“Fantastic, Amazing… 5 Stars!”

Utilizing Natural Language Processing to Bolster Active Management

Kunal Ghosh
Managing Director,
Portfolio Manager

Lu Yu, CFA, CIPM
Managing Director,
Portfolio Manager

Lina Zhao, CFA, CAIA
Assistant Vice
President, Analyst

Brandon Michael,
Senior Product
Specialist

Imagine you are visiting a city for the first time and looking for a great dinner spot… where should you go? One of the most common ways of deriving a restaurant recommendation is via crowd-sourced content such as Yelp[1], whose user-provided reviews consist of a written review of the experience and a final star score ranging from 5 stars (high) to 1 star (low). All else equal, you would likely select a restaurant with a high star rating which is indicative of positive experiences from previous customers. Yet, we show that intended and unintended biases do not guarantee a great dinner if you blindly follow the star ratings.

One approach to assess the potential quality of a restaurant would be based solely upon their written review as many users are extremely generous and blindly hand out 5 stars, while others are stingier about providing such a high score. Clearly, reading each individual review would not be the most effecive use of time given the hundreds or potentially thousands of restaurant-review combinations – resulting in a near impossible task.

Natural language processing (NLP) is a much more sophisticated yet elegant method to approach this problem. Natural language processing utilizes computers to analyze and understand language, based upon key words, which can be trained to come up with a highly predictive result. The process distills written text into a quantifiable score which is useful in the measurement and independent comparison of language. Using natural language processing, users could have a high level of confidence when a restaurant written review aligns with the numerically provided star score, whereas those with differences would warrant further investigation.

- Natural language processing (NLP) –made possible by advances in artificial intelligence and computational linguistics – utilizes computers to analyze and understand language
- This insight can identify inherent behavioral biases in qualitative information, including analyst sell-side reports, transcripts of management calls, and central bank interest rate announcements, among others
- The AllianzGI Systematic team’s philosophy and process is designed to capture inefficiencies caused by human behavior, and NLP represents a wealth of future opportunities to further enhance their time-tested investment approach

A parallel to restaurant reviews can be drawn to investment management, whereby Street analysts provide similar information on their viewpoint for stocks in the form of a review and final score (ratings, estimates, target prices, etc.). The AllianzGI Systematic team (“the Team”) has been embarking down this road of natural language processing to further augment their behavioral finance-focused investment approach. Rather than simply using the summarized investment views at face value, the Team is developing a way to analyze Street analyst reports and come up with their true viewpoint on the stock. The goal: a differentiated, 5 star approach to active management.

1. The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available in both JSON and SQL files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

Appetizer Course: Why Natural Language Processing?

In today's fast paced and data intensive world – where every 2 days human beings globally create as many exabytes of information as they did from the dawn of civilization until 2003[2] – numbers provide the natural advantage of brevity necessary to process vast amounts of information. An article or editorial backed with numbers or statistics automatically makes it more credible and convincing. Numbers, regardless of their relevancy or accuracy, can be easily compared to one another. In contrast, a descriptive text or opinion often captures emotions and "gut feel" of the author, which is much more subjective and open to interpretation. As a result, qualitative language is much more cumbersome to compare.

Natural language processing harnesses advances in the fields of computer science, artificial intelligence, and computational linguistics. Rather than reading volumes of written information to come up with a viewpoint, natural language processing exploits computational power through the input of qualitative text to derive a quantitative score.

Simply put, NLP translates text into numbers which allows for the analysis and comparison of written responses to one another. Natural language processing also provides the added benefit of stripping away biases associated with a purely quantitative score. The result is a process which balances society's reliance on numbers along with the informational advantage derived from language.

Soup Course: Data Analysis & Training

The Team implemented natural language processing of Yelp restaurant reviews to show how a codified number scale can be further improved through the utilization of the often ignored written descriptions.

A critical component of natural language processing is the data analysis and training required to derive a score which is both intuitive and has a high predictive power. User provided written reviews and final star scores can be easily loaded from any platform of text, such as online text or even PDFs, which is converted into a txt or xml file. From here, it becomes the computers function of crunching information based upon the coding and training administered by the Team.

To get a flavor of how the process works, let's first start with how the information is analyzed, which is done in three steps.

1 Turn sentences into words to process the text
This is essentially adjusting the language into a format the computer algorithm can analyze. The sentence: "I want to learn how basic text cleaning works" is cleaned to remove key stop words or those with no productive power (such as "I" and "to") to become: ["want", "learn", "how", "basic", "text", "cleaning", "works".]

2 Analyze the text by turning words into vectors
This is necessary to run regression models required to analyze the words and come up with a solution. There are broadly three types of vectorization methods which can be used, ranging from simplistic to complex.

Bag of words model, or count vectorizer, is the simplistic method which sums up the number of occurrences for each word in a matrix. The two sentences examples "food was delicious" and "food quality was poor" would be processed in the following matrix:

	Food	Delicious	Quality	Poor
Example 1	1	1	0	0
Example 2	1	0	1	1

Using information provided by Yelp, the Team translated reviews into vectors which could be analyzed via logistic regression to predict positive reviews (4 or 5 stars) or negative reviews (1 or 2 stars). The largest coefficients of a positive score were associated with the words "delicious" "excellent" and "incredible" while the negative scores were tended to use the words "mediocre", "worst" and "horrible".

Natural language processing exploits computational power through the input of qualitative text to derive a quantitative score

Term frequency – inverse document frequency (TF-IDF model is a more refined method of analyzing information, which assigns higher weights to terms that appear often in a particular review, but not often across all reviews. In the previous example, the word "food" appeared in both reviews and as such is less relevant in predicting the overall star score. Replacing the bag of words model with the TF-IDF model yields a slightly different outcome. "Delicious" retained the highest coefficient, and instead "amazing" and "pleasantly" surfaced to the top, while negative scores still included the words "worst" and "mediocre", which was joined by "horrible".

N-grams model offers the most complex method used to count adjacent words in a sentence. This can consist of a bigram (two words, such as "not working") or even trigrams (three words such as "not an issue"), which can yield very different results than measuring the words independently. After including bigrams, the largest coefficients of a positive score were "never disappointing" "amazing" and "delicious", while negative scores often used the words "at best" "worst" and "not worth". Bigrams save users from the pitfall of diametrically opposite interpretation. "Never disappointing" becomes a positive attribute, similarly "at best" and "not worth" are rightfully identified as a negative attribute.

2. According to Eric Schmidt, Chairman of Alphabet Inc (Google)

3 Translate vectors into a regression model to calculate a final star score

There are three types of regression models which can be used in this analysis. 1) a logistic regression (classification) to assign a binary outcome, such as good or bad, 2) a linear regression which can predict a specific star score, such as 4.5 stars, or 3) a neural network which is a non-linear, artificial machine learning process used to solve complex problems.

For the purposes of this paper, the Team compared results via the combination of TF-IDF and linear regression, which offered the most intuitive relationship between reviews and final star score and the ability to predict the restaurant review from 5 stars to 1 star.

Through multiple model iterations, the Team was able to train the model to analyze user reviews and come up with our predicted star score. The predicted star score was then compared to the actual star score assigned by the user. The approach developed by the Team had a high predictive power of matching the text of user reviews with their final star score.

Salad Course: Calculating & Comparing Scores

After the model was trained, the Team loaded user reviews and produced a predicted star score based solely on the written response. The three examples below show the user written review, predicted star score based upon natural language processing of that text, followed by the actual score provided by the user.

User review #1:

"My Favorite place to eat in Las Vegas! The BBQ Ribs are phenomenal, along with the $1 beers. Such a great Brewery!"

Predicted stars: 4.9
Actual stars: 5

User review #2:

"This place is mediocre at best. I have been here so many times and it was delicious. Now not good. The last four times we were here it wasn't good in anyway. The service from the moment we walked in the door was bad. They screwed up my salad and poured salt--they called it seasoning--all over it. Then the steak was way over cooked (I ordered medium). The fries were over cooked and crunchy. The water was never filled. They were terrible. I'm so disappointed. Eater beware this place is no good anymore."

Predicted stars: 1.1
Actual stars: 1

User review #3:

"First off let me just say a restaurant which only serves pb&j sandwiches is a concept doomed to failure from the start. I think their target audience is pregnant women and kids. I had the chunky peanut butter with raspberry jam and chocolate chips toasted. It was a good sandwich, but good pb&j sandwiches do not justify spending 8 bucks. I could have taken that 8 bucks, bought some bread, pb, and j, and made my own damn sandwich, just how I like it. I would have tried the PB&J BLT but I'm neither pregnant (im a guy) or adventurous enough (its weird.) The sandwich itself was awesome, like I said.... but I'll finish with saying EIGHT FRIGGIN dollars."

Predicted stars: 3.0
Actual stars: 3

The model was trained to accurately predict star ratings in the vast majority of cases – for positive, negative and neutral reviews – as outlined in the three user reviews. This provides validation of the potential to turn sentences into text, analyze text by turning it into vectors, and build a regression model which has a high predictive power. At times, content provided in reviews may differ from their actual star score, which is where things get more interesting.

The approach developed by the Team had a high predictive power of matching the text of user reviews with their final star score

Main Course: When Behavioral Biases Influence Scores

In a utopian world, the user review would be perfectly aligned with their final star score. However, this is certainly not the case with most things in life and there are numerous examples of statistically significant differences between the restaurant's written review and the star score given the dataset of nearly 5 million Yelp reviews which can be analyzed. These differences represent an opportunity, where the market is not efficiently including all available information.

The difference in predicted score from actual score can be from simple user error (such as when the user accidentally clicks on 1 star instead of 5 stars for a highly rated restaurant) or because of behavioral biases which cannot be as easily spotted.

Behavioral biases come into play when the user had certain emotional biases which were reflected in their star score, including the hope a small mom-and-pop restaurant will succeed, when the user has had a positive past experience, or there is an certain emotional affinity which overrode a logically-driven decision-making process as cited in the following example:

User review #4:

"I have great expectations, but they let me down, i order the carne asada salad, w beans, rice, tomatoes letuce and steak, but the dressing is to watery, too runny, the steak was blend and the rice mushy, also get the guac and chips, the chips way to salty, and the guac was bland, need a little lime , cilantro onions or seasoning, just taste like a mash avocados, the staff was very friendly, thats why i give the 5 stars, i hope you fix it all and for my next visit u will get the corn, (they didnt have it this time), good luck ..."

Predicted stars: 2.1
Actual stars: 5

Clearly the reviewer had their stars crossed when giving this restaurant a 5 star rating! Based upon the laundry list of let-downs, few would agree the review above is indicative of such a rating and instead the emotional impact of the staff being very friendly led to the overstatedly high review. If this review was considered in isolation, this restaurant would certainly not be among the top choices given the artificial 5 star score.

The difference in predicted score from actual score can be from simple user error... or because of behavioral biases which cannot be as easily spotted

Dessert Course (Everyone's favorite): Translating NLP to Active Investment Management

The Yelp examples offer insight into how natural language processing can be implemented to provide insights beyond that of a singular number. When the predicted and actual scores matched it suggests a high level of confidence in a particular review, whereas when scores differed it was a more interesting and represented an opportunity to find out why.

The investment world carries similar distortion of commonly used investment metrics from intended or unintended bias. The Team has devised a model which can strip away biases in an analyst's ratings, estimates and target prices using natural language processing. Instances where the actual score differs from the predicted score offer a differentiated viewpoint and represent opportunities for active management. For example, a dogmatic analyst who falls in love with the past successful investment thesis refuses to change their rating whereas the language in their report is likely to articulate their true view. Conversely, a progressive analyst may have a favorable viewpoint but hesitation to move away from peers may lead to an unchanged rating while the language in their report is likely to pick up on the masked exuberance.

Natural language processing for the first time should allow the Team to every day read through information, comprehend and capture the sentiment of ~12,000+ stocks in the investment universe. It will also open avenues to quantify a host of new pieces of information which could previously could not be measured, such as using the transcript of a management's call to identify the true conviction in their business. Other examples include the ability to predict central bank's interest rate views or calibrating consumer surveys based upon written responses. The scope of natural language processing is endless.

The Team's behavioral finance-focused investment process captures biases from Street analysts, company management and investors. Eventually, the Team envisions extending the NLP capabilities to all news flow harnessable around the planet. A common criticism of a quantitatively-focused process is the reliance on backward looking data; NLP will instead give the Team the ability to process vast information in real time. Natural language processing represents a way to take readily available information and translate it into a viewpoint which is more nuanced and refined. While numbers afford the speed and objectivity of processing information, natural language processing brings the much ignored and maligned subjective viewpoints which can be a treasure trove of future alpha opportunities.

WP-0319 | 424957 | 4207