Remy’s Bias Looking Glass - Chapter One

Nik Bear Brown

11/6/2020

# Remy’s Bias Looking Glass

Remy’s Bias Looking Glass is a tutorial on analyzing and visualizing bias and fairness on Twitter. The book follows the adventures of Rembrandt X Einstein, Remy, and his partner in crime Angela G Hopper, Angi, as the meander through Twitter Land.

They are both “AI Skunks,” that is fellows at a think tank of the same name. They are both skunks. Angi is a professor with a PhD in Artificial Intelligence. Remy is an entrepreneur and Tesla-like inventor. In spite of being soul mates in many ways they are opposites. She is a liberal and he is a conservative. She is a little younger and he is a little older. She is female and he is male. She is an atheist and he is spiritual. As a skunk Remy has a little more white than black and Angi has little more black than white. Angi has freckles too, which some skunks look down upon.

This primar will play with the primary statistics of sentiment analysis and standard statistics for measuring bias. That is the statistical parity difference, equal opportunity difference, average odds difference, disparate impact, Thiel index, partial dependence, and SHAP values.

The focus of this journey is on how to transform these statistics into data visualizations and to understand bias and how it differs from fairness.

## Chapter One - Twitter Bias Statistics

In chapter one Remy and Angi explore the statistics of bias. Later chapters will explain these statistics in more detail. More sophisticated analysis such as partial dependence plots, ICE plots, SHAP force plots, and SHAP summary plots will occur later in the journey.

## Am I Biased? Creating a Twitter Looking Glass

After tweeting in support of Remy’s favorite politician, Mayor McCheese, several people responded calling Remy such names as a snollygoster, a Bieber, a lickspittle, and a ninny. That if Remy weren’t such a tosser and twit he would support Coriolanus Snow.

Worst of all Remy was accused of being biased and unfair, as he asserted they were. They got into an argument about who is more biased. Rather than argue, Remy decided to investigate. To accomplish this Remy wants to build an AI looking glass which would allow people to visualize bias in writing, in particular in Tweets. The AI looking glass can also be used as a mirror to look at one’s own bias.

## Rembrandt X Einstein, Remy

By the way, my name is Rembrandt X Einstein, or Remy as my friends call me. I am an AI researcher and professor at the Smarty Pants Institute. I build predictive models using AI all of the time but sometimes wonder if the predictions they make unfairly affect people. I also wonder what my own biases are. It is time to use math to find out.



Tesla MacBook Remy

## Can Math Detect Bias?

Let’s start with a simple definition of bias. If the probability of an outcome is equal given the same inputs then this is unbiased. For example, if two people with the same qualifications have the same probability of getting a job or having the same salary then there is not bias. Conversly, if the probability differs significantly then there is not bias. For example, if two people have the same qualifications (e.g. years experience, education, etc.) they should have the same chance of getting a job.

In math, we call the outcome variable the target or dependent variable. Convention has a one call the target variable , and in particalur it is usually called or “y-hat” to indicate is is an estimate or prediction or as the true is rarely known. The input variables, thing like years experience, education, techincal skill, etc are called input variables or dependent variables and usually called , we use a captial rather than a lower case to indicate that there can be many input variables, or a vector rather than a single scaler value.

This leds to one possible mathematical definition of bias.

The above reads the probability of y-hat is the same if the inputs are the same. This is a common sense notion, that is there is no bias when the given the same input one gets the same result.

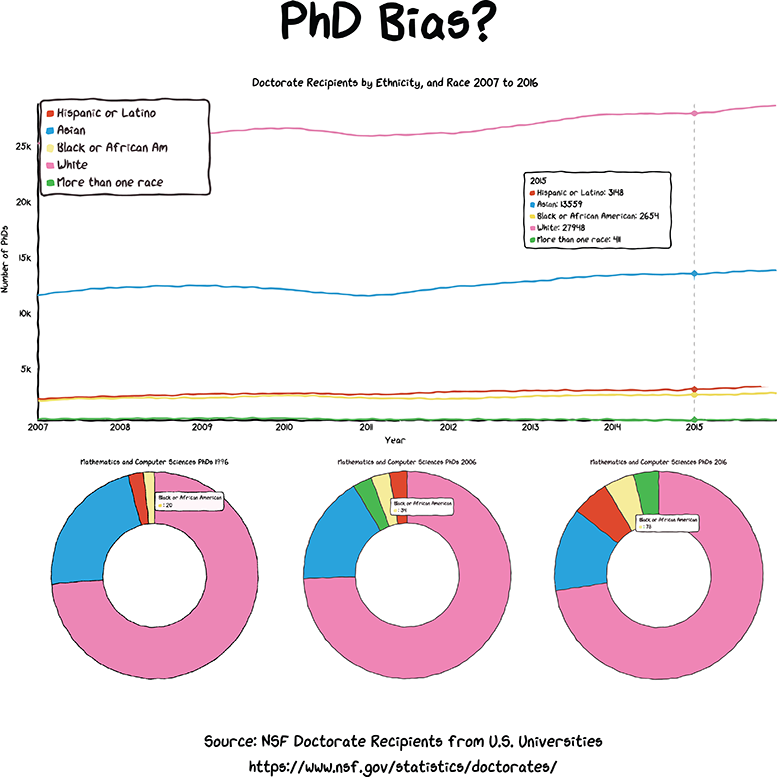
We will continue to come back to this idea the outcome being different when the relevant input is the same as we continue to refine our exploration of bias and fairness.



Same Content with a Different Response indicates Bias

## Can We Visualize Bias?

If our notion of bias is disparate outcomes the one should be able to visualize bias. One just needs visualizations that show disparate outcomes in the data.



PhD Bias

The graphic above shows that those who self-indentify as White and Asian are awarded the most PhDs in the United States. This disparity has existed for decades. But is it bias?

According to the US Census those who self-indentify as White make up 76.3 percent of the population, and those who self-indentify as Black or African American make up 13.4 percent.

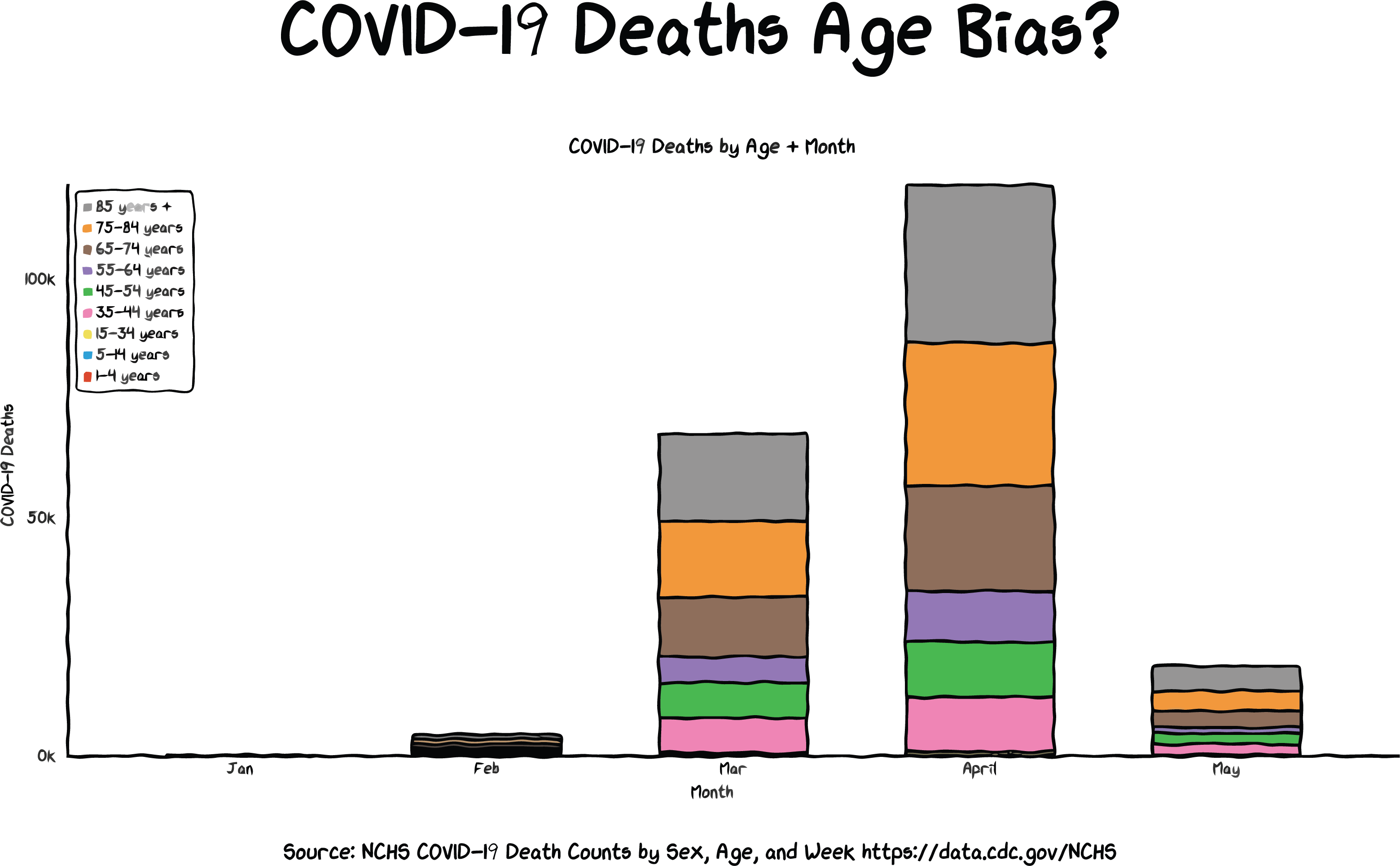
A logical expectation would be that those who self-indentify as White make would be awarded around 76 percent of the PhDs, and those who self-indentify as Black or African American would be awarded around 13 percent. The expectation is called a *model*. The model essentially states that individuals of all races and ethnicities should be awarded PhDs with equal probabilites, so the proporation should reflect the number of tries. If one accepts this model, does the graph indicate bias? According to the [NSF Survey of Earned Doctorates](https://www.nsf.gov/statistics/srvydoctorates/) African Americans are awarded around 5-7 percent of all doctoral degrees to students who are U.S. citizens. This means one would expect African Americans are awarded around 14 percent, but are actaully awarded around 6 percent. Further in certain fields, expecially STEM fields Blacks are vastly underrepresented among doctoral degree recipients.

For example, African Americans earned only 1.8 percent of all doctorates, 3.8 percent of all mathematics and statistics doctorates, 3.7 percent of all doctorates in computer science, and only 4.1 percent of all doctorates awarded in engineering disciplines. In 2017, there were more than a dozen fields—largely subfields within science, technology, engineering, and math—in which not a single doctoral degree was awarded to a black person anywhere in the United States.

## Is Bias Always Bad?

Bias just relfects the inequality of outcome. Sometimes that reflects the reality of the world. Basketball players are biased towards being tall. One reasonable model of the world would be the height helps one play basketball so this makes sense.

COVID-19 shows a clear bias with respect to age. Older people are far more likely to die from coronavirus infection.



Age Bias COVID-19

In spite of a median age of 40 years old in the United States, nearly all of the COVID-19 deaths are 65 and older. There is a large age bias with respect to COVID-19 deaths. While this is, by definition, bias it is not necessarily unfair as one would expect this population to be more vulnerable based on biology. While fairness is related to bias, it is not the same thing.

## Bias versus Fairness

Bias is when there are unequal outcomes. Fairness is when the unequal outcomes are undeserved. For example, if a man and a woman were to apply for to a prestigious company for an artificial intelligence engineering job and the man had a PhD in Computer Science and the woman didn’t graaduate college most would not consider that unfair. However, if they both had essentially the same qualifications and were both highly qualified and the man were far more likely to get the job based on the only difference being gender then most would consider that unfair.

## Visualizing Ones Own Bias

One can explore this idea of responding to the same input differently based on something immaterial to the outcome or target of interest in order to develop algorithms and visualizations of bias, including understanding ones own bias. For example, if two politicians tweet essentially the same idea and one response depends more on whether the tweeter was liberal or conservative than the content of the tweet that would indicate bias.

We are going to call the variable that is immaterial to the outcome . This would lead to a mathematical expression of fairness as follows:

The equation would mean that the probability of the outcome would not be affected by the input .

For example, ones salary would not be affected by ones gender, but rather other input such as education, years experience, skill level, etc.

## Remy’s Taco Bias

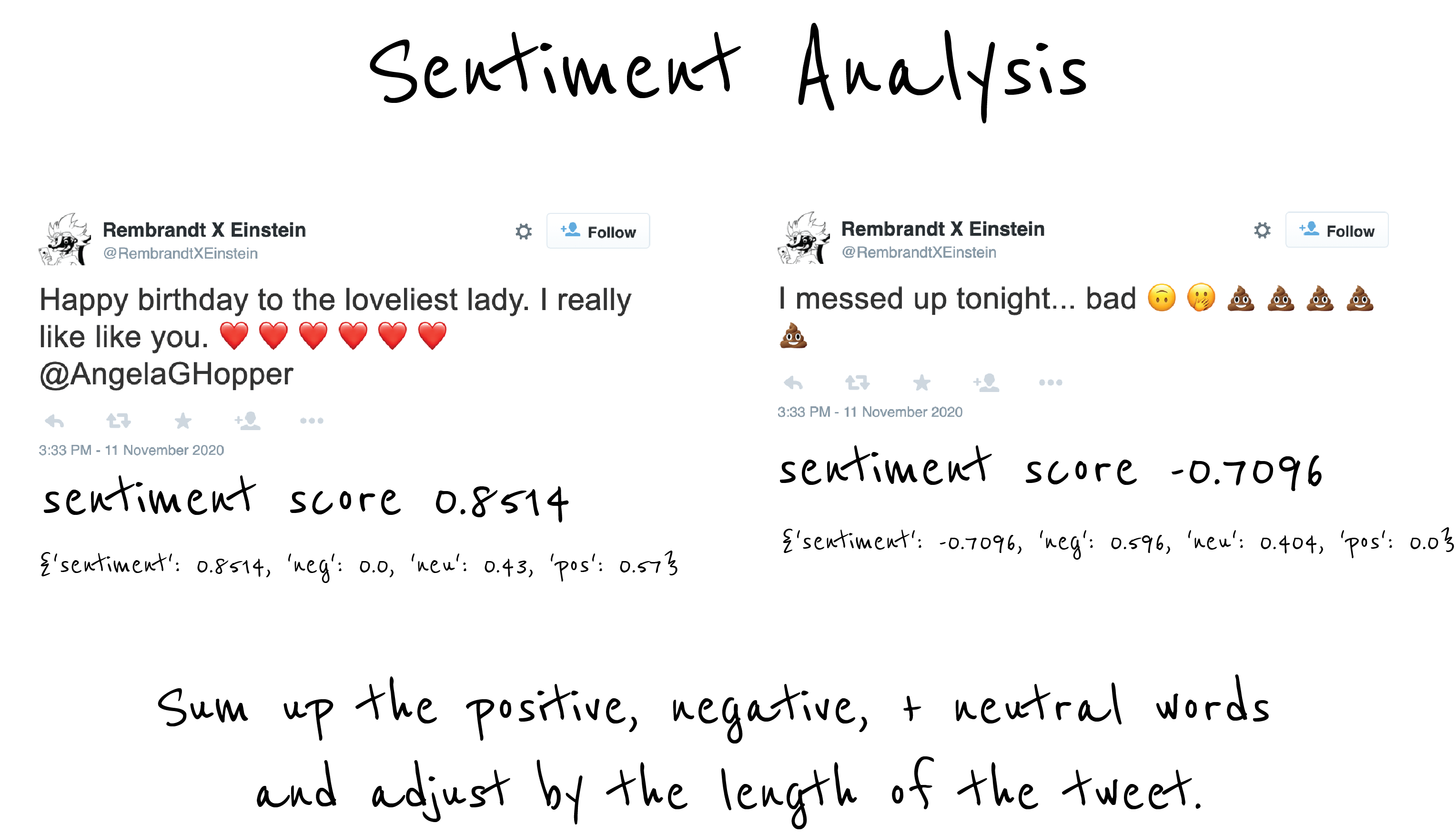
Remy and his BFF Angi eat togther but she is a vegan, and Remy is an omnivore. They came up with the following compromise for their dietary schedule.

* Meatless Monday
* Taco Tuesday
* Waffle Wednesday
* Tofu Thursday
* Stir Fryday
* Weekend Wheatgrass Juice
* Weekend Whisky & Wieners

Both Remy and Angi express their food sentiment on Twitter.

 They decided to plot their Twitter food sentiment for a week. To get a baseline they also decided to plot their followers food sentiment for the same period and average those as well.

The sentiment analsyis is pretty straight forward. We classify tweets with a probability of being related to food, accept those which have a very high probability of being food related and then calculate the sentiment of those tweets by summing up the positive, negative, + neutral words and adjust by the length of the tweet.



Sentiment Analysis

## Classifying Food Tweets with Naive Bayes

We classify tweets with a probability of being related to food using the Naive Bayes algorithm. [Naive Bayes classifiers](https://en.wikipedia.org/wiki/Naive_Bayes_classifier) are a family of simple probabilistic classifiers based on applying Bayes’ theorem with strong (naive) independence assumptions between the features.

Abstractly, naive Bayes is a conditional probability model given a problem instance to be classified, represented by a vector representing some features (independent variables), it assigns to this instance probabilities

for each of possible outcomes or classes .

The problem with the above formulation is that if the number of features is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes’ theorem, the conditional probability can be decomposed as

In plain English, using Bayesian probability terminology, the above equation can be written as

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on and the values of the features are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

which can be rewritten as follows, using the Chain rule for repeated applications of the definition of conditional probability

Now the “naive” conditional independence assumptions come into play assume that each feature is conditionally statistical independence|independent of every other feature for , given the category . This means that

.

Thus, the joint model can be expressed as

This means that under the above independence assumptions, the conditional distribution over the class variable is

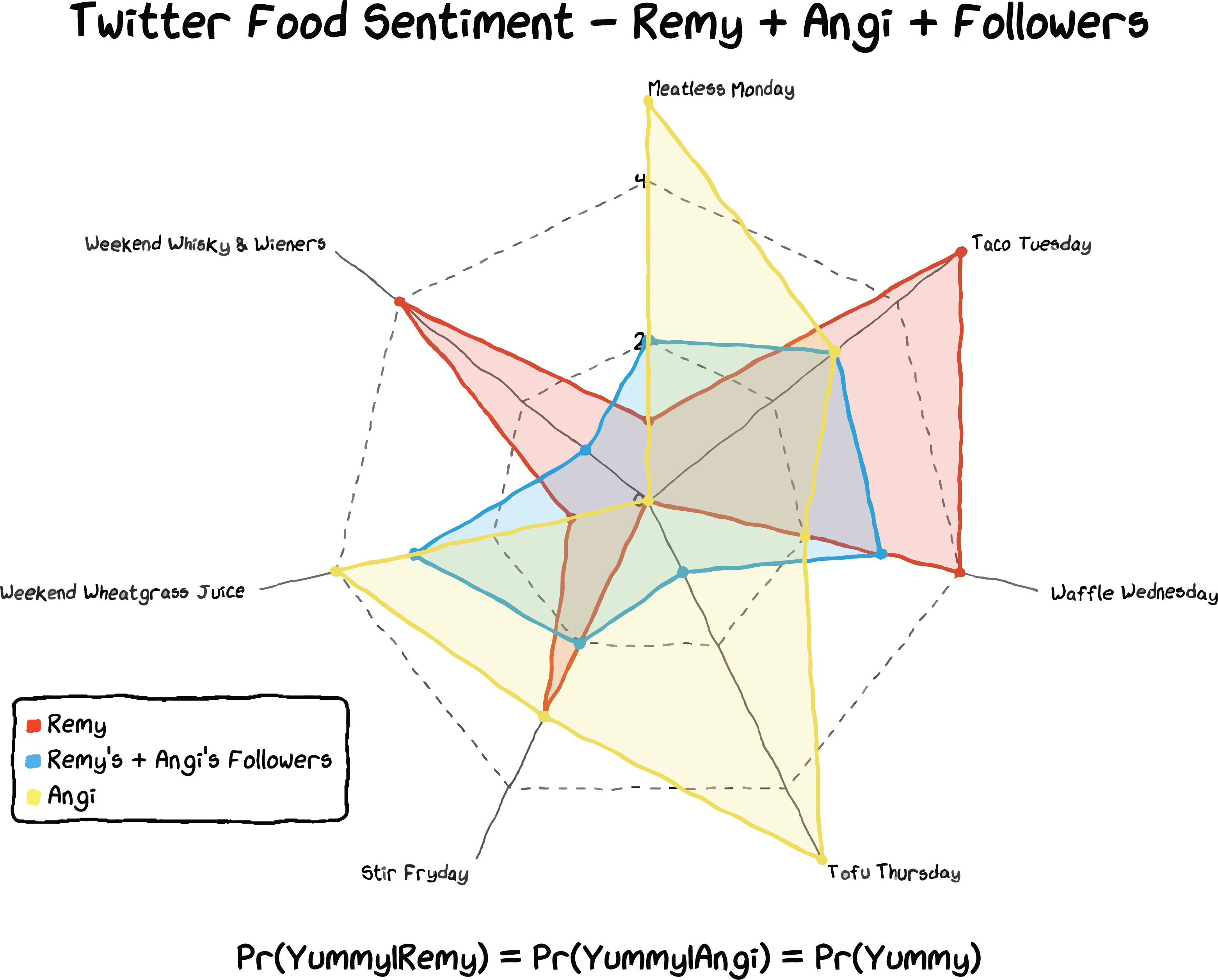
where the evidence is a scaling factor dependent only on , that is, a constant if the values of the feature variables are known.

*Constructing a classifier from the probability model*

The discussion so far has derived the independent feature model, that is, the naive Bayes probability model. The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the ‘maximum a posteriori’ or ‘MAP’ decision rule. The corresponding classifier, a Bayes classifier, is the function that assigns a class label for some as follows

## Remy and Angi’s Food Sentiment

Remy and Angi’s food sentiment is as one would expect. Remy loves tacos and waffles and is not fond of his veggies. Angi, a vegan, loves her tofu and greens.



Twitter Food Sentiment

## Remy and Angi’s Political Sentiment

Remy and Angi’s food sentiment is as one would expect. Remy loves tacos and waffles and is not fond of his veggies. Angi, a vegan, loves her tofu and greens. While this food preference is a bias it rarely leads to arguments.

What about their politics? In spite of being soul mates in many ways they are opposites. She is a liberal and he is a conservative. While they rarely argue over food they often argue over politcs.

## What is Next?

In our next adventure we will look into the difference between bias and fairness. Remy and Angi will try to understand *fairness* and *discrimination*. Discrimination is the unequal treatment of individuals of certain groups, resulting in members of one group being deprived of benefits or opportunities. Common groups that suffer discrimination include those based on age, gender, skin color, religion, race, language, culture, marital status, or economic condition.

Remy and Angi study the statistics of *fairness* and *discrimination* including statistical parity difference, equal opportunity difference, average odds difference, disparate impact, and the Thiel index.

```