# Dataset: We Feel Fine

We Feel Fine, by Jonathan Harris and Sep Kamvar, automatically scrapes blog posts every ten minutes, looks for "I feel" or "I am feeling" statements, and tries to extract single-word feelings from those posts (along with any other demographic and geographic information it can reasonably derive). It's a large dataset (a fairly conservative estimate puts it at over 38 million entries\(^1\)), accessible via a REST API.

Notable limitations and peculiarities:
1. Feelings are single words: phrases are far more expressive. Feeling extraction, also, is sometimes wrong.
2. We didn’t find any data for November 2012.
3. Most entries seem to be from the United States. State information, at any rate, is predominantly American when present.

Officially, We Feel Fine offers 14 attributes, most of which, we found, were often (and unfortunately) missing. After feeling our way through various smaller slices, we decided to focus on larger trends and point anomalies for more recent years. Our final dataset has a little over 1.2 million entries, and traces posts for 43 assorted feelings\(^2\) from January 1st, 2010 to September 30th, 2013.

---

\(^1\) We Feel Fine apparently grabs at least 15,000 feelings per day. We’ve seen entries dated as early as August 2005. To simplify, let's just estimate the number of entries collected from 2006 to 2012: 15,000 feelings per day x 365 days per year x 7 years = 38,325,000 feelings—quite a bit, if you ask us.

\(^2\) In alphabetical order: accomplished, alive, alone, angry, awkward, bad, better, blah, bored, comfortable, confident, dead, down, empty, fat, frustrated, good, guilty, happy, helpless, lonely, loved, lucky, nervous, ok, okay, overwhelmed, pretty, proud, sad, safe, scared, shit, sick, sorry, special, stressed, stuck, stupid, tired, trapped, useless, worse. We tried to pick a variety of feelings that were (1) frequent enough to meaningfully analyze (see [here](#) for some counts), and (2) not too ambiguous (eg. "well").
# The yearly feeling "landscape" actually stays much the same

Figure 1 above first divides by year, then allocates one box per feeling, then colors and sizes each box by post count. Larger areas and darker colors indicate a larger number of posts.

We exclude seven saturated feelings (bad, better, down, good, guilty, happy, sad) to avoid confusion. Said saturation results from a technical limitation: we could only retrieve at most 1,500 entries per API request, and could only make a finite number of requests, so some of the more commonly expressed feelings hit an artificial ceiling.
Figure 1 provides an overview of most feelings in the years 2010, 2011, 2012, and 2013. Apparently, our feelings are more similar than they are different: box colors and sizes, both representations of post count, pretty much stay the same from year to year (except for 2013, which is lighter because it isn’t over yet). We didn’t expect this. This may, granted, be because of the particular feelings we chose. The similarity is striking, nonetheless.

Figure 2: Tracking frustration across time

(a) Monthly post counts for "frustrated" since January 2010

(b) Monthly post counts for all 43 feelings since January 2010

# We were really frustrated this past September (part 1)

Figures 2a and 2b (left) depict monthly post counts for each month from January 2010 to September 2013. Figure 2a follows one feeling, frustrated; Figure 2b instead covers all 43. The precise counts aren't the important part here: the trends are.

Most feelings, we've found, follow this rule: when there are more posts for all feelings, there are more posts for that particular feeling. This follows from the previous observation about the relatively static yearly "landscape".

"Frustrated", shown in Figure 2(a), obeys this rule almost perfectly—until it hits September 2013. There, it turns around and goes the opposite direction.

As an aside: note the downward trend for overall post count in 2013 (see Figure 2b). Perhaps people are migrating to newer mediums that Harris and Kamvar aren't scraping?
# We were really frustrated this past September (part 2)

Figure 3 above focuses on the 30 days in September 2013. The counts here are naturally smaller than the counts in Figure 2, but the mid-month jump is still very clear.

Drilling down a bit more reveals that the spike in frustration isn’t an all-of-September phenomenon, but a last-half-of-September phenomenon. Why? We can’t say for sure. One sentence doesn’t provide enough context for us to make any confident claims. But it is rather coincidental, small counts notwithstanding, as those were the weeks right before the government shutdown.
# We feel (predictably) "loved" and "special" on some days, but unusually overwhelmed on others

Figure 4 (above) and 5 (next page) examine daily post counts for three specific feelings: "loved", "special", and "overwhelmed". Points are sized and colored by post count to reduce noise. Larger shapes and darker colors indicate larger post counts. Of note: the odd November 2012 gap mentioned previously is visible in all plots, as are some single-day anomalies.

Continuing further into the data, it's clear that there are not all days are equal. "Special" and "loved", shown above, unsurprisingly show little blips around Valentine's Day. Looking at "proud" reveals one such blip on July 4th. These all have apparent, plausible explanations; Harris and Kamvar found similar things.
Not all such blips are easy to explain. Consider, for instance, "overwhelmed".

We have yet to find a satisfying explanation for why we felt particularly overwhelmed around June 25th and 26th 2011, marked above. Digging around events at that time hasn't unearthed anything particularly compelling. It's possible that this is merely random noise, but it, and the other unexplained blips we came across, seem worth investigating.
# Spotfire Critique

We used Spotfire to analyze our data. It performed admirably overall.

Pros:
- It was easy to jump into. Datasets loaded quickly, with minimal fuss, and basic controls were generally discoverable. Default color palettes were sufficiently distinct and reasonably pretty.
- Visualizations were generally (perhaps surprisingly) responsive. Trellising, swapping axes, changing attributes and so on rarely incurred a significant delay even with our 1.2 million rows.
- There's a fair amount of more advanced customization, both visual (colors, fonts, shapes, etc.) and functional (some sort of SQL-esque interface for limiting data ranges, for example).

Cons (and suggestions):
- Context menus. Right-clicking on different areas of the visualization pulls up different menus. Right-clicking on the visualization opens one menu; right-clicking the legend gives you another; right-clicking the axes gives still another; and the list goes on. It's not at all clear where these context divisions are, however. Drawing in thin lines would help here. Or perhaps lightly shade the background for the different functional areas. Better yet, try to standardize the context menus so I don't have to keep guessing.
- There are lots of icons, and not that many shortcuts. The latter is becoming increasingly frustrating. Please add shortcuts for more common actions!
- Possible bug: at one point, we were trying to create a line graph that had missing data. There's an advanced option to break the line when there's no data, but we checked the option and nothing happened (far as we can tell).