Identifying online display of Food Hygiene Rating Scheme ratings: Assessing online display

Methods and challenges

Once a sample of websites had been generated, these were passed to an image matching pipeline. This pipeline obtains the underlying code for the website, identifies the images within the code, extracts them, and applies image matching algorithms to determine whether each image is an FHRS rating. It also travels to each page that is one ‘click’ into the website, and evaluates the images it finds on each of those pages. All images are downloaded temporarily while the website is being analysed, and then deleted at the end of this process. Only the URL of the closest matching image is retained.

The matching algorithm compares each image against a set of reference images – i.e. those that are known to be FHRS ratings. Four versions of each rating were used, encompassing both the digital and ‘window sticker’ formats, and including a bilingual version.

Each website image received a score from 0 to 100, representing the confidence of the match, against each reference image. The highest scoring images from each website were compiled, and all images scoring higher than a particular threshold (in this case 30) were examined manually to establish whether they were ratings or not.

A key challenge in implementing this process across a large sample was the time taken by the image matching algorithm. These issues were addressed by:

- adding a quicker initial matching step to eliminate all but the most likely matches
- restricting the number of images that can be sent to the slower, more powerful matching algorithm (at a cost of potentially missing a rating image in a very large website – our quality checks suggested this is unlikely to affect many websites however)
- utilising the computer’s processing capacity more efficiently, allowing multiple matching processes to occur at the same time
- once the pipeline was fully developed, running the full sample through it on a slightly more expensive, but more powerful computer

These innovations resulted in a process that can be run end to end within a day.

Results

Full sample

Analysis of the full sample suggested that the prevalence of online display of FHRS ratings is around 3%. This is the proportion of businesses with an online presence that are displaying a

rating. It represents around 1.7% of all in scope businesses (as only around half were found to have an online presence).

The matching process identified 21 websites in the sample that were displaying an FHRS rating. We believe based on our subsequent quality assurance checks (see next section) that this represents about 80% of what is truly present. Therefore the final total is likely to be around 26 websites, or 3% of the sample of 803.

Of the 21 websites that were uncovered, half were establishments in England, with a quarter each in Wales and Northern Ireland respectively. As the nations were roughly equally represented in the sample, this suggests that online display is disproportionately common in England.

Table 4 shows the proportion of each establishment in the original sample, compared to the 21 establishments that were found. Restaurants and hotels were represented in the same proportion as they were in the sample, with pubs under-represented, and takeaways over-represented.

Table 4: Online display by business type

	% in sample	% in found
Hotel/bed & breakfast/guest house?	12	10
Pub/bar/nightclub?	18	5
Restaurant/Cafe/Canteen	21	33
Takeaway/sandwich shop	12	10

All the websites were displaying a 5 rating, apart from one, which turned out not to be the business's own website. A comparison with the actual ratings on ratings.food.gov.uk found that two of these businesses were rated 4, despite their professed dedication to hygiene.

Quality assurance

There are two types of error in using a predictive algorithm; it predicts that an image is a rating when in fact it is not (a 'false positive'), or it predicts that an image is not a rating when in fact it is (a 'false negative'). The only way to truly know these numbers would be to go through every website in the sample manually; the outcome we are trying to avoid. However, the analysis of smaller sub-samples can help us to estimate the extent of these two error types, and therefore how much confidence we should have in our results.

False positives

The image matching process does not make a 'yes' or 'no' decision about whether the image is an FHRS rating; it assigns a score between 0 and 100, where 0 is definitely not a match and 100 is a definite match. Therefore a decision needs to be taken by the analyst of what threshold should be taken to imply a (possible) match. For this analysis, a threshold of 30 was used.

However, this does not mean that every image above this threshold was an FHRS rating. The analysis returned 48 unique images, across 66 websites, above this threshold. Each of these images was examined manually to determine whether it was a rating. 27 of these were not ratings; they were 'false positives' in this process. This is quite a high false positive rate of around two thirds.

The false positive rate can be reduced by increasing the threshold score. For example, raising this threshold to a matching score of 50 would reduce the number of false positives to just 5. However, the disadvantage is that it increases the false negative rate; in this case it would only have identified 18 of the true images instead of 21. Therefore, given the fairly low prevalence of

high or even moderate scores (only 8%, or 66 websites in this case, scored over 30), it is better to set the threshold low and manually review. It takes considerably less time to run the analysis and review 48 images (1 day) than it does to review 800 websites (which could take several weeks), so a great deal of time could be saved.

It is difficult to say why an image matching algorithm has assigned the score it has. However, an examination of some of the false positive images suggests that circular features within an image – like the circles around the numbers in a rating image – may be tricking the algorithm.

False negatives

False negatives are harder to quantify, as you do not know the true extent to which images are displayed online. However, we used two approaches to try and quantify the false negative rate.

The first was to look at a sample of 100 sites to which the algorithm had assigned a very low score (below 15), or for which it had been unable to obtain images or calculate a score. Of these, 2 were in fact displaying a rating. In both cases, the scraper and image matching pipeline had run successfully, but had failed to identify the FHRS image. In one case this was because the image was bundled in with others, and in the other because the image was not of a type that the image matching pipeline could process. In 72 cases the pipeline had worked as expected, but the image was not present. Of the remaining 24 that did not work, 8 of the links were broken or inaccessible, and in the remaining cases, the scraper was unable to retrieve the images. Therefore we could say that the scraping and matching pipeline failed for 18 of the 92 working links, or about 20% of the time.

The matching pipeline was also run for an additional sample of 50 websites that had already been identified as displaying a rating; these were discovered through reverse Google image search or by chance. In 40 of these, the correct image was identified and scored above the threshold (it was detected in a further 2 cases, but with a score below the threshold). Again this suggests that the pipeline is failing around 20% of the time.

Taken together, these suggest that the matching process is likely to be detecting around 80% of the websites displaying FHRS images, and there is unlikely to be widespread online display beyond what we have been able to estimate here.

Factors affecting the score

The score assigned to an image is in part beyond our control; these algorithms are complex, externally developed libraries of code. There is therefore little that can be done about false positives; the algorithm has decided that the Instagram logo looks like the FHRS one, and it cannot be trained to believe otherwise. However, there are a number of factors around image quality that seem to result in the algorithm giving an undesirably low score to images that are actually ratings.

The highest scoring images received an almost-certain score of 98.9. They were universally high quality images of the digital version of the badge. The ‘window sticker’ style images are also capable of receiving a high score when they are good quality, which received a score of 98, but a poor quality version scored just shy of the threshold.

The matching algorithm also found it much easier to match standardised images. Images that did not pass the threshold included: one website that had made its own version; another that had a photo of a sticker in a window (although elsewhere these were picked up if they were clear, straight on photos); a website that took a screenshot from ratings.food.gov.uk (with too much surrounding noise); and a rating image that was bundled together with other graphics in a single image.