Turing School of Software and Design

Module 4 Capstone Project – 1703 BE

September 2017

Wakaru

Customer Relationship Management Tone Aggregator

Carl Richmond

Repository

Abstract

Curating user experiences has become central to many industries in the digital age, where presenting a unified story, across company and product, is central to a holistic marketing strategy (Zeiser 316). Customer Service represents one of the pillars of business and controlling a user's experience in that area is important to presenting a unified front.

With the emergence of accessible natural language processing technologies, such as IBM's Watson Tone Analyzer and Google's Natural Language API, companies can now track the language and tone of their front-facing employees — such as customer service agents.

The present work is focused on developing methodologies to manage the data received from these natural language processing programs, in such a way as to infer whether front-facing employees are contributing to the story of a business or hurting its overall brand.

Wakaru, the tone analyzer app, aims to present these conclusions to businesses in an easy to understand manner.

Introduction

The history of natural language processing began in the early 1950s with Alan Turing proposing, in his article *Computing Machinery and Intelligence*, that one method to determine a machines intelligence would be to play an imitation game, where the computer must fool an interrogator into believing it is human. At a fundamental level this involves a program breaking down human sentences and text into machine understandable data.

The work around Wakaru does not involve contributing any new ideas to the field of natural language processing, but rather how a business can deal with the resulting data in a statistically sound way. It aims to answer the question of whether a business can use the data from IBM Watson's API to inform its practices in Customer Service departments.

Sample Selection

The sample size for Wakaru was determined using the following equation for a company of 200 employees, confidence level of 80% and margin of error of 5%.

$$\frac{\frac{z^2 \times p(1-p)}{e^2}}{1+\left(\frac{z^2 \times p(1-p)}{e^2 N}\right)}$$

Population size = N | Margin of error = e | Z-score = z

To reach the goal of 91 samples, surveys were sent out and recipients were prompted to reply as if they were customer service agents. All the surveys and results can be found here. The surveys were split into several categories to focus the language of the respondents and help identify trends in different customer service interactions. Twelve different surveys were created, each with a unique prompt, to cover a variety of different scenarios.

Following the study done by Software Advice (Software Advice), the categories were set as: granting requests with a good tone, denying requests with a good tone and denying request with a bad tone. The study found that customer's perception of tone changed dramatically between granting and denying interactions. Specifically, customers prefer a casual tone when being granted a request, but find it offensive when being denied one. The categories of the surveys, as with the prompts, were designed to focus in on the tone differences between granting and denying interactions.

IBM Watson Tone Analyzer

IBM Watson's Tone Analyzer returns 10 metrics per segment of text that you provide: disgust, fear, joy, sadness, anger, openness, conscientiousness, extraversion, agreeableness and emotional range. Each metric has a score between 0 (lowest) and x (highest). By themselves, they were found to mean very little – especially with professional language. The words we use in casual conversations are much different than the words we use in a professional conversation, even a casual one, and the tone scores indicate difficulty in classifying professional interactions.

For instance, many good toned responses taken from the sample surveys scored as low as bad toned responses across IBM's metrics. Disgust for instance, had a STDEV of 0.02 across all categories, which indicates the metric is statistically insignificant and cannot help inform whether an interaction was ultimately good or bad – by itself. Anger, fear, openness and emotional range had similar results. The lack of deviation across the tonal categories highlights the difficulty in classifying language and the need to look at all the results together, and focus on how the scores correlate to each other.

Classification

To differentiate the good emails from the bad, the tonal metrics with the strongest correlations were combined into scores. The Pearson Correlation Coefficient or Bivariate Correlation were used to rank the relationships between the different metrics.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Through linear regressions and looking at the overall breakdown of the sample data, Joy and Sadness stood out as being moderately correlated to the categories of the emails, while also being moderately negatively correlated to each other. Their p-values scored 0.004 and 0.07, respectfully, against the null hypothesis, which was that their scores are not related to the category of an email. The p-values of the other Watson metrics were far outside of the scope of what is considered statistically significant — 0.05.

Joy and Sadness' relationship with one another, calculated using the Pearson Correlation, and their other strongest relationships are displayed below. Green represents only the good emails, red represents only the bad emails and blue represents a mix of good and moderate emails.

	Sadness	Openness	Conscientiousness	Extraversion	Agreeableness
oy	(-0.416)(0.231)(-0.314)	(-0.277)(-0.307)(-0.177)	(-0.089)(-0.251)(0.167)	(0.575)(-0.120)(0.558)	(0.456)(-0.184)(0.373)
adness	(-0.416)(0.231)(-0.314)	(0.001)(-0.126)(0.082)	(-0.096)(0.230)(-0.073)	(-0.098)(0.013)(-0.395)	(-0.327)(-0.009)(-0.237)

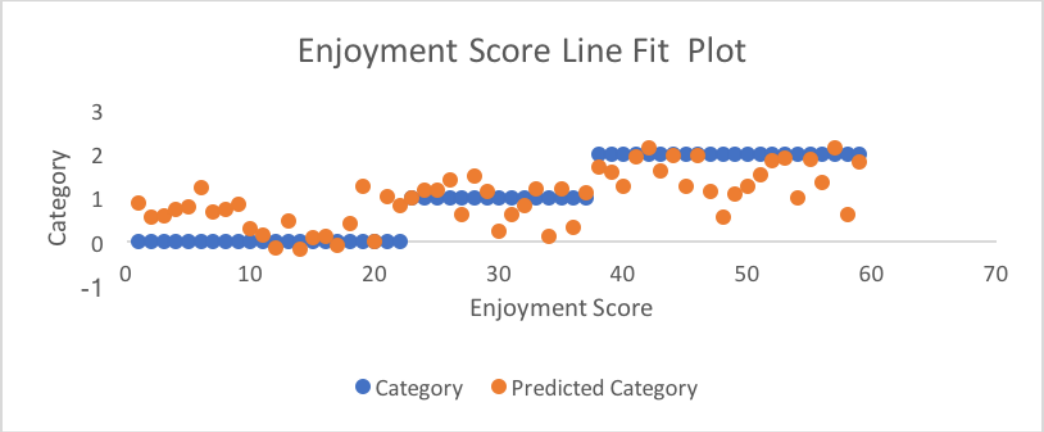
What stands out in the data returned from Watson, is there are few obvious positive relationships that help identify how an email should be classified. To get around this, Wakaru creates scores out of the information that is moderately correlated. Each score has a graph associated with it, to display the results from a regression analysis using it. They are useful in understanding how the score and category of an email are related.

Because Joy was found to be a solid indicator of category, an Enjoyment Score was created using the following formula:

$$r = (\mathrm{E} + \mathrm{A}) * j$$

extraversion = E | Agreeableness = A | Joy = J

The Enjoyment Score is designed to use Joy as a coefficient to the sum of the other metrics that positively correlate with it, to accentuate the relationships that would not be easily identifiable otherwise.



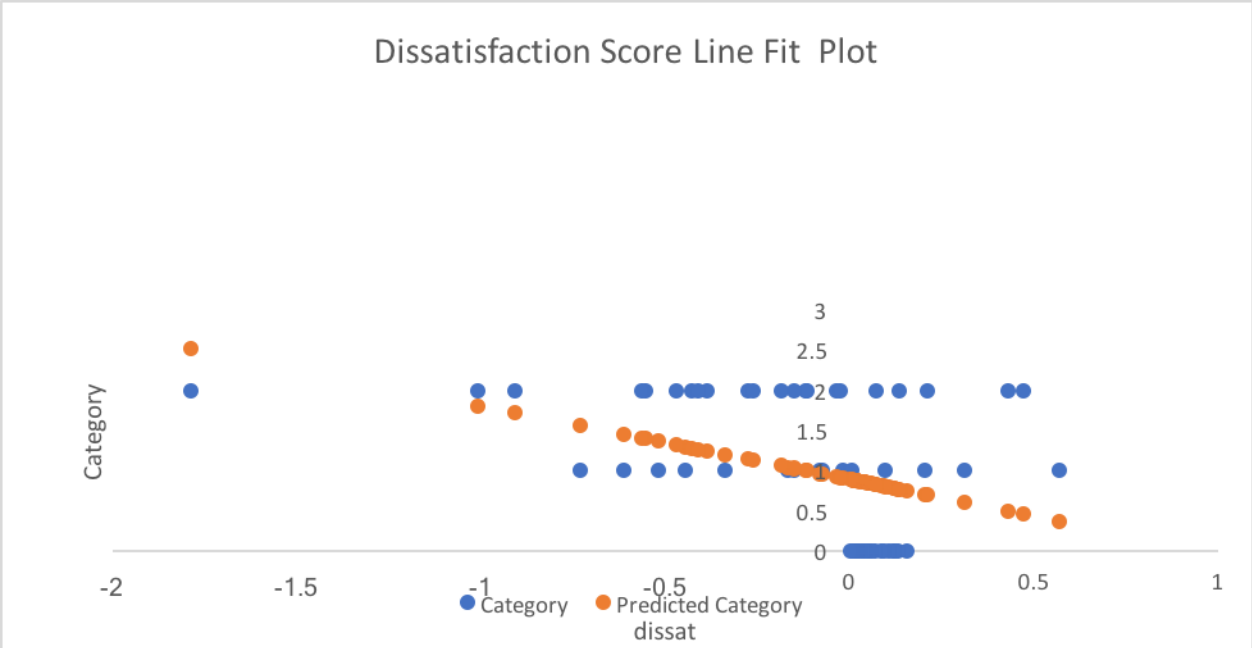
Category 0 = good | Category 1 = moderate | Category 2 = bad

Sadness was found to be the second most important indicator of category, so a Dissatisfaction Score was created using the following formula:

$$r = (\mathrm{E} - e - A) * S$$

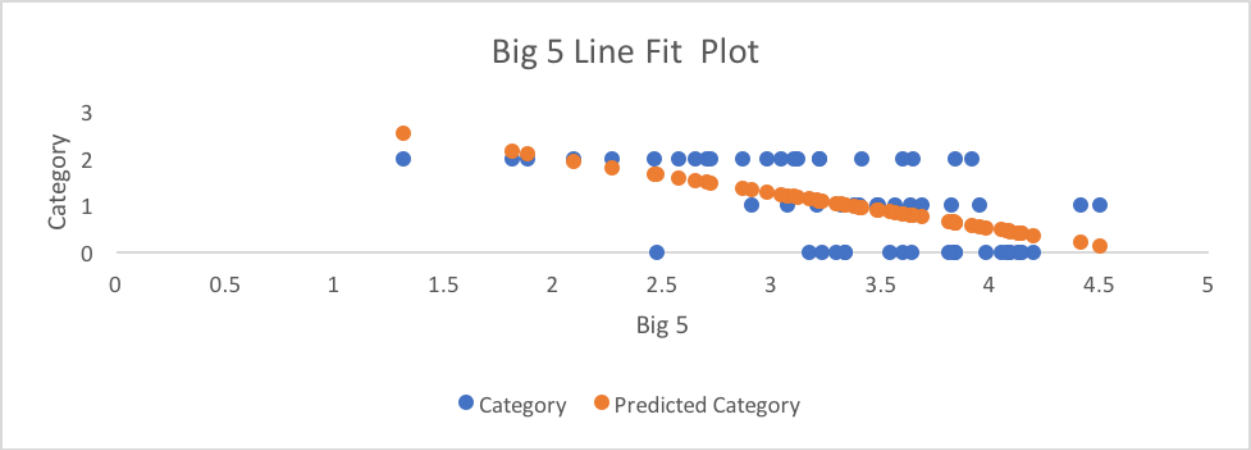
Extraversion = E | Enjoyment Score = e | Agreeableness = A | Sadness = S

The Dissatisfaction Score is designed to use Sadness as a coefficient to the difference of the other metrics that negatively correlate with it, to accentuate the relationships that would not be easily identifiable otherwise.



Category 0 = good | Category 1 = moderate | Category 2 = bad

In the IBM documentation for Watson, they reference the Big 5 emotions—openness, conscientiousness, extraversion, agreeableness and emotional range—as being significant. Because of the importance put onto those categories by the creators of Watson, a Big 5 score was created by adding their scores together to act as a guard against the other assumptions made in the project.



Category 0 = good | Category 1 = moderate | Category 2 = bad

Wakaru looks at the percentile rank of each score—enjoyment, dissatisfaction and big 5—within each email category (good, medium, bad). Each score is given three different percentile ranks, which indicates its placement in relation to the other emails of that category.

A very well written email may have an Enjoyment Score breakdown of: 50th percentile rank in category 0 (good), 75 percentile rank in category 1 (medium) and a 100 percentile rank in category 2 (bad), which indicates that the email's Enjoyment Score is higher than all the bad emails, but only higher than half of the good emails. We can therefore deduce that in the area of Enjoyment Score, the email fits squarely into the middle range of a good email and the high end of a moderate email. This same process happens to the Dissatisfaction Score and Big 5 Score. Once the percentile rank of each score has been calculated, the findings are summarized based on this decision three.

```
1.  def low_medium_high(number)
2.      if number >= 75.0
3.        "high"
4.      elsif number >= 45.0 && number <= 75.0
5.        "medium"
6.      elsif number <= 45.0
7.        "low"
8.      end
9.    end
```

This results in three different scores of either high, medium or low for each of our aggregate scores (enjoyment, dissatisfaction and big 5).

These scores then go into another decision tree to further summarize the information and create two sentences, one to breakdown the overall enjoyment and another to breakdown the effect on the brand. The Enjoyment and Big 5 Scores are used to determine the enjoyment factor of an email, from the customer's perspective, while the Enjoyment and Dissatisfaction Scores are used to determine the brand impact. These generalizations are intended to guide users on how to interpret the raw percentile rank returns displayed by Wakaru.

Wakaru also has the ability for users to set fixtures, which allow them to identify emails that they know to be good, medium or bad. This in effect teaches Wakaru the users style and preference. Fixtures are treated as anchors in Wakaru's analysis, in that they help define email categories. When a user sets a fixture, every email in the database is reclassified using the new parameter. Through the use of fixtures, a users results will be tailored to their specific needs and reflect their professional requirements. This flexibility was created with brand image in mind, as every companies' perception of good or bad is different.

Conclusion

Wakaru is a proof of concept that businesses can use the aggregate data from Watson in ways that can result in actionable insights. Even within the short timeframe of this project (two weeks), it was possible to make a model that can accurately determine extremes of good or bad. With more time and resources, it would be possible to use machine learning and create a much more accurate statistical model. Within a different framework, other than Ruby on Rails, it would have been possible to add-in more statistical analysis as well. The only indication of the standard error rate of Wakaru is the regression models used to make determinations on what Watson metrics to use, and those are not representative of the accuracy of the final product. Much more testing would be necessary to determine how accurately Wakaru picks up the nuances of a moderately bad or good email, for instance. At its core, Wakaru is a very small peek at what we can be achieved through natural language Processing and its potential to inform business practices.

Bibliography

Software Advice. "The Best Tone for Email Customer Support." 2014. *Software Advice.* <http://www.softwareadvice.com/resources/the-best-tone-for-email-customer-support/>.

Zeiser, Anne. *Transmedia Marketing: From Film and TV to Games and Digital Media*. New York: Focal Press, 2015.