimization, we were left with 6,700 words as features. We then explain why there are misclassifications and often that the happiness moment can either fit into another category different than the ground truth, the moment is too short with not enough meaningful words, and/or somebody incorrectly answered the question with multiple moments. We finish our discussion with a qualitative approach to exploring dataset and future works.

SECTION II. RELATED WORKS

The purpose of our report is to analyze the sources of happiness and look further into the classifications of happiness sources based on the text. To perform this research, we relied on previous studies conducted on crowd-sourcing, the behavior of online decision making, and the study of human factors of happiness. These such areas are crucial to the social science of human psychology and understanding how crowd-sourced events that users have shared that bring them joy. To understand what causes happiness, one can use data from people's real experiences to better target these events and make a lasting impact.

One of the main motivations behind this area of study is the growing issue and importance of mental health. From the research study "Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health", researchers Althoff, Clark, and Leskovec share that, with mental health being one of the most critical health issues, affecting 45.6 million adults every year [2], the main question arises with our limited exposure. Our knowledge surrounding mental health issues and how to have fruitful conversations in this space is extremely limited due to a lack of data regarding happiness and emotions. One study suggests that data, taken from Short Message Service (SMS) texting [2], in the events of depression, self-harm, and suicidal thoughts, allows us to understand how mental health can better be treated through counseling and conversation, similarly as our happy moments' data. Thus, using this data from SMS texting, one can hone in on the skills and strategies necessary to predict the effectivity of counselor conversations and predict the outcome of such conversations. This understanding is crucial to the future of mental health, counseling, and the corresponding area of human psychology, as it allows us to understand the importance of happiness and human emotions. This insight will help us in our study, as it largely plays into the HappyDB dataset we have of the worker's input from the AMT crowd-sourcing platform.

Considering these motivations for research on happiness, we examined studies and research on happiness to analyze the specific events and activities that bring people joy, specifically through crowd-sourcing platforms. Crowd-sourcing is fundamental to explaining what makes one happy because this platform has emerged as one of the most efficient methods for analyzing online decision making. Crowd-sourcing allows people who contribute, known as workers, to capture ideas and provide their input to companies who are looking for future insights and provide feedback [5]. Payment from paid crowd-sourcing platforms is meager. For example, one study identified that the hourly wage for Mechanical Turk was around $1.38 [6]. Considering this, the user's motivation is called into question. In a study done on the crowd-sourced platform Threadless, an open source platform that allows their users to contribute ideas for art and creativity. Results of this study showed that other motivations to participate in crowd-sourcing included the desire to improve their skills, future opportunities on the site, and community involvement [7]. Studies like these show that crowd-sourcing is exceptionally useful. In the Threadless example, we see how a platform can obtain solutions to tasks that are easy for humans but difficult for computers [1]. This understanding helps find solutions that are crucial to today's world to know.

Building off this, for our study on happiness, by asking the workers on the AMT platform repetitive questions about their own experience on happy moment events and activities, the platform in return can re-adjust to better target the individual workers with the best content that applies to them based on their experiences [1]. Also, with this motion of the platform, we can learn what other components are included in an activity that is classified as happy. The main limitation with a crowd-sourcing platform is the issue of diversity among the model; although a good model may be selected, one must consider the variety of each task and worker who is using the crowd-sourcing platform.

Also, it must be called into question how honest people are online; since the physical presence of an individual is not required, it is not verified if one is honest in their responses and classifications. Therefore, in a research study done about the honesty and information provided online [8], an equal number of men and women in college were given a survey about their online usage in a chat-room. Here, it was discovered that people who spend more time online are more comfortable with the platform and therefore more likely to provide honest responses and invest in the platform. In contrast, men are more likely than women to lie, trying to present themselves as intelligent, successful, and a strong figure. We must consider this for our study on happiness, as people could be faulty providing their demographics and sharing their happy moments.

Using the data we got from a crowd-sourcing platform and each worker submitting their individual responses, we must analyze what goes into people's happiness level and the study of social science behind such a level. From the study Simple Statistical Method for Measuring How Life Events Affect Happiness, it is evident that life events affect people's happiness levels [3]. Such events have a broad spectrum, ranging from health, relationships, employment, and money. All such components that correspond to these events must be taken into account, but the weight of all these events may vary. This calls into question how you would measure the size of a particular event in relation to others, with the goal to compare happiness levels in order to get the greatest happiness possible. According to Clark and Oswald from A Simple Statistical Method for Measuring How Life Events Affect Happiness, the issue is that there is "no acceptable way to measure the size of events upon human happiness and psychological health" [3].

Therefore, the quantitative question arises: "what is better or worse." For example, is getting married or going on vacation greater? Is divorce or unemployment worse? Is walking the dog or going on a run better? As shown through research, is not possible to compare such questions, as the weight and human impact vary.

In the study conducted: Are you Happy While You Work?, it was shown that work is the single lowest happiness level of any activity that an individual could engage in, besides being sick [9]. Researchers Bryson and MacKerron share that there is a tradeoff; to have the money needed to spend to do the things one enjoys, which brings happiness, you must work. Thus, with the goal of bringing about happiness, which is better to do; having a job to bring home income or participating in leisure activities? We are not sure about the answer, but the data presented has allowed us to examine this question, and others similar to it, which we will address later.

Another critical component to our research of happiness through text-based analysis is the use of effective words within a sample of words. Through a study done conducted on individual words and their impact on the more significant outcome trying to come across in the greater sentence, paragraph, or writing sample, there are four different categories of words based on the emotion of the word [10]. These types are categorized as the following: valence, the degree of emotion evoked by a work; arousal, the intensity of feeling; dominance, the degree to which the word provides some emotion that leaves the reader feeling dominant; and finally, part-of-speech. Thus, depending on the event and the emotion evoked, each category of words will have higher or lower overall totals of these words depending on the impact to the entire body of text. This is because each event may leave people experiencing different feelings of emotion, which results in a mixed-use of words. Within this study [10], it was observed that crime would have the lowest level of valence, thrill-seeking events

present high valence, whereas having a relaxing vacation has a low level of arousal [10] because of vastly different activities and the words used to describe such emotions.

Different approaches have been done for the research of happiness and causes of it [11], [12]. One study collected body sensor data via smartwatches [13] to detect what makes someone happy. Different algorithmic models [14] have been applied for fitting similar datasets. Regarding our particular study, this calls into question how workers respond to the crowd-sourced questions. Therefore, through this research, we can formulate the following research questions for our paper:

1. Can we create a useful list of activities that are happy? This could be helpful for recommending additional actions to the user.

2. Can we reliably and consistently remove extraneous words? For example, if someone says: "I am happy because my friend is pregnant" versus "My friend is pregnant," can we effectively get rid of these words? For example, "I got to spend time with my son" vs. "I spent time with my son." calls into question expectation. This question focuses on the importance of words.

SECTION III. Research Questions

The dataset for our investigation in this research work is the HappyDB database. Since this database is very relatively newer and not many research work has been done on this, it is an excellent candidate for further explorations and thorough analysis. For this investigation, our research questions are:

1. (RQ1) Can we predict based on this data what makes someone happy?
   a) What are we able to predict based on semantic of what made a person happy in the past 24 hours or 3 months?
   b) What was misclassified?
2. (RQ2) By looking at specific subgroups (certain country, age, gender, etc.), which specific subset is happy about other subsets? Or in general happy?

SECTION IV. HappyDB Database

A. Description

HappyDB is a collection of 100,922 crowd-sourced happy moments, with the goal to understand what makes people happy from words. In this dataset, we worked with two tables: "cleaned_hm" and "demographic" [4]. The "cleaned_hm" file contains 100,535 observations and 9 variables, containing the cleaned version of happy moments crowd-sourced from the users. The variables in this file are:

1. happy moment's ID
2. participant (MTurk worker) ID
3. reflection_period (24 hours or 3 months)
4. original_hm (the user's original happy moment entry)
5. cleaned_hm (the user's edited happy moment entry (spelling check, punctuation fix etc.)
6. modified (boolean flag for edited)
7. num_sentance (the number of sentences)
8. ground_truth_category (ground truth category label, null if missing)
9. predicted_category (happiness category label; one of seven categories predicted by the classifier).

The "demographic" file contains 10,844 observations and 6 variables, and contains the demographic information of the participants. The variables in this file are:

1. participant (MTurk worker) ID
2. age
3. country
4. gender (male, female or other)
5. marital (married or single)
6. parenthood (yes or no)

B. Cleaning process

The aim of this section is to go over the cleaning and preparation processes in order to discuss what have been done to the data before using it. Primarily, the following variables were omitted from clean_hm file: original_hm, modified and num_sentence. These variables were omitted because they do not serve the focus of this paper. Second, changes were made to the demographic file since there were multiple issues that needed to be resolved in order to make use of these variables properly. Age variable had multiple incorrect data-entry, for example, invalid age ("227", "233") or random strings ("ca"). If the variable had meaningful text (like, "60yrs"), it was fixed, otherwise set to null. Based on the age value, we categorized the variable, with age groups of 17-20, 21-30, 31-40, 41-50, 51-60 and 60+.

The two tables were then merged based on the participant (MTurk worker) id. This ended up in a table containing 100535 observations and 14 variables.

C. Exploratory Data Analysis

In order to develop general understanding of the dataset, it is essential to perform basic analysis of the dataset. This general analysis will help us gain an understanding of the dataset from a broader scope, which will help with the analysis process later on. It is important to note that about 86% of the dataset came from the United States, 8.9% from India, and the remaining from other countries [15]. There are 10,841 unique users in the dataset, which suggests that multiple entries were generated from each or most of the users since the dataset contains 100,535 observations.

D. Contributors vs. participants

Since participants generate multiple observations, a comparison between variables from the two files is needed in order to gain a more in-depth insight about the dataset and the behavior of the variables. Upon examining gender representation in the dataset, it seems that males are responsible for generating about 57% of the observations (57,690 obs), females are responsible for generating 42% of the observations (42069 obs), and 0.0077% observations were generated from 'other' or null categories (776 obs).

The given figure 1 provide a visualization of the gender representation in the dataset:
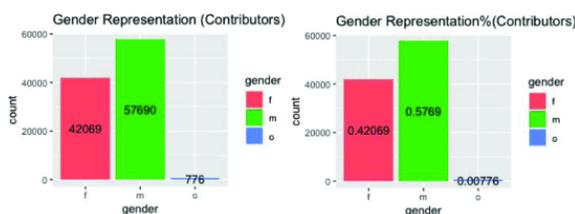

Fig. 1. Gender representation

It is important to mention that examining gender representation in the "demographic" file, 54% (5,445) of the participants are females, 53% (5311) of the participants are male, and 0.009% (88) of the participants are classified as "other" or null.
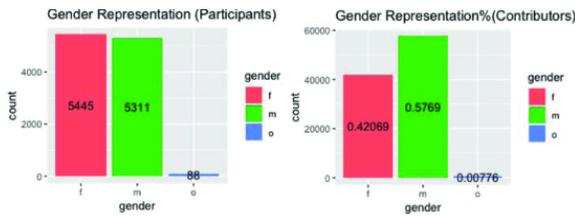
Fig. 2. Gender distribution in participants vs contributors

Thus, while the number of female participants is more or less equal (slightly higher by 134 participants) than the number of male participants, male participants contributed more to the dataset than female participants. Males' contribution average is about 11 observations, females' contribution average is about 8, and others' contribution average is about 9 observations.

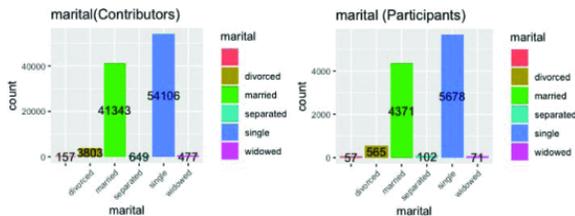As for the martial variable, it seem consistent in both cases, the following graphs illustrate that clearly.

Fig. 3. Marital status distribution in participants vs contributors

Parenthood variable have the same situation with the marital variable as there are no noticeable changes in both cases, the following graphs demonstrate that clearly.

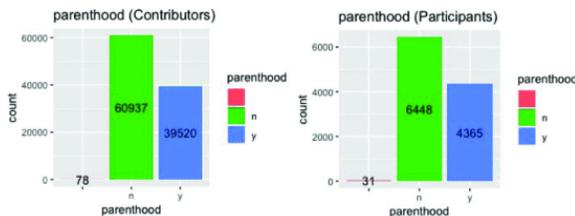Fig. 4. Parenthood distribution in participants vs contributors

Finally, the following jitter plot provides a general insight about the dataset by comparing the "age_category" variable and the "predicted_category". This plot is highly insightful in the sense that, it illustrates the representation of each age group and the predicted happy moments category.
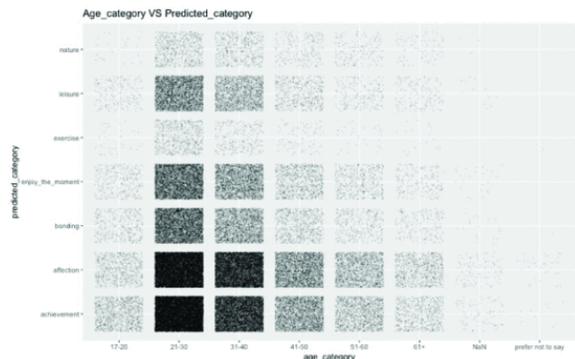

Fig. 5. Age category vs predicted category

SECTION V. Data Methods

A. Dataset

In this section, we will focus on our first research questions (RQ1). As previously mentioned, our dataset has people's sentiment of happiness from the past three months or 24 hours as prompted. There is a predicted category for each happy moment. There are 7 different categories: affection, achievement, enjoying the moment, bonding, leisure, nature, and exercise. These categories come from the original research paper [15] as they can be quoted saying "We publish our crowd-sourced labels as part of HappyDB to provide a ground-truth for researchers". This allowed us to use a predictive model to determine somebody's happiness strictly from the response they gave. Some categories are similar but have slight differences. For instance, the difference between affection and bonding is spending time with family/loved ones/pets and friends/colleagues, respectively.
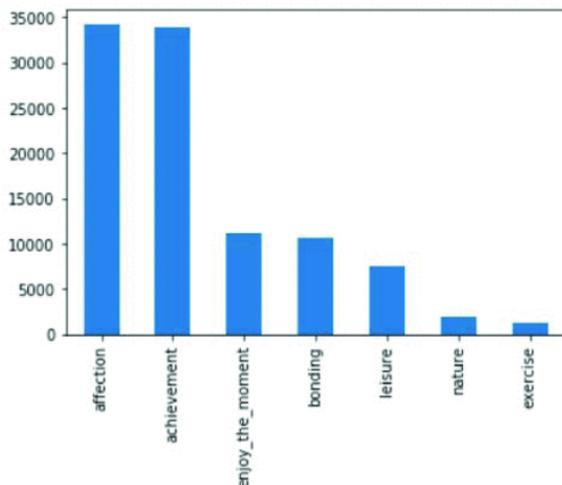

Fig. 6. Count vs ground truth category of happiness

The focused dataset of the model used is strictly the sentiment of happy moments and their happiness category. The happiness category distribution can be seen in the bar graph in figure 6. Affection and achievement make up two-thirds of our happiness category. In the context of happiness, this makes sense. Feeling loved and appreciated will make an average person happy as well as accomplishing something.

B. Model

Due to our dataset consisting of only text, many numerical classifiers will perform very poorly without any prior manipulation. To extract features from our text, we use the bag of words approach. Instead of having one paragraph in each column, our dataset was now represented by having roughly 25,400 columns with one column representing one word in our dataset. This was the basis for us to use a Linear Support Vector Classifier. This is a multiclass classifier based on a one-vs-one scheme. Furthering our dataset preprocess, we used a count vectorizer with each response to be represented by the count of every word in its corresponding column. The count vectorizer takes in a parameter of maximum features which allows us to limit the number of columns or words in our case that we want represented in our new dataset. This new dataset was further preprocessed by going through a TF_IDF transformer, or a term frequency-inverse document frequency which is a numerical statistic reflecting the importance of a word.
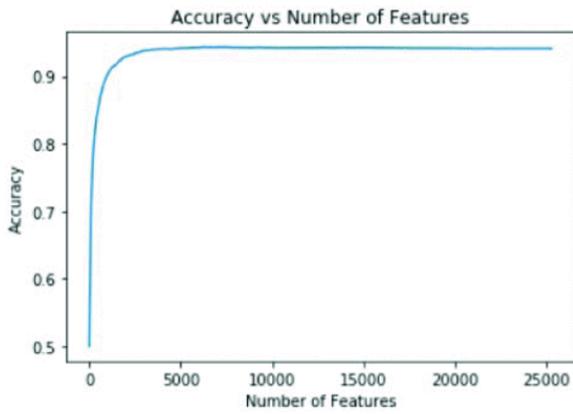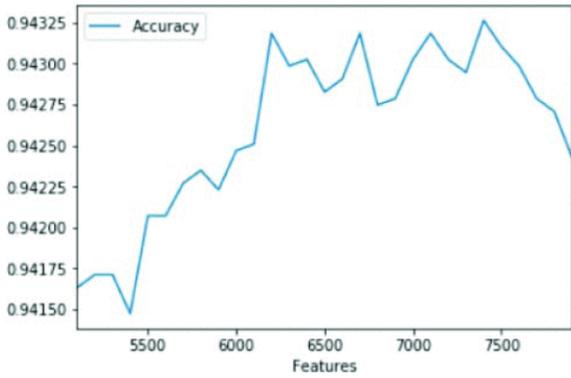
Fig. 7. Accuracy vs number of features

Fig. 8. Accuracy vs number of features (5000 to 8000)

By varying the number of maximum features from 1 to 25,400 by one hundred and training our model on a 75/25 train and test split, we can plot the accuracy vs. the number of features as seen in figure 7. Furthermore, at roughly 5,000 features we see our accuracy plateau. Our figure 8 shows the range of 5,000 to 8,000 closer, and it can be seen that the accuracy varies by roughly .002.

SECTION VI. Results

A. Metrics

By using 6,700 features, our model obtained the best accuracy. With our ideal number of features, we performed a 5-fold cross validation with our Linear SVC, as previously mentioned, resulted in the array [0.9405271 0.94107409 0.94221205 0.93966375 0.93389046]. Our model is precise and accurate around the 94% mark. Further classification results can be seen in table 9.

	precision	recall	f1-score	support
affection	0.94	0.95	0.95	8606
enjoy_the_moment	0.96	0.98	0.97	8507
achievement	0.97	0.95	0.96	2743
bonding	0.88	0.86	0.87	2657
leisure	0.91	0.87	0.89	315
nature	0.91	0.87	0.89	1844
exercise	0.92	0.89	0.90	462
avg / total	0.94	0.94	0.94	25134

Fig. 9. Accuracy table

More important than our accuracy and statistic results overall, is in what context was our predicted happiness category successful. It turns out, in many categories as our accuracy implied and as seen in our heat map confusion matrix in figure 10.
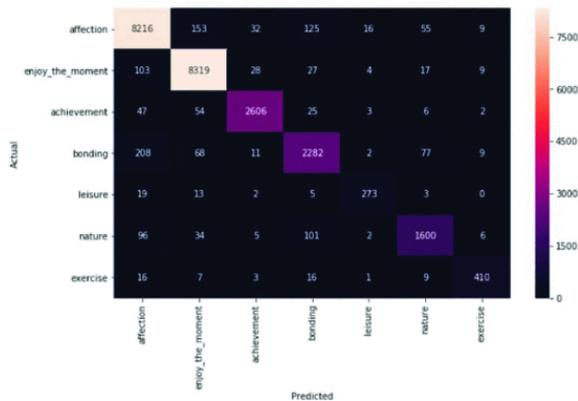

Fig. 10. Confusion matrix of predicted categories with original categories

One primary concern we had was that because of two-thirds of our categories being made up by affection and achievement that there would be much bias towards these classes and our accuracy for the less represented classes were done poorly.

B. Misclassification and Context

We have looked at some of our misclassifies to make sense of our model in context:

"My students gave me a card"

It was misclassified as an achievement rather than bonding. There are two key takeaways from this statement. First, due to the short length of the response, the emphasis of weights will be placed on the words "students," "gave" and "card." The other words will be with minimal to zero weights as the TF-IDF transformer will correct them. The phrase "gave me" can be interpreted as the person receiving something which is where the model probably inferred it was an achievement. For instance,

"My work colleague gave me the news that I was promoted to a higher rank."

Of course, this sentence also includes "promoted" which is a direct achievement, but this is an instance in which "gave me" is used differently. Another few examples of misclassifications are described below:

1. "I took a shower in warm water."
   a) Ground truth: Leisure
   b) Predicted category: Enjoy the Moment
2. "Ran my fastest 5K ever!"
   a) Ground truth: Achievement
   b) Predicted category: Exercise
3. "I received a massage from a professional which alleviated a lot of the pain I was experiencing."
   a) Ground truth: Enjoy the Moment
   b) Predicted category: Achievement

So, the most significant inference made from our misclassified sentiments is that the shorter the sentence and the less meaningful words used, the more likely the response is misclassified. Of course, this makes sense

because in a broader classification ideology, the less meaningful information one has, the harder it is to predict the classification. Furthermore, in the examples above (reference table) the predicted category is not necessarily wrong in context, but rather wrong in the ground truth. Showering is an in the moment happiness, running a 5k is exercise, and alleviating pain is an achievement. Therefore, we can conclude that our model is rather good at understanding the responses given.

There is also another reason a response was misclassified. Simply put, the person responding listed multiple moments in the same response and did not follow the original directions. Examples of this are:

"Being told you made someone's day. Being able to rub your eyes because you're not wearing mascara. Exploding with laughter after holding it in."

and

"my first day office experience party with new friends my first month salary".

Both were labeled as "achievement" but predicted as "enjoy the moment" and "bonding", respectively. The first example was most likely predicted as enjoying the moment because of the term "laughter" and the second example as bonding due to the term "friends" which is part of that category's definition in this dataset.

Overall, our Linear SVC model was able to accomplish the understanding of happy moments and fit it into one of our seven ground truth happiness categories.

SECTION VII. Dissecting happiness

In this section, we will be discussing our second research question (RQ2) of this paper, looking for general reasons on what makes a specific subgroup happy, and how do they compare with each other. We will do so by analyzing a few case studies and discussing a few happy moments out of those cases.

In this stage, it would be important to discuss the features and labels that we currently have at our hands. Although our database is separated into multiple separated files, during the cleaning process we merged them and had come up with a single united file. This new joined table contains features and labels as follows:

Features	Labels
Age group	Ground truth category
Country	Unigrams
Gender	Bigrams
Marital status	Trigrams
Parenthood	
Reflection period	

By specifying different values for the features, we can focus on a specific subgroup and then using unigram, bigram or trigrams, and we can explore further what affects one single group's happiness.

A. Case study 1: Married vs. Single

In our first subgroup analysis, we have tried to compare the reasons for the happiness of a group of people who are married versus those who are single. Specifically, we have looked into people who are from 'USA' and talked about happiness in the category 'enjoy the moment' in the last '24 hours'. Within this group, we have looked in people who are married vs. who are single. In our database, we have:

Subgroup	Count of happy moments

Married	220
Single	402

Now, we looked into the top 10 trigrams of these reviews. We used trigrams since unigrams mostly contain of the words 'happy,' 'made,' 'evening' etc., which does not convey enough information for us to draw conclusions. Bigrams are somewhat good, but trigrams were the most effective in this scenario. We have represented the top 10 trigrams of each group in figure 11.
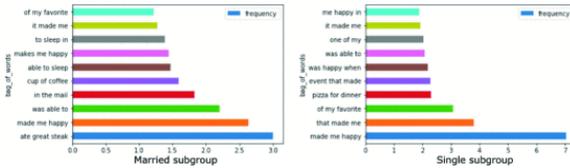

Fig. 11. Married vs single group trigram comparison

From here, we kind of dive into the trigrams and discuss what they mean. Most of the bag of words are pretty straightforward and meaningful, like 'makes me happy,' 'made me happy,' 'was happy when' etc. These are all expected. However, if we look into the top-most bag of words in the married group, we see that it talks about food. It is understandable since food is one of the core things that we do have daily and is more likely to make us happy in a time-span of 24 hours.

Following that, when talking about food, people in the single group had only one item mentioned, which is 'pizza for dinner.' This is an interesting finding. People in the married group is happier when they had 'steak,' compared to that of the single group being happy with 'pizza.' It is indicative of multiple points and can be a result of different reasons:

1. **Food priorities:** Singles are more likely to skip cooking a meal and have a pizza instead.
2. **Financial conditions:** Steak costs fundamentally more than pizza. Married people have a family to maintain, so they are more likely to have a fixed source of income, and thus can afford expensive food.
3. **Social status:** Married people might be done with their 'pizza' days, and might just have moved onto costlier foods.

Now we will have a quick look at some of the reviews discussing these foods:

- "The delicious **steak** that I had for dinner tonight made me very happy."
- "I ate a **steak**, green beans, and corn on the cob for supper."
- "I ordered two of my favorite pizzas from **Pizza** Hut and it was cooked just right."
- "I made myself **pizza** for dinner and sampled some home brewed beer."

B. Case study 2: Parents vs. Non-parents

Our second case study involves people who are parents versus who are nonparents (people who do not have any child). To be specific, our target group in this scenario is people who are from 'USA' and talked about happiness in the category 'enjoy the moment' in the last '24 hours'. Within this group, we have looked in people who are parents vs. who are non-parents. In our database, we have:

Subgroup	Count of happy moments
Parent	210
Nonparent	447

Again, we looked into the top 10 trigrams to see what make these people happy, and we got the list in figure 12.
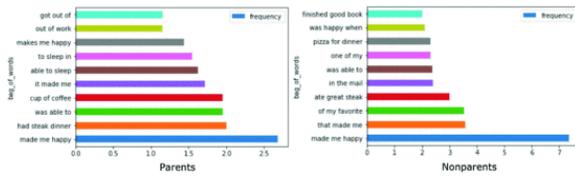

Fig. 12. Parent vs nonparent group trigram comparison

Here, one thing that needs to be discussed is that nonparents are a combination of multiple subgroups. They potentially have people who are single, who have married but no children, who are of older age etc. This can also be seen in the bag of words of the nonparents segment. We can see a variety of topics including reading 'good book,' going to 'mall,' having 'pizza,' 'steak' etc.

The parent group, on the other hand, has something in common, they have children. Their bag of words also shows similar, diverse topics, but we can see 'able to sleep' and 'to sleep in' in the 6th and 7th bag of words. Both of these have the word 'sleep' in it. This discovery is also an interesting finding, and the potential reason could be parents desire for sleep (or rest) a lot more than other subgroups since they are mostly busy taking care of their children and families. When they get a little bit of time to rest, that substantially makes them happier.

Here are a few reviews from this particular subgroup that talks about sleep:

- "I got a full night of **sleep.** That does not often happen with a 3 month-old in the house."
- "I got to **sleep** in this morning and it made me feel rested and happy."
- "It was Saturday morning so I got to **sleep** in a little bit!"

The codes implemented in this project are influenced by the article "Multi-Class Text Classification with Scikit-Learn" [16] by Susan Li.

Similar to these, we can forward these to multiple case studies, and that would be giving us more and more insight into people's mind and what specifically makes them happy.

SECTION VIII. Ethics Checklist

In this section, we will be discussing ethical perspectives of our analysis along with scrutinizing the generation and maintenance of HappyDB dataset. The argument for doing so is that, without properly analyzing ethical perspectives of the implications of our understanding of a data-based research, we cannot be prepared for the multifaceted outcomes of our application in the society, as well as make necessary amends if required. The questions are based on an ethics checklist for data scientists, Deon checklist [17]. The project is open-sourced, and the mentioned list has organized the most commonly asked questions in 5 broad categories.

A. Data Collection

1) Informed consent
The specific dataset that we are dealing with, has been collected as a survey from users working in the Amazon Mechanical Turk (MTurk) program [18]. Part of the agreement as submitting the survey responses were giving consent to the data submitted, and it is expected that the users clearly understood what they were consenting to, this is also explicitly mentioned in Amazon Mechanical Turk's privacy notice [19].

2) Collection bias

The survey was collected via MTurk, so there is a known source of bias induced in the survey. The dataset that we have is comprising of people who are technically sound, smart enough to surf the internet and is in need of cheap and fast money. There were no specific steps taken to mitigate this issue since this was an easier way to collect this amount of information over the world within a short time-frame.

3) Limit PII exposure

The dataset itself does not contain any personal identification, email or anything related to the MTurk workers. However, it is not possible for MTurk workers to completely hide their identities. The survey itself did not ask for workers name or id, but since MTurk manages this, the surveys are stored along with their worker identifiers. So, technically, the survey is not fully anonymized, but, by hiding workers' name and other information, it limits PII exposure to the people who created and managed the surveys. There were other ways to limit PII exposure. One could have posted in MTurk with a link to an external survey. Once the workers have completed the external survey, they could be given a completion code, which can be entered into the post in MTurk. This wouldv'e stopped MTurk tracking the worker and their corresponding response in the survey with the cost of being a little more complicated.

B. Data Storage

1) Data security

The dataset, along with its anonymity of users, is stored in Github and made open-sourced [4]. Github, as an open-sourced git based platform, maintains the sanctity of the dataset and lets people access the dataset as needed. It also keeps track of the people forking the dataset; however, it does not know when the dataset is downloaded and distributed separately.

2) Right to be forgotten

MTurk workers, as part of their agreement, submitted their response fully knowing that they cannot edit or delete the response after submission. We do not expect there was any provision for removing personal information or submitted responses from the dataset upon request.

3) Data retention plan

The dataset maintainers can quickly close down the GitHub repository and related website if needed. Although a cancellation and deletion of the dataset can be executed, it is not expected, since the researchers want this dataset to be promoted and reach to the maximum number of researchers possible.

C. Analysis

1) Missing perspectives

Our analysis in this research work is twofold: quantitative (looking to get maximum accuracy of predicting happiness category based on the submitted text) and qualitative (looking at texts of subgroups to find reasons that make people happy). We have not discussed our analysis and blind spots with affected communities, although we discussed the planned approaches and potential outcomes in between the team members. Also, we presented our research work in related theory class and collected informative feedback from fellow researchers. The feedback helped us rethink some of our approaches.

2) Dataset bias

The survey design was done keeping this issue in kind. In general, to reduce the maximum amount of bias, the researchers tried to keep a more or less equal number of male and female workers. The survey also has happy moments with an equal number of reflection periods (twenty-four hours and three months). However, the dataset is still biased based on age (most workers are of the age group 20 to 40), country (majority from the

USA) and marital status (majority are single) [15]. Other than these issues, there are no other significant sources of bias in this dataset. Moreover, since we have collected minimal information, there is a greater possibility of missing confounding variables.

3) Honest representation
The visualizations shown in the article are honest and correct. One might always argue that visual representations could be made better or more meaningful. The visual representations in this research paper are interpretive enough to make our works and approach understandable.

4) Privacy in analysis
The dataset does not have any specific personally identifiable information, instead has information collectively as a group ("married," "30 years", "USA" etc.). So, PII is neither used nor displayed in our analysis.

5) Auditability
Our analysis is done in two programming languages, Python and R. Both of the scripts are stored in Github and is readily available to the public [20]. Thus, the entire work is reproducible. It is also well documented, in the sense that, the python codes and R scripts contains necessary comments along with an explainable line of thoughts.

D. Modeling

1) Proxy discrimination
We have not specifically looked into any causal inference models and checked for any confounding variables that might have hurt our analysis. Given more time and effort into the research problem, we can explore it and potentially extend it as future work.

2) Fairness across groups
Fairness across groups seems an excellent point to explore. However, we have not tested our model results concerning different affected groups.

3) Metric selection
We have tried to compare our analysis and output metrics with those of the primary article. We have yet to consider other factors that could optimize for our defined metrics.

4) Explainability
We can explain and justify our selection of model, features, and labels up to a certain level. The machine learning models used in this analysis are not too complicated; thus can be explained to mass people.

E. Deployment
This following section discusses the deployment of our developed model in real-life and relevant ethical questions. Now as we did not deploy our analysis for public usage, we will be simulating the scenario where our assumed model is deployed in a server system. We will assume a system with functionalities, where people can upload a text block containing their happy moments, and our system will effectively detect the category of their happiness.

1) Redress
A software project can always go in in multiple directions. It can effectively be harmful to an individual or a community. For this reason, we should build a contingency plan that can be followed in case of similar issues. The plans could be (1) pulling off the server with the software, (2) releasing a press report, (3) building an investigation team that can look into the issue at hand and (4) taking effective and immediate decision to prevent further loss.

2) Roll back

In general, if the system is a web-based system, we can always turn off the server and close access to it. However, the issue is a serious one when we roll out the system in a mobile client (Android or iOS applications). There is no possible way of rolling back or uninstall a smartphone app remotely. Nevertheless, one can turn off the server or keep a flag at the server in order to terminate the connection at any moment.

3) Concept drift

Resistance to concept drift can be achieved by regularly monitoring user submitted response in the applications, and updating the model over time based on the newly submitted texts.

4) Unintended use

Whenever a software system is deployed, there are certain ways of tracking user activity on that system, using analytic platforms, like Google analytics, Fabric analytics, and other related services. We can incorporate those to identify and prevent unwanted uses of our system.

SECTION IX. Discussions

In this section, we have discussed our general takeaways from the research work, the lessons we learned, the shortcomings that we found and ways to extend this work.

A. Defining happiness

One of the controversies of doing research work in this dataset was our approach to explore happy moments and its relation to happiness. We, in no way, have claimed to define 'happiness.' Our goal is to look into reasons that make people happy and use a broader view of categories to categorize them. We have tried to dissect specific reasons for happiness and applied known machine learning algorithms to predict the categories of happiness. The definition of happiness is not something our research has focused on, in fact, this requires a more philosophical approach than computational.

B. Detecting 'category' of happiness

Our applied algorithm detects categories of happiness with a pretty good accuracy score. Accuracy is a subjective term, and accuracy in machine learning depends more on the dataset provided, the inherent biases of the dataset, number of classes, types of classes and other related factors than on the algorithm itself. One can always get better accuracy with binary classification than multiclass classification. In that sense, we have not made any strong claim from our end about this research being a superior one; however, it is a novel one and a good starting point for further explorations.

C. Reasons for happiness

The second research question in our article discusses specific reasons for happiness and how similar events have a disparate impact on distinct subgroups. The reasons of happiness vary a lot based on age, gender, social status, country, etc. Our research takes a primary step and looks for potential relationships between these features.

D. Social and ethical implications

This research work has a significant impact socially. We are effectively looking inside people's head and exploring potential reasons that make them happy. This exploration also has implications in marketing campaigns (e.g., targeting emotions that make people happy during holidays), in product development (e.g., offering new products based on the happy moments), or in psychological therapies (e.g., what makes similar minded people happy). However, we need to take into consideration that we might be missing out confounding variables. We are already missing out significant variables, like exact timestamp of happiness, social and

financial status of an individual. We can safely assume that financial status has a notable impact on factors that make us happy. Similar other variables might be missing out from the dataset too. This way, our analysis has the potential of being misdirected and falling into wrong conclusions.

E. Future works

The dataset is one of its kind and a relatively newer one. This opens multiple possibilities for further explorations. One good way could be analyzing the whole dataset, and explore a better classification system of happiness. We could also try to build similar datasets for other emotions, like sadness, anger, fear, disgust, etc. Another possible extension of this research could be on an application level; we could effectively build a sentiment analysis tool and deploy it online to let more people experiment and collaborate. Similar experiments have been done in MIT Moral Machine web applications [21].

SECTION X. Conclusion

This research paper analyzes the moments and activities that make people happy, based on a collection of happy moments. We are focusing on specific happy moments from a collection of text responses that people have shared through the crowd-sourcing platform: Amazon Mechanical Turk (MTurk). Using crowd-sourcing to collect the data allows us to advance our understanding of the cause of happiness, by focusing on words and real human experiences. Workers of MTurk were asked to reflect on events making them happy in a given period and share three specific moments in complete sentences.

Through text-based analysis, we have explored what other components have a role in making a specific event happy and further analyze how we can classify such words. By using a bag of words approach and taking the top 6,700 words out of the total 25,400 words in the dataset, we were able to predict the happiness category of the moments with a 94% accuracy. The model we used was a Linear Support Vector Classifier. Although affection and achievement make up two-thirds of our happiness categories, we still were able to successfully predict the less represented categories up to a reasonable extent of accuracy.

Irrespective of some misclassification by MTurk workers, our models were able to accomplish the goal of understanding a happy moment and fit such a moment into one of the seven ground truth happiness categories we set at the beginning of this study. Also, we dive deeper into specific subcategories of classifiers in an attempt to find insights about their happiness level based on specific factors.

REFERENCES

1. A. Slivkins and J. W. Vaughan, "Online decision making in crowdsourcing markets: Theoretical challenges", *ACM SIGecom Exchanges*, vol. 12, no. 2, pp. 4-23, 2014.
2. T. Althoff, K. Clark and J. Leskovec, "Large-scale analysis of counseling conversations: An application of natural language processing to mental health", *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 463, 2016.
3. A. E. Clark and A. J. Oswald, "A simple statistical method for measuring how life events affect happiness", *international Journal of Epidemiology*, vol. 31, no. 6, pp. 1139-1144, 2002.
4. *rit-public/happydb*, [online] Available: https://github.com/rit-public/HappyDB.
5. M. Hossain, "Users' motivation to participate in online crowdsourcing platforms", *Innovation Management and Technology Research (ICIMTR) 2012 International Conference on*, pp. 310-315, 2012.
6. N. Kaufmann, T. Schulze and D. Veit, "More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk", *AMCIS*, vol. 11, no. 2011, pp. 1-11, 2011.
7. D. C. Brabham, "Moving the crowd at threadless: Motivations for participation in a crowdsourcing application", *Information Communication & Society*, vol. 13, no. 8, pp. 1122-1145, 2010.

8. M. T. Whitty, "Liar liar! an examination of how open supportive and honest people are in chat rooms", *Computers in Human Behavior*, vol. 18, no. 4, pp. 343-352, 2002.
9. A. Bryson and G. MacKerron, "Are you happy while you work?", *The Economic Journal*, vol. 127, no. 599, pp. 106-125, 2017.
10. D. Brett and A. Pinna, "The distribution of affective words in a corpus of newspaper articles", *Procedia-Social and Behavioral Sciences*, vol. 95, pp. 621-629, 2013.
11. H. A. Schwartz, M. Sap, M. L. Kern, J. C. Eichstaedt, A. Kapelner, M. Agrawal, E. Blanco, L. Dziurzynski, G. Park, D. Stillwell et al., "Predicting individual well-being through the language of social media" in Biocomputing 2016: Proceedings of the Pacific Symposium, World Scientific, pp. 516-527, 2016.
12. J. Bollen, B. Gonçalves, G. Ruan and H. Mao, "Happiness is assortative in online social networks", *Artificial life*, vol. 17, no. 3, pp. 237-251, 2011.
13. P. Budner, J. Eirich and P. A. Gloor, "making you happy makes me happy"-measuring individual mood with smartwatches", 2017.
14. J. Li, S. Roy, J. Feng and T. Sim, "Happiness level prediction with sequential inputs via multiple regressions", *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 487-493, 2016.
15. A. Asai, S. Evensen, B. Golshan, A. Halevy, V. Li, A. Lopatenko, et al., "Happydb: A corpus of 100000 crowdsourced happy moments", 2018.
16. S. Li, *Multi-class text classification with scikit-learn*, [online] Available: https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f.
17. D. Data, *Deon: An ethics checklist for data scientists*, [online] Available: http://deon.drivendata.org/.
18. *Amazon mechanical turk*, [online] Available: https://www.mturk.com/.
19. *Amazon mechanical turk privacy notice*, [online] Available: https://www.mturk.com/worker/privacy-notice.
20. "Social and E. I. of Data Science: Marquette University", *Mscs-5931-happydb*, [online] Available: https://github.com/Nathanlang14/MSCS-5931-HappyDB.
21. *Moral machine*, [online] Available: http://moralmachine.mit.edu/.