"I Am Not Racist, But..."

What Your Reddit Comments Say about Your Actual Opinions, According to Sentiment Analysis

by Ege Sagduyu

Introduction

I am sure you have encountered that phrase before. Maybe it was a colleague or a friend who uttered it with a careful whisper after looking over his or her shoulders, or maybe it was your obnoxious drunk uncle who blurted it out at a family reunion. Regardless of the context, the phrases that start with this kind of self-justification almost always expected to be followed by a negative, insensitive remark and an awkward moment. However, as a student of statistics, I must ask: is this usually the case? Do we use this phrase solely as an excuse to vent out our horrible biases and prejudices, or are they also used to initiate an open discussion without instant criticism? Are these phrases mostly used to convey negative sentiments (as we have come to expect them to), or might the reality be different? Or, as Reza Farazmand puts it, can we still be considered while writing this sort of stuff, or does "the search" continue?

These were the questions that popped in my head when I saw that a fellow data scientist scraped all Reddit comments from May 2015 that included the phrase "not racist, but" shortly after the dataset was made available on Kaggle. I decided that best way to answer these questions is to perform sentiment analysis on each of these comments to see whether the average sentiment they conveyed was negative or not. Needless to say, the results painted a really interesting picture.

What a Comment Feels Like

But how does one go about figuring out the sentiment of a comment? On the sentence level, sentiment analysis is the process of looking at the structure of a sentence, determining the relationship between the words, and then assigning it a polarity score based on the strength of the opinion. This strength is usually determined by looking up the polarity of a word on a "polarity lexicon", an interconnected cluster of words and their synonyms which have polarity values (ranging from -1.0 for the most negative, to +1.0 for the most positive) manually assigned by expert taggers.[1] The polarity of individual words, combined with the information conveyed by the sentence structure (which adjective acts on which noun etc.), gives adequate information on the emotional polarity of the sentence.

[1] Usually, there are also scores associated with the subjectivity of a text, but for the purposes of this article, I will focus only on the emotional polarity. For a more technical and exhaustive introduction, check out Bing Liu's excellent tutorial here.

In order to obtain my data set from this corpus, I used the sentiment analysis feature of an easy-to-use Python library called TextBlob and obtained a polarity score ranging between -1 and +1 for each of the 285 comments that contain the phrase "not racist, but". Since what I wanted to look at was the comments with actual sentiments (whether it is positive or negative), I then discarded all the neutral comments (i.e. the ones with a polarity score of 0), thus reducing my sample size to 152 polar comments. With these polarity values, I first calculated the corresponding z-scores to see if the normal assumption is reasonable on a normal Q-Q plot (Figure I). Additionally, I plotted a histogram and compared it against the curve of a normal distribution using the same sample mean and standard deviation to see how it fits the data set. (Figure II)

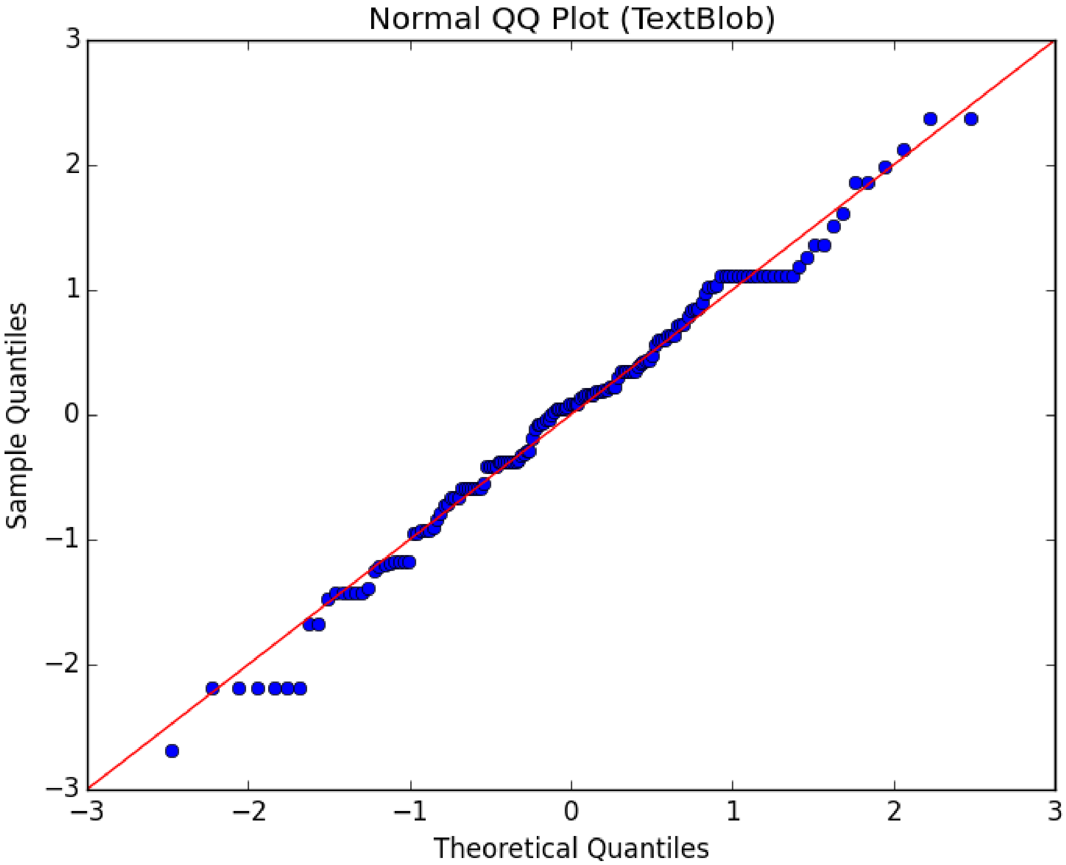

Figure I: A normal Q-Q plot comparing the normalized polarity scores against standard normal distribution. From the linearity of the fit, we can say that normal assumption is reasonable.

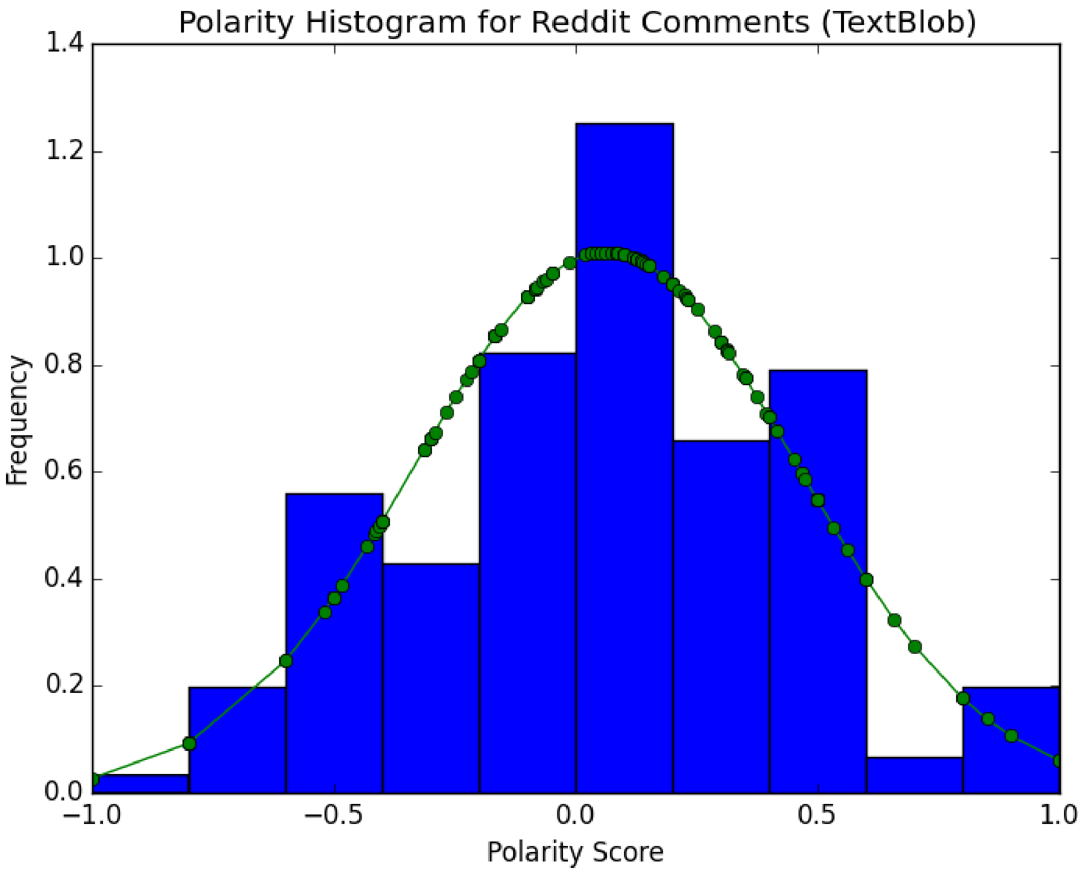

Figure II: Histogram for our polarity score data, with the corresponding true normal distribution plotted in green for comparison ($N = 152, \bar{x} = 0.0631, s = 0.3948$)

Let's Delve Deeper

As you can see from the fit of the plots, the normal assumption is reasonable. So we can finally construct a 95% confidence interval to see where the true mean lies most of the time. Using the interval function of scipy's stats package for normal distribution we obtain the following confidence interval (CI) for this sample when $\alpha = 0.95$:

$$(0.00033, 0.12584)$$

Now let's step back and see what this CI tells us about our initial question. We started by stating that most of us think the comments that include the trigram "not racist, but" would probably be end up conveying a negative sentiment against a group of people. In terms of hypothesis testing, this means that our null hypothesis was the true mean of polarity scores would be negative. However, the fact that our CI is strictly positive for this sample points suggests that the null hypothesis might be incorrect.

We can now delve even deeper by calculating the necessary p-values. Using a one-sided t-test (where our null hypothesis is $H_0\colon \mu \leq 0$ and our alternative hypothesis is $H_1\colon \mu > 0$) via stats.ttest for scipy, we obtain $Pvalue = 0.0257$ and since $Pvalue < \alpha$, we reject the null hypothesis. So we can be happy to say that not *all* Reddit users are horrible.

But What Does This Tells Us?

In the case of Reddit comments that contain the phrase "not racist, but", we have seen that our sample suggests not all people use this phrase to say terrible things and instigate hatred. Now, since the capabilities of sentiment analysis is somewhat limited when it comes to detecting advanced nuances in a language, it is maybe too soon to claim we are a civilized species capable of rational thought and well-mannered discussions at all times. However, the main part of what makes the Internet so amazing is that people from every walks of life can chime in, contribute, and cultivate a community, especially when it comes to something of Reddit's scale. And while it is certainly true that there are a lot of bigoted Internet users with bad intentions and misanthropy, I think this experiment shows us how one phrase cannot be the sole determinant of a person's conversational abilities, so taking the time to empathize and understand that person's viewpoints become all the more important. Thus, the search for intelligent life *must* continue.