Concepts vs Keywords

Machine Learning can be used in text analytics to classify text based on 'concepts' rather than using keywords and rules. This paper shows the differences between these two techniques with a real case study and dataset.

Contents

1.0 Overview

Everyone who uses text analytics, directly or indirectly, intellectually understands the theoretical differences between machine learning concepts versus keywords, yet there is still a disconnect in the market place between what each technique is good for and the capabilities and limitations. Which one should I use? What difference will it make? There are also some fanciful claims on proponents of both sides which only add to the confusion.

The objectives of this research were to qualify and quantify the differences between ML concepts versus keywords based around exemplary datasets. Whilst this paper was sponsored by a machine learning company, the research was conducted at arm's length using two distinct datasets and blind controls. The other text analytics companies are market leaders. For the purposes of clarification, we will be using the word 'label' to describe each count of the issue for both techniques. This is synonymous with 'tags' in other literature.

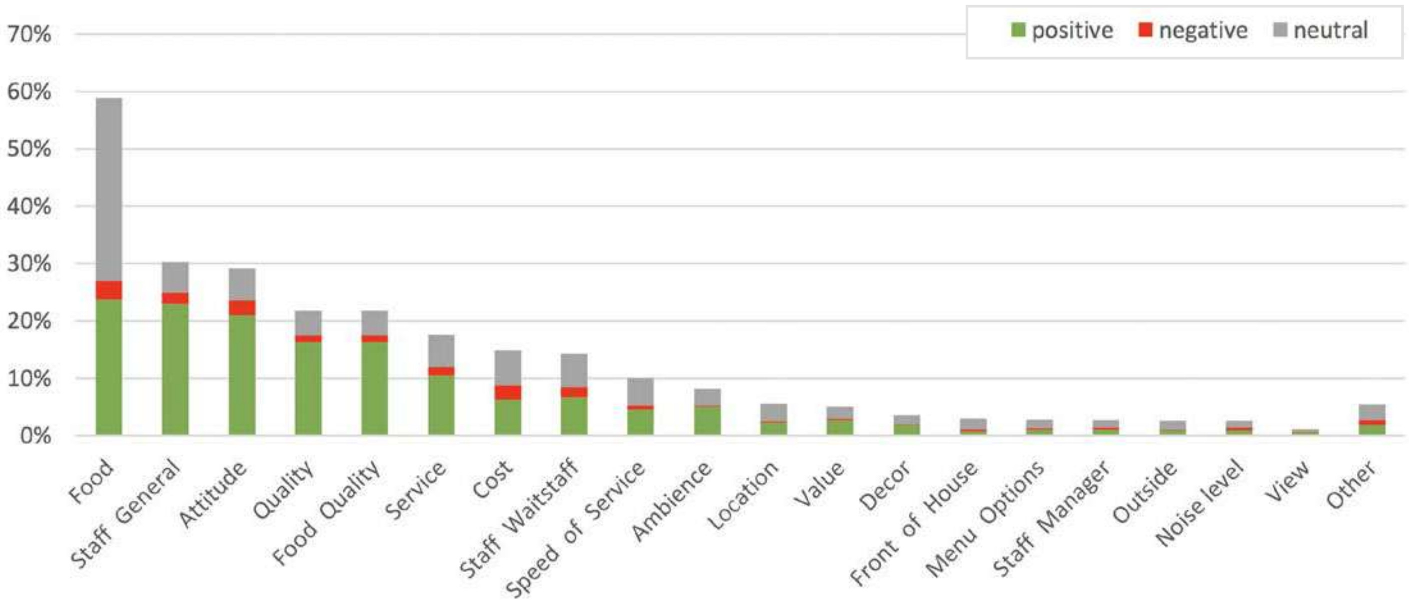
2.0 Restaurants Dataset

The first dataset chosen was a set of TripAdvisor reviews for restaurants. These were aggregated and run through a leading text analytics company using its rules and keywords, and sentiment analysis. By comparison we used some software called PrediCX (from C-Centric) to quickly generate ML concepts. The reason for the dataset was that both the keywords text analytics company, and the ML concepts company had an industry-specific model relating to restaurants. We comment on the results below.

2.1 High Level Comparison

Below we show the results from comparing both techniques at ‘First Level’ i.e. the highest level of any hierarchy. For the keywords, you can see below in **Figure 1** how these top categories break down by volume and sentiment.

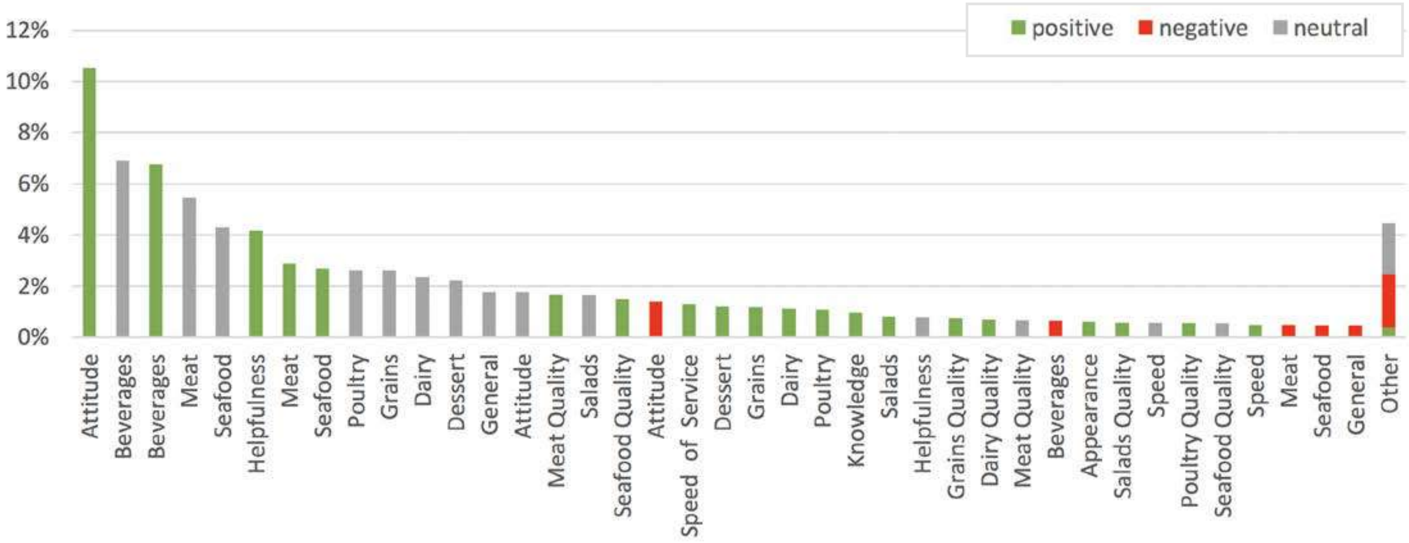
Figure 1. First Level Keyword Analysis of TripAdvisor Reviews



Some comments are that obviously Food is the main one with the most labels, and with most of the commentary being neutral followed by positive. You can then start to see some unclear things, e.g. what is the distinction between “Staff_General”, “Attitude”, “Service”, “Staff_Waitstaff”? If you make a positive comment for one, don’t you intrinsically make a comment for another? Similarly, what about “Quality”, “Food_Quality” and “Food”? You can resolve this initially by suggesting that these labels can be collapsed into respective hierarchies of your choosing but it’s not as simple as that, as the labels don’t trigger in a consistent way e.g. “Staff_Waitstaff” appears 18% of the time that “Staff_General” appears (begging the question on who the other 82% of the staff referred are) but it also appears another just over the amount equivalent of the 18% elsewhere when “Staff_General” doesn’t appear at all. You can of course choose to do some post-processing and filtering, although when you start to read the reviews and take judgements it can become quite convoluted quite quickly.

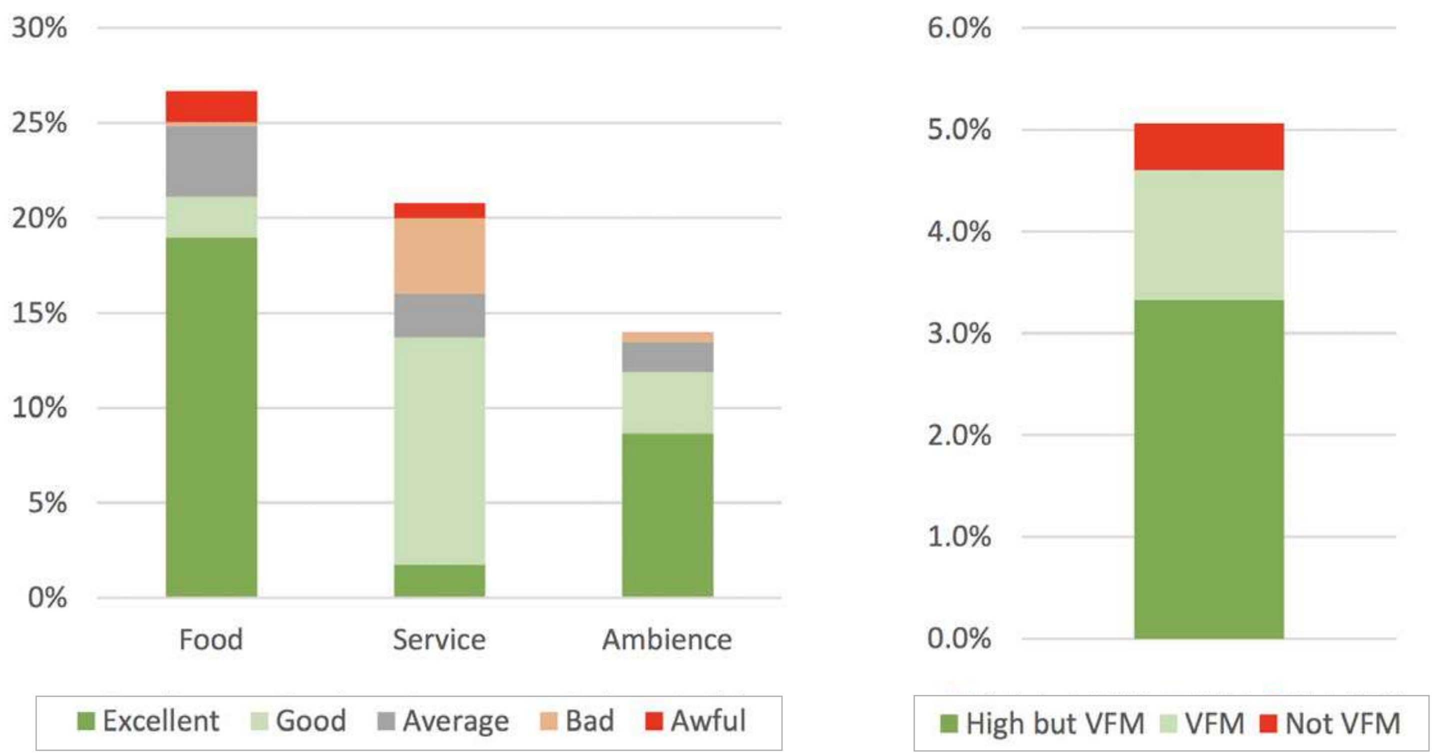
The second level analysis is where the text analytics provider has listed their taxonomy as nested within the first level above. We’ve plotted these below as a non-stacked histogram with each sentiment label separated, which will become easier for comparison later.

Figure 2. Second Level Keyword Analysis of TripAdvisor Reviews



Commenting here, the overwhelming labels are “Attitude” and “Beverages” if you add together the positive and neutral labels (second and third from the left). It then lists some types of food and other attributes of the experience. We will return to this below as we examine the ML Concepts analysis.

Figure 3. First Level ML Concepts Analysis of TripAdvisor Reviews



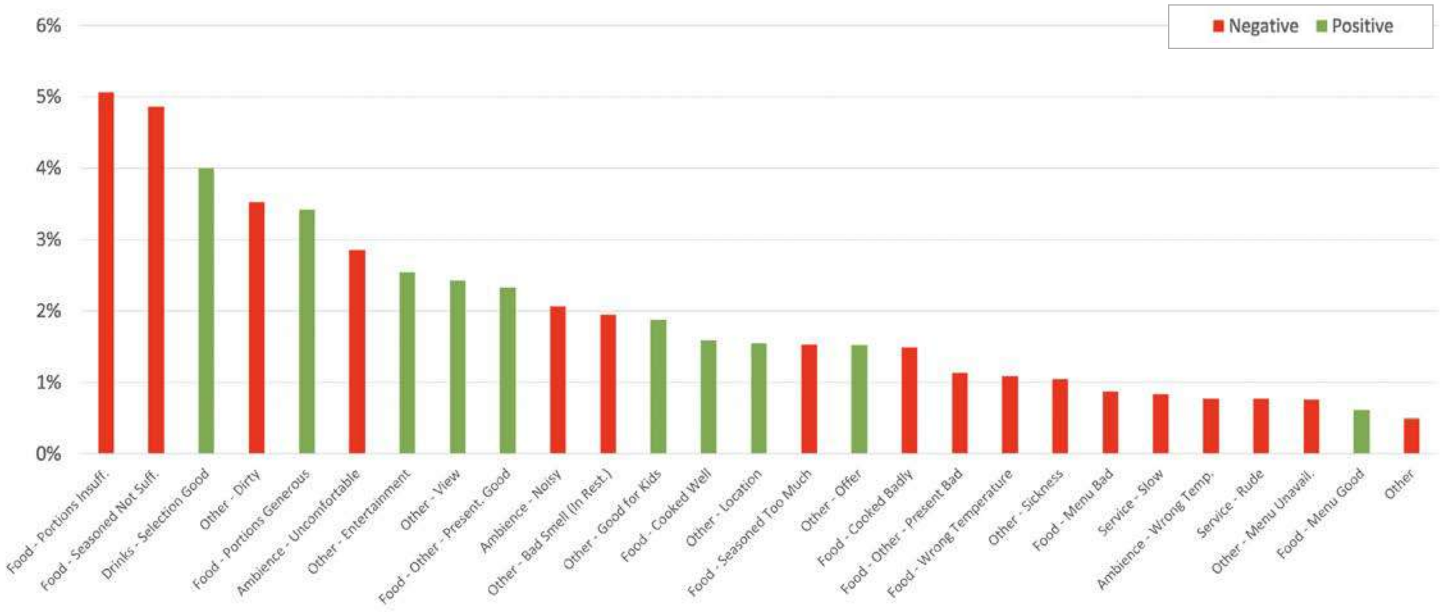
The first observation is that there are two graphs as it’s not possible to add the figure at right i.e. “Value for Money” directly onto the high-level concepts at left as they are not equivalent, i.e. “High but VFM” refers to comments where customers say they thought the meal was value for money even though it was high. This is useful, actionable information and a strong signal, but it is not equivalent to the overall opinion terms at left.

The second observation is that the graph at left shows a scale of five opinion terms. It is notable that “Average” here is not equivalent to “Neutral” in the keyword analysis above. “Average” here refers to terms such as “OK” and “fine” whereas “Neutral” picks up mentions of the word and synonyms of “Food” whether or not the word has an opinion term associated. In the sentence: “We asked again when our food would arrive as we had waited over an hour”, the count of the word “food” doesn’t offer a lot of meaning, neutral or otherwise. The meaning here is that the service was slow. In ML concepts, it won’t count this unless customers are talking about the food, not just mentions. We will also refer to a discussion of sentiment analysis comparison later on.

The third observation is that the hierarchy is well-structured i.e. all the relevant second-level concepts about food can be nested within the first level concept of “Food”. Actually in reviews, it is often the case that first level concepts are mentioned on their own, e.g. “the food was fantastic” and it may not give any second level information why e.g. “the chicken was so tasty”. You can then choose to split out first level concepts from second level, which can be useful to prevent double-counting for sentiment purposes. For the avoidance of doubt, the graphs shown above are in this mode.

“You can split out first level concepts from second level, which can be useful to prevent double-counting for sentiment purposes.”

Figure 4. Second Level ML Concepts Analysis of TripAdvisor Reviews



We can compare visually the second level of labels from ML concepts versus the keywords above in Figure 2. In the keywords, the common topics are “Attitude” (positive) and “Beverage” (positive and negative). It is difficult to proscribe a direct action associated with this information, particularly without further analysis and reading some of the reviews first-hand. With the ML concepts, you can see that customers are mostly concerned with the food portions being too small and the food not being seasoned enough i.e. being too bland. These have direct actions associated and no further reading of reviews is required to do this. There is also a holistic observation in perusing the dataset which is that customers often use quick superlatives when saying something is positive: “the restaurant was great” whereas they provide more information in a negative situation. Heuristically this supports Bill Gates’ famous quote: “Your most unhappy customers are your greatest source of learning”. Note the stark contrast of the sea of green (Figure 2) versus the sea of red (Figure 4).

2.2 Intents

Reviews are potentially a rich source of insight both in terms of the opinion of the reviewer as well as their intents, i.e. whether they would recommend it or not, and whether they would return. This is quite interesting in itself as well as secondary analysis, i.e. trying to get a handle on root cause of why people have those intents.

Below, we lay out four charts for intents for both keywords and ML concepts. There are two further charts in the Appendix.

Figure 5. Keywords: “Recommend” Intents (all Levels)

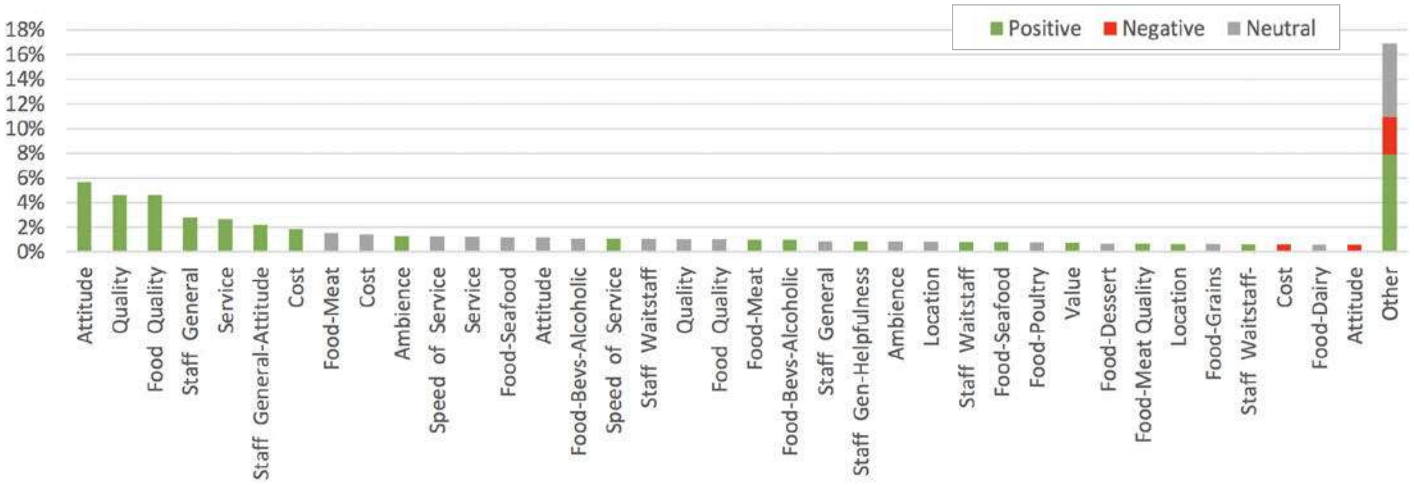


Figure 6. Keywords: “Quit” Intents (all Levels)

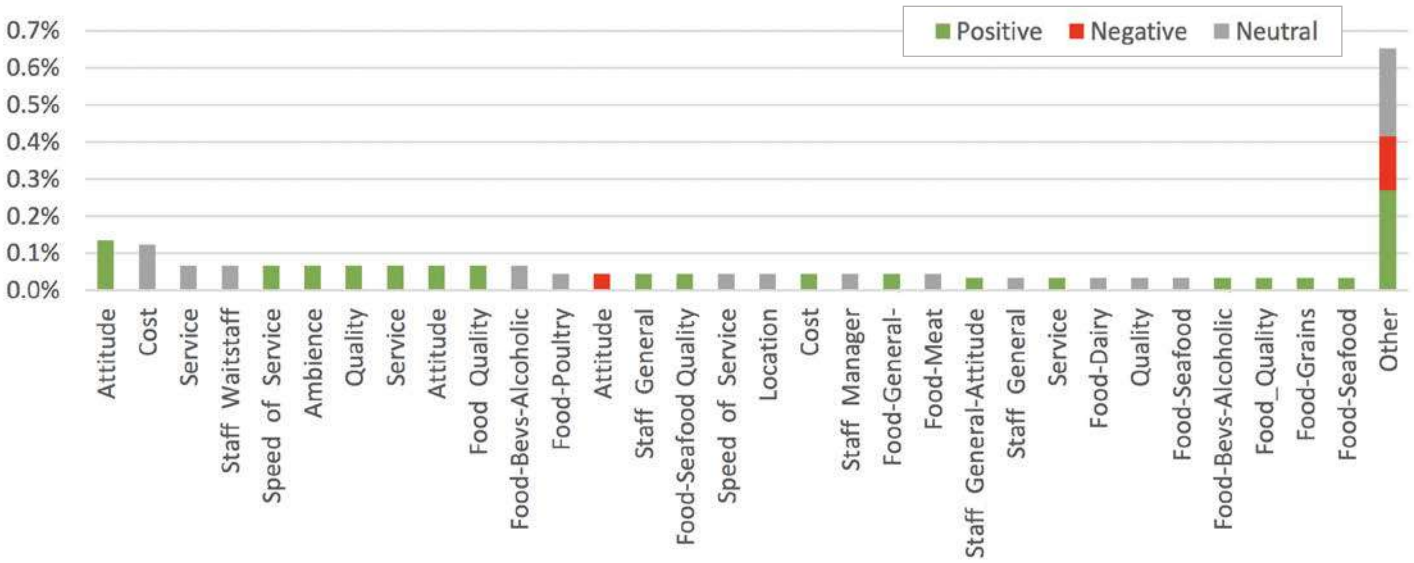


Figure 7. ML Concepts: “Will Return” Intents (all Levels)

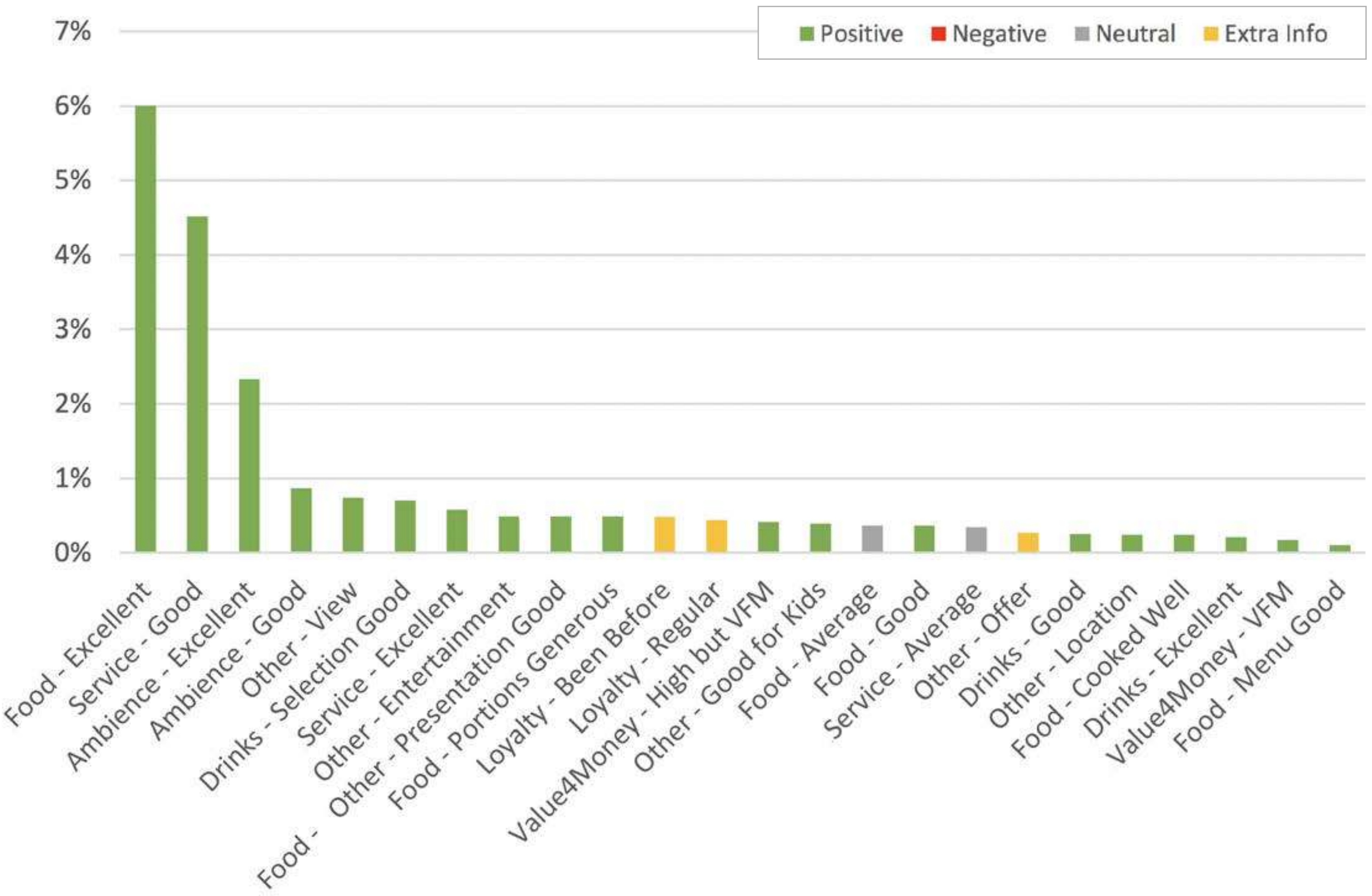
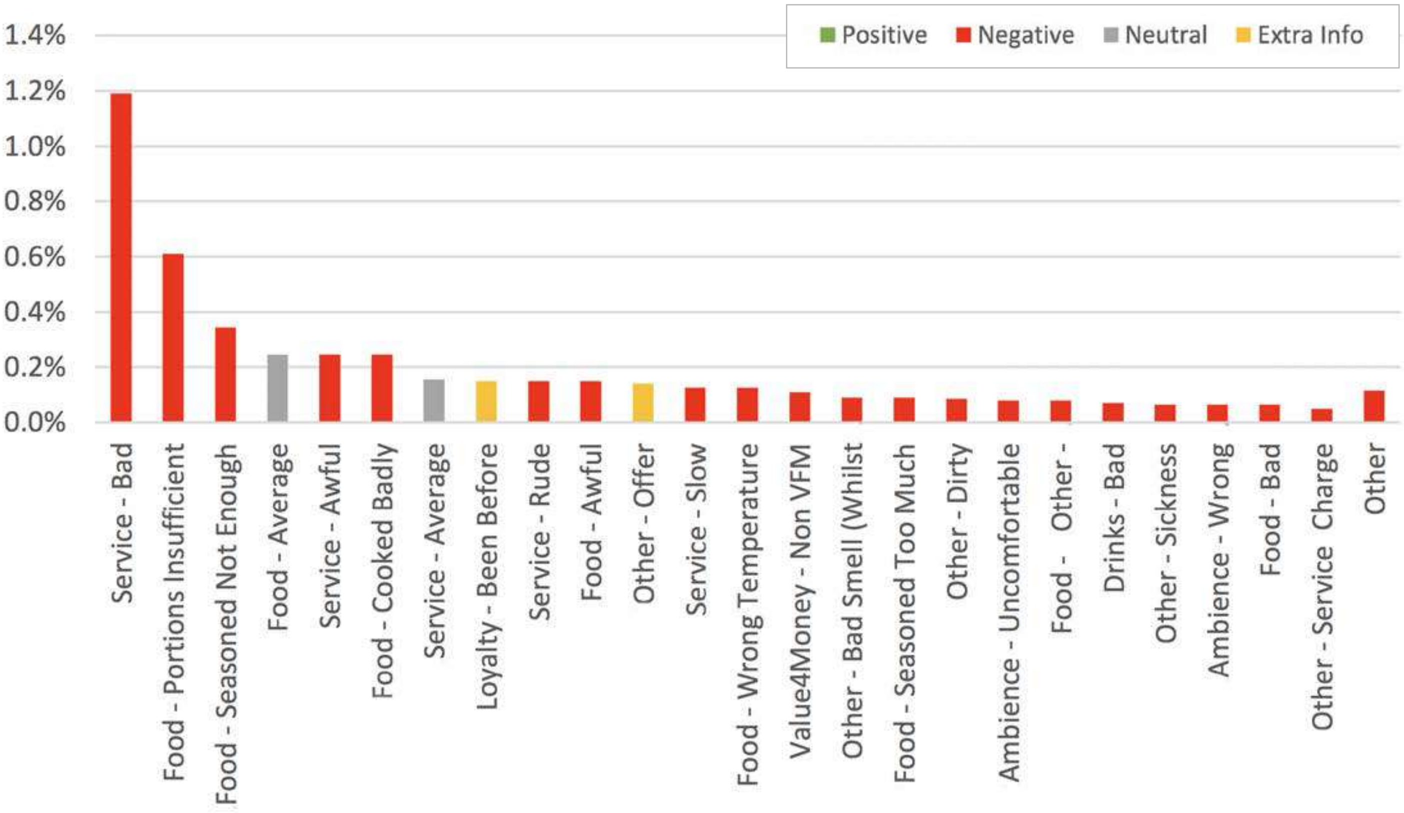


Figure 8. ML Concepts: “Won’t Return” Intents (all Levels)



Hopefully the comparisons between keywords and ML concepts here are clear. We spell out the main observations:

- The charts for keywords are mostly green, even for “Quits” whereas the ML concepts for “Won’t Return” are mostly red. The reasons for this are to do with accuracy (detailed further below) as well as the fundamental way that these two approaches are working. With keywords, for “Quits” you get less of the positive mentions of the dominating labels whereas for ML Concepts “Won’t Return” you actually get the concepts that get labelled and you get a clearer, more actionable picture. From ML concepts it is clear that at a high level, excellent food and ambience drives positive recommendations (more than excellent service) but poor service drives negative ones. At lower levels, a good view and entertainment can make a positive different whilst vice versa for small portions and bland food. For the keywords it is difficult to come to such clear and specific conclusions and there is a huge tail of ‘Other’.
- The other observation is that there are different intents from the ones listed that can be discerned. For ML concepts, we have “Will Return”, “Won’t Return”, “Will Recommend”, “Won’t Recommend”, “Regular Customers”, “Been Before”. For keywords there’s just two. In Appendix 1, we show that there are some subtle differences in actionable insight even between “Will Return” and “Will Recommend”: “Generous portions” and “High Price, but VFM” drives “Will Recommend” more, whereas the “View” and “Entertainment” are higher for “Will Return”. Also “Will Return” was mentioned significantly more for “Will Recommend” than the other way around.

Overall you can see how ML concepts differ fundamentally and practically, particularly for secondary analysis. If you were to substitute the intent for a star rating they also gave, or the price/tip they paid, then you can see that ML concepts would be able to yield actionable and predictive insight whereas keywords wouldn’t.

“You can see that concepts would be able to yield predictive insight whereas keywords wouldn’t.”

2.3 Commentary on Sentiment Analysis

As mentioned above, the sentiment analysis for ML concepts is explicit and simple to grasp, i.e. each concept has an associated sentiment, graded or otherwise. You can then do what you like with this e.g. tot them all up, filter for certain topics or whatever. If you get the sentence: “the food was great but the service was terrible and I’m never going back” then you don’t get into a spiral of averaging the sentiment (unless you want to). You have one marker of positive, one of negative and a clear negative intent. Also if the fish was cold but the ice cream was warm then they are both negative because the terms “cool” and “warm” have opposing sentiment depending on the context.

However the sentiment for the keyword text analytics provider (and indeed similar text analytics providers across the board) isn’t straightforward and the pitfalls are often overlooked by users. Essentially it is driven by a black-box algorithm which uses various techniques to try to match the opinion terms close to the keyword terms (from -1 to 1) and then combines them in a way within sentences. The procedure is highly technical, and this provider referred to “log odds ratios” and “lexical chaining”. Many companies use their literature to promote this black-box approach as a secret-sauce and a benefit. The authors of this report wholeheartedly disagree that any lack of transparency is a good thing, as well as the principles of Occam’s Razor.

For the quoted sentence referred above you might get what you expect or not depending on the algorithm and other sentences around. However the way people talk about things in reality, with sarcasm, negatives and context, it can present a lot of challenges that many people who use sentiment analysis overlook. The claimed performances can be wildly optimistic and indeed subjective. Performance figures are provided in the section below. Some examples of where sentiment can go quite wrong are presented in Appendix 2.

2.4 Performance

Before we delve into the actuals, it is worth talking about performance. Firstly we can see that it’s not just about accuracy, there are qualitative differences as well as quantitative.

Secondly, the nature on how these two techniques, or any technique, performs depends ultimately on measuring the labels by a human who understands the domain. As Boris Elveson of Forrester said: “Don’t take vendor claims about product accuracy at face value. The only way to verify accuracy is for a human to manually mark up sample data sets”. The performance can change depending on the dataset and indeed the analysis and measurer. We performed this critique arms-length for this TripAdvisor dataset and also for Twitter messages as well critiquing against a different text analytics company (we haven’t spelled this out in this paper but we present the performance below too).

Also performance is often misunderstood and indeed in the author’s view, obfuscated by other partisan commentators. Without a detailed commentary, a key point is to distinguish “precision rate” from “recall rate”. The precise definition is presented in Appendix 3, but essentially the former is about false positives and the latter is false negatives. Put bluntly, commentators often refer to accuracy when they mean “precision rate” and the consequences of this is that there’s typically a large pool of records unlabelled i.e. appearing in “Other” as well as more difficult to spot where records are labelled but missing certain labels because the taxonomy doesn’t include them, for example if someone says “my coq au vin was cold” then the lexicon of the provider needs to have “coq au vin” listed, and indeed typos of the same, otherwise it will be missed and a false negative. It’s easy to see that reporting a high “precision rate” as accurate alone could be highly misleading, particularly if new things, or new expressions appear in the market. This could mean that not only is accuracy low but it is unknown what the accuracy is.

Performance of Sentiment Analysis

The precision rate for the keyword labels was 58%. This was compared to 76% for ML concepts.

The recall rate for keywords was unknown. There were 46% of records which were completely unlabelled. For ML concepts the recall rate was explicitly known and it was 51%. There were no records unlabelled.

> “The precision rate for the keyword labels was 58% compared to 76% for concepts.”

By observation, it’s clear that measuring the recall rate is a problem for keyword analysis.

Performance of Intents

On the intents, the keyword precision was 68% although it was 54% if you took into account the difference between recommendation and saying they would return. There were also a lot of instances of inconsistency e.g. where both "recommending" and "quitting" were predicted together as well as duplication by 14% i.e. double counting.

Estimating the recall, the keywords picked up intents in 23% of records for this dataset, albeit with the accuracy noted above, so less in fact around 13% to 16%.

ML concepts picked up intents 28% of the time. The precision was 88% and recall 45%.

Some observations:

- For keywords it labelled 23% of records as "recommending" but less than 1% as "quitting". For ML concepts, it labelled positive intents 23% and negative intents as 5%. This meant that it was significantly better (1000%) at picking up negative intents and arguably these carry the most insight.
- The low recall was because of the relatively small occurrences of intents, particularly negative ones.

Performance Commentary for a Twitter Feed

Elsewhere, a comparison of keywords versus ML concepts was made for UK telcos, in other words, customers enquiring and commenting to their telco provider. Again this was against industry-tuned models on both counts specifically for telcos. Without rehearsing the above TripAdvisor, the overarching point was that the key insights were missed by keywords but picked up in the ML concepts. This is covered in a case study elsewhere.

> "On Tweets the precision rate was 38% for keywords versus 70% for concepts."

In terms of precision rate and recall rate they followed the same pattern as TripAdvisor, i.e. 38% for keywords versus 70% for concepts. There were 11% of the Tweets in the 'Other' category compared to zero with ML concepts.

Lastly, and slightly differently to the TripAdvisor case, the average number of labels per record was 2.3 with keywords but only 1.05 with ML concepts (for TripAdvisor both were about 5 each). This is important as a Tweet usually has one intent and so keywords become less useful for picking this up as much of the verbiage will be context.

3.0 Conclusions on Performance

Many observations have already been made. The conclusions are that performance for keywords is driven by the a priori lexicon and taxonomy. Performance isn't explicit and has to be manually measured. The performance for ML concepts hasn't been discussed in great detail here but relies on an algorithm and some training data applying to the domain concerned. It can be tuned and may or may not require intervention from a data scientist. We note only for the purposes of completeness, that PrediCX was the software used here and the performance shown was achieved by a human-in-the-loop i.e. a non-data scientist spending two man-days from scratch to generate the machine learning models. As part of the human-in-the-loop process, the performance is explicit as it invites human intervention to train the models by requesting validation on records where it has high uncertainty. There is always scope to tune and develop models further.

When asking what is 'good' performance, this is a harder question to answer than it first appears. One benchmark is to compare against humans and this can vary widely depending on the data, domain and appropriateness. It can range anywhere from 70% to 90% precision with recall less well known but much lower (it's easier to miss things than to correctly or incorrectly judge something). Another benchmark is to have a reliable self-measurement metric, noting that some labels will have better performance than others, most likely the larger classes and/or those with less variable description possibilities.

4.0 Overall Conclusions

There are a range of options for analysing text. We have endeavoured to make an objective comparison of keywords with ML concepts around two different datasets. The benefits of keywords are that it's intrinsically easy to conceptualise a lexicon. The benefits of ML concepts is that they can provide actionable and accurate insight. There are also providers out there which obviate the work and technical skill level required to get the most out of ML concepts.

Appendix 1

Figure 9. ML Concepts: "Will Recommend" Intents (all Levels)

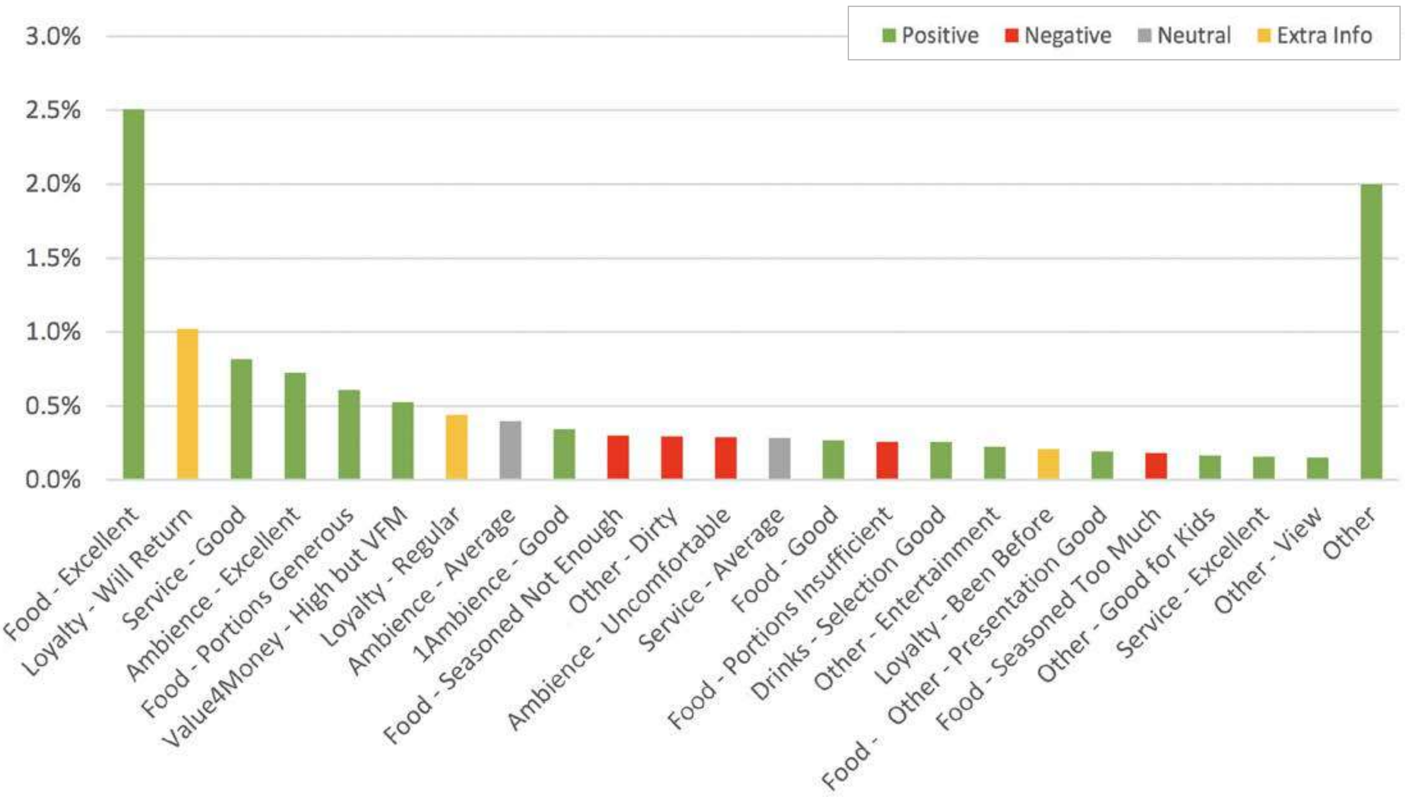
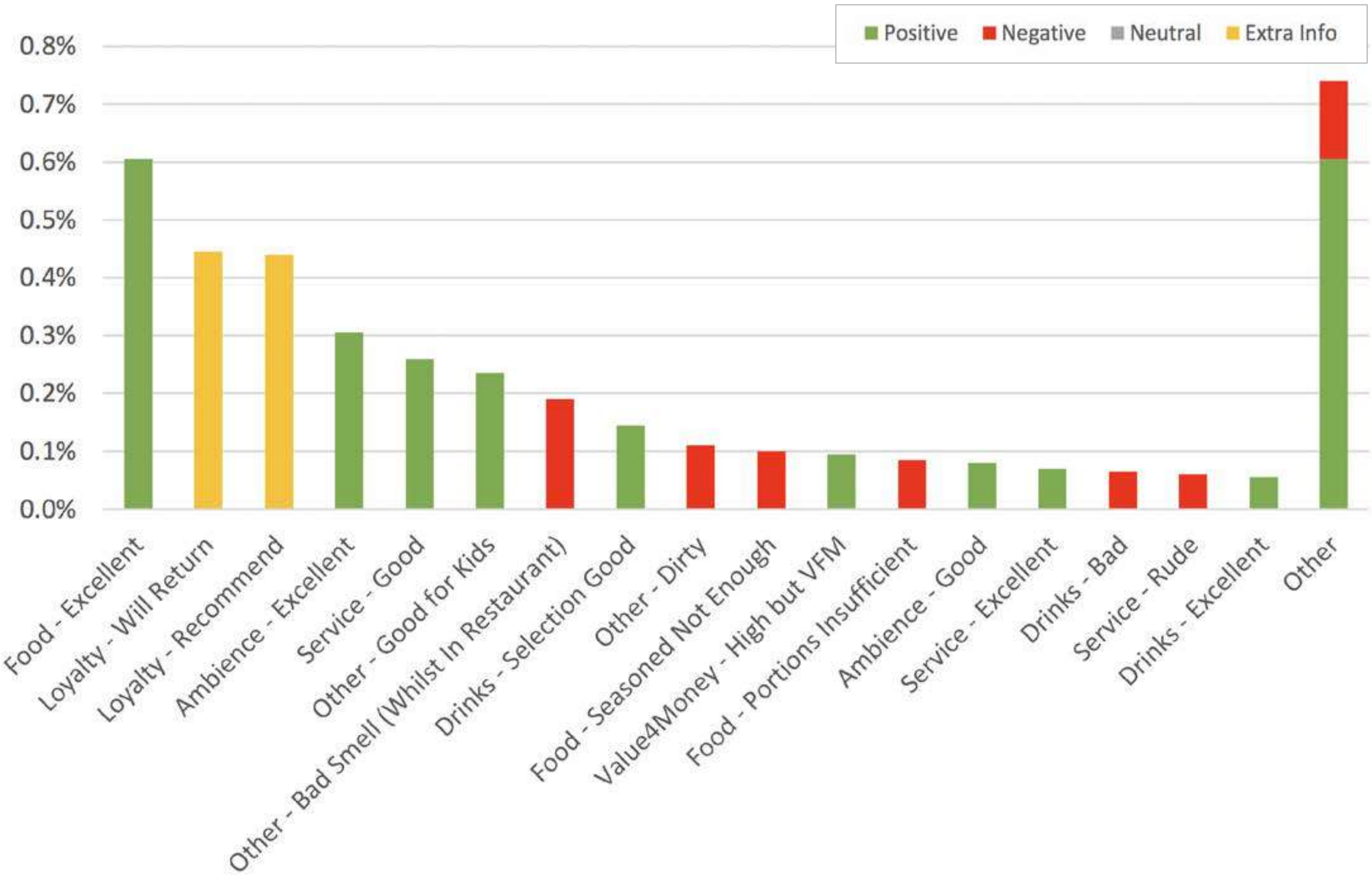


Figure 10. ML Concepts: "Regulars" Intents (all Levels)



Appendix 2

Below is a sample of reviews and the keywords picked up:

"However unfortunately Sticky Mango did not light our fire! Our table was squeezed into a corner with just enough room to squeeze in and out yet close enough to feel like you were dining with the adjacent table. We chose the set menu and although a couple of dishes we superb generally they were mediocre. We chose the tasting menu as the al a carte was limited. Our quest for a good London Thai continues... PS pricey too"

Label	Comment
Menu_Options (negative)	Yes (assuming the second mention)
Menu_Options (neutral)	Yes … but the first menu mention just context
Cost (neutral)	No, it was negative
Attitude (positive)	Irrelevant
Food_Quality (positive)	No, although there was a mix of opinions
Speed_of_Service (neutral)	Not mentioned
Service (neutral)	Not mentioned
Quality (positive)	Ditto to Food Quality above. Duplicative
	No label was picked up for being crammed in

"Very nice for a healthy and quick snack, you have good salads, good sandwiches and fresh juices."

Label	Comment
Food-Salads (neutral)	No, positive
Attitude (positive)	Not mentioned
Food-Salads Quality (positive)	Yes (albeit duplicated)
Food_Quality (positive)	Yes
Speed_of_Service (positive)	Not relevant … quick snack is not quick service
Quality (positive)	Yes
	No label for "healthy", "sandwiches", "snacks" or "juices". Just "salad"?

"Maybe it all depends on the food you ask for, the dish, the day, the hand of the chef that day. I ate only once, so, my first impression, though very good was not excellent, and in the end I was a trifle discontented with the restaurant: maybe it was because I went for the pork's belly, a dish not always easy to cook and even more difficult to chew on (lol). The rind was really difficult to eat and digest. There were excellent surprises though: the rest of the meat was tasty, very well seasoned (anise included), wish mashed potatoes seasoned with watercress, and sour cherry's sauce. The dessert won it all: a black forest cake with custard, whipped cream and morello sauce. The staff were kind, available and surprising attentive. The wine recommended was a must, so we were quite happy but the way the rind of the pork belly was cooked, if anyone can cook it well. Finally last but not the least, the restaurant is right in the middle of Covent Garden's market, providing for a jolly, lively and beautiful atmosphere. We'll be back some day, just to taste the rest!"

Label	Comment
Food-Dessert (neutral)	No, the dessert "won it all" very positive
Attitude (neutral)	No, lots of relevant staff opinions
Ambience (positive)	Yes ("beautiful atmosphere")
Food-Meat (neutral)	?? There's two references to meat, one neutral and one positive
Staff_Waitstaff (neutral)	No, positive
Food_Quality (neutral)	?? It's not clear – there's positive and negative. Overall they are coming back to taste the rest so positive but says "trifle discontented"
Staff_General-Helpfulness (positive)	Yes (contradicts "Staff_Waitstaff" above)
Staff_General-Attitude (positive)	Yes (ditto)
Food-Beverages-Alcoholic (neutral)	No, the wine "was a must"
Staff_General (neutral)	No (see above)
Quality (neutral)	No (see above)
	Location wasn't labelled

Appendix 3

Precision Rate is “how many selected items are true” as Recall Rate is “how many true items are selected”. Precision Rate can sometimes be thought of as the absence of false positives i.e. if you are blindfolded and walk around a farm guessing the animals from feeling it, you might guess pigs 90% accurately and that would be Precision Rate. Recall Rate is the absence of false negatives i.e. you walk around the farm blindfolded and guess 90% all the pigs correctly, but you don’t locate all the pigs and miss out 50% of them. Therefore you have a 90% Precision and 50% Recall.

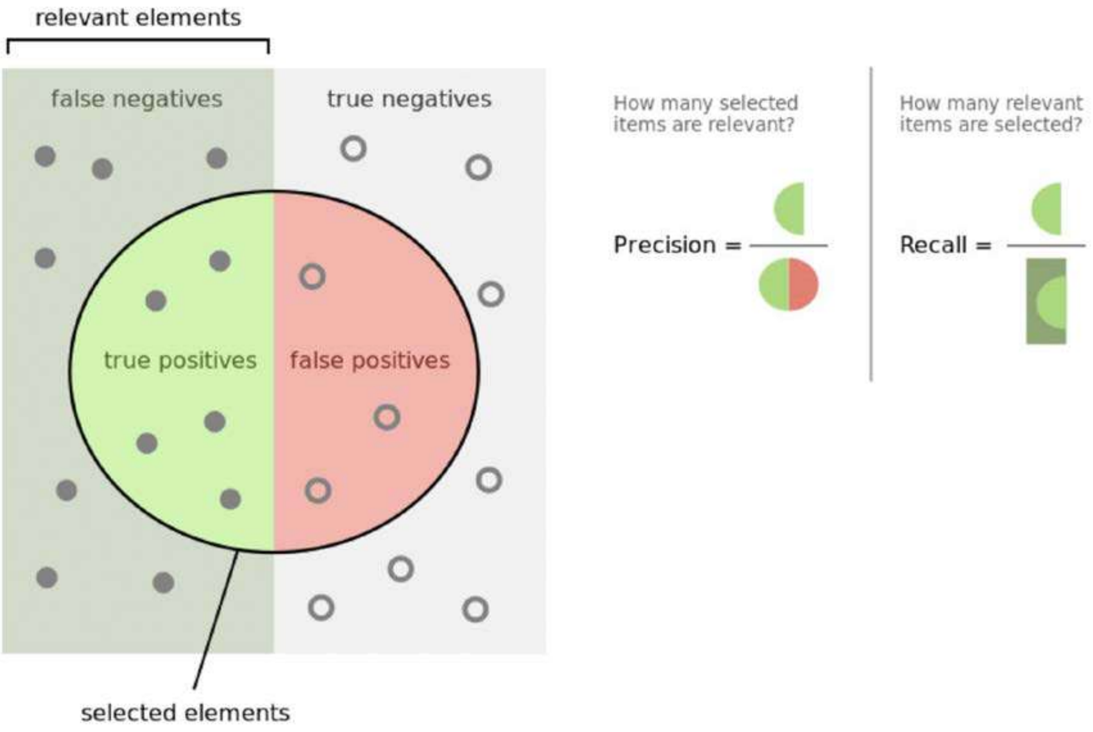
So here (tp = true positive and fn = false negative etc.):

$$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{Recall} = \frac{tp}{tp + fn}$$

Accuracy in its strictest data science definition is this context is a combination of the two

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

There is a helpful illustration for Precision and Recall below in the diagram (courtesy to Wikipedia):


relevant elements
false negatives
true negatives
true positives
false positives
selected elements
How many selected items are relevant?
Precision =
How many relevant items are selected?
Recall =