

# Classifying brands on Facebook using supervised machine learning



General Assembly DSI : Capstone Project

by Sam Ho  
[sam@samho.co.uk](mailto:sam@samho.co.uk)

# The Agenda : ‘CADET’

Context

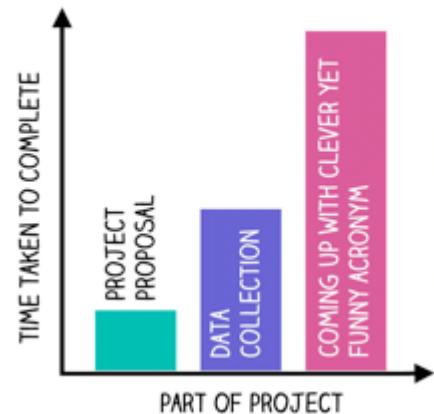
Aims

Data

EDA | Modelling

Implications

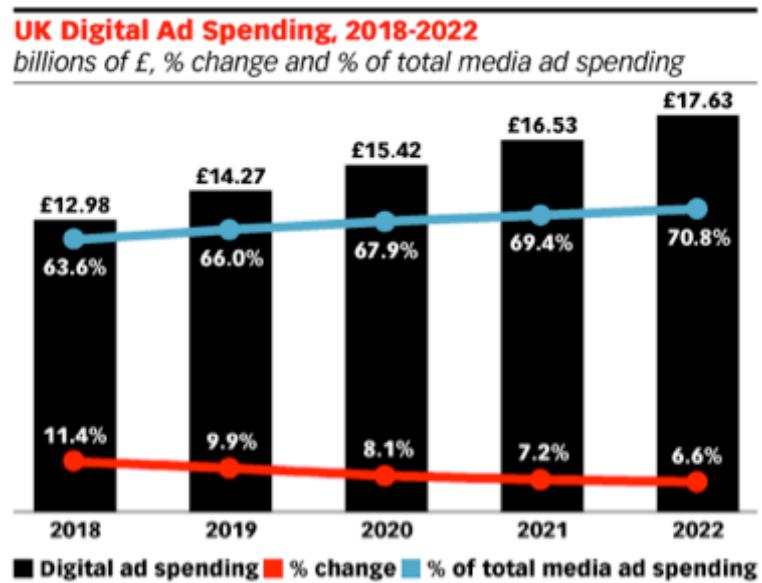
HOW SCIENTISTS SPEND THEIR TIME



# CONTEXT



In 2018, almost two thirds of all advertising spend was on digital



# Why does this matter?



More spend =

...more content  
...more noise



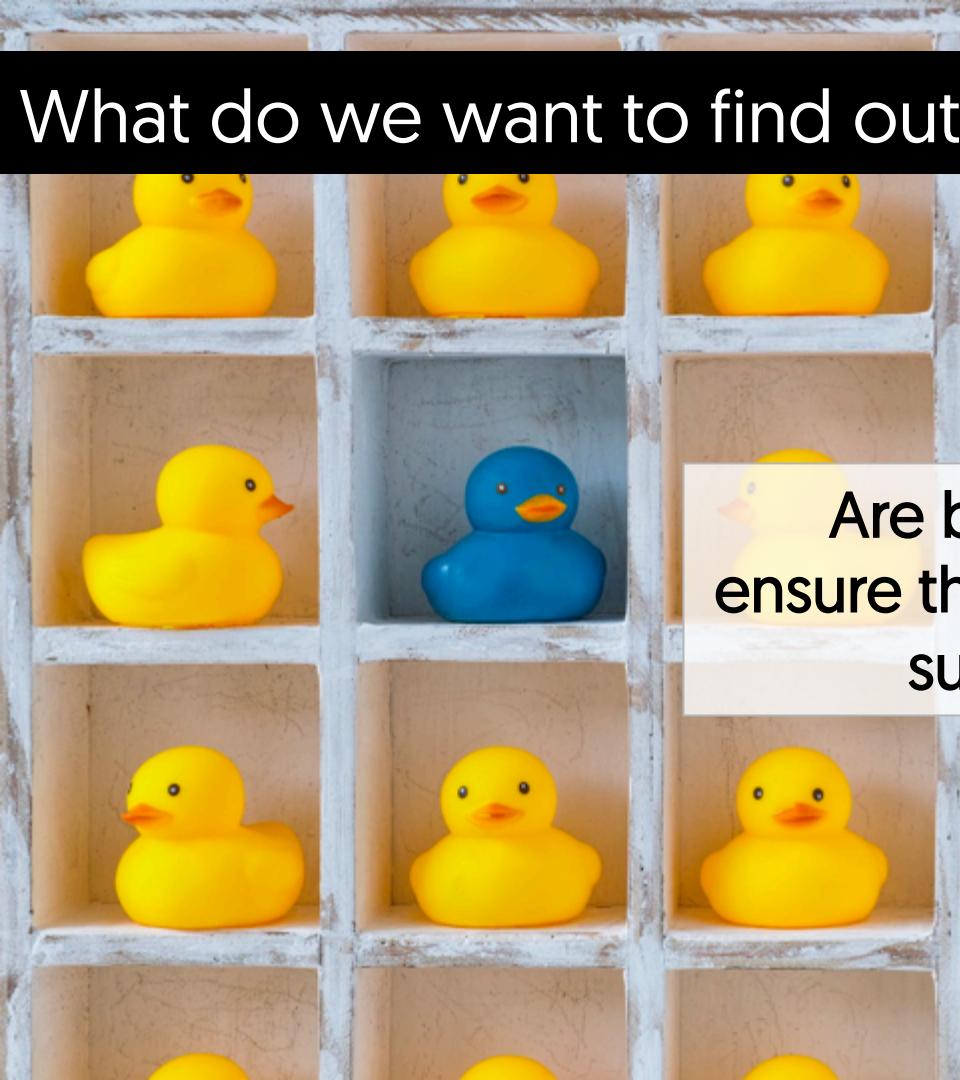
Increasing pressure to

...stand out  
...to differentiate

The background image shows a wide-angle aerial view of the City of London's financial district during sunset. The sky is a warm orange and yellow. Several iconic skyscrapers are visible, including the Gherkin (Swiss Re Building), the Walkie-Talkie, and the Cheesegrater (20 Fenchurch Street). The city extends into the distance, showing a mix of modern glass buildings and older brick structures.

AIMS

# What do we want to find out?

A photograph showing a grid of yellow rubber ducks in wooden boxes. There are two rows of four boxes each. In the second row, the third box from the left contains a single blue rubber duck, while all other boxes contain yellow ones.

Are brands doing enough to  
ensure their Facebook content is  
sufficiently differentiated?

# How do we use Data Science to answer this?

Can we make a robot smart enough  
to be able to tell the difference  
between brands on Facebook?\*



\*Can we use supervised machine learning to allow us to classify different brand content on Facebook?

# DATA



# One category, seven brands



Waitrose



Sainsbury's



ASDA

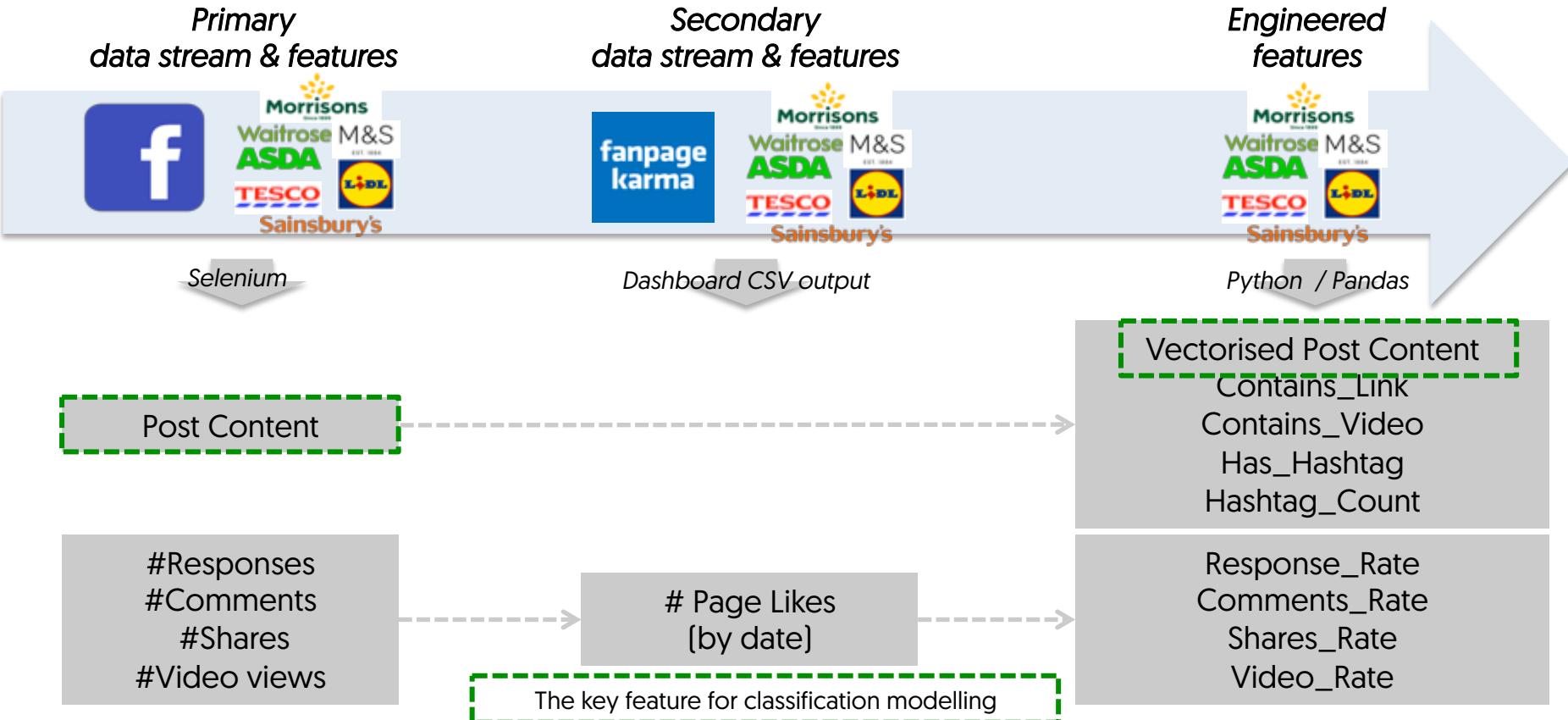


Higher End / Premium

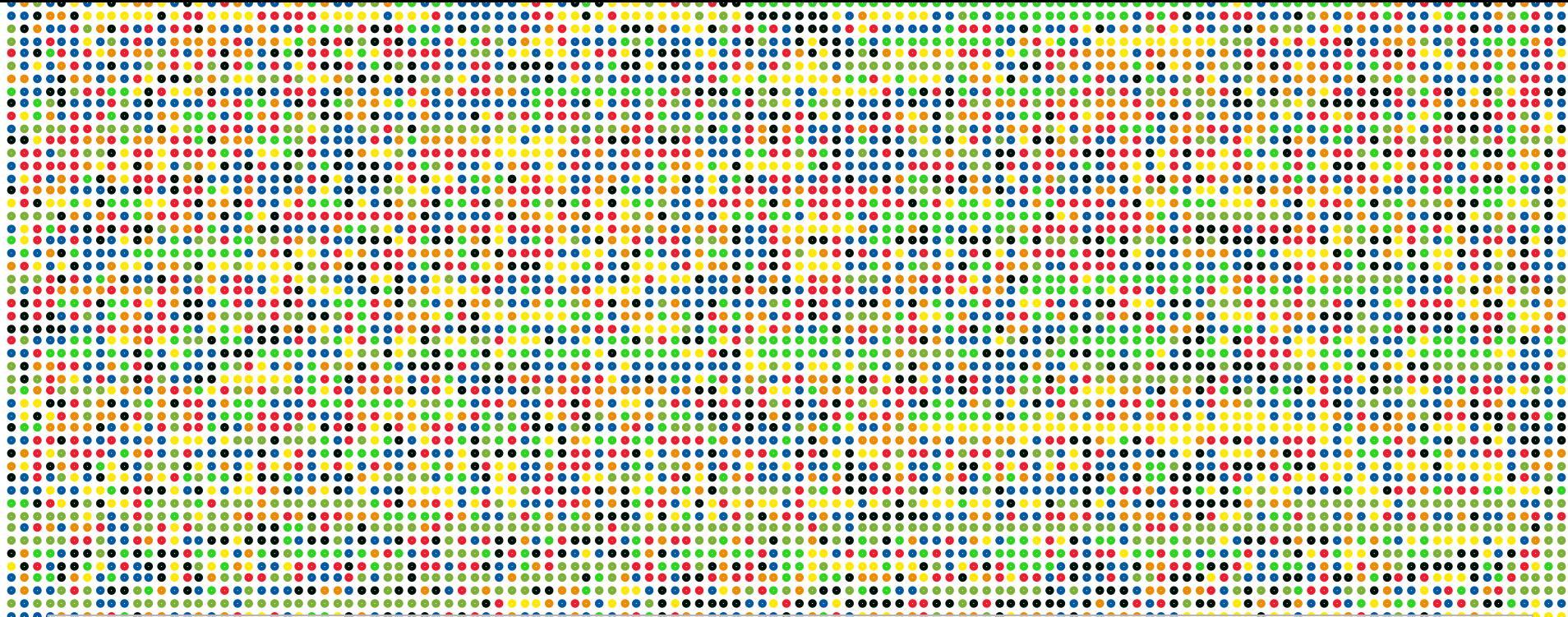
Lower End / Budget

All on Facebook. All posting consistently. All with some purpose.

Two main data streams followed by feature engineering



# 6350 social media posts were scraped



Waitrose Sainsbury's **TESCO**

**ASDA**

**Morrisons**  
Since 1899



# Classes were mostly balanced



Waitrose Sainsbury's **TESCO**

**ASDA**

Morrisons  
Since 1899



870 posts  
(0.13)

777 posts  
(0.12)

714 posts  
(0.11)

1068 posts  
(0.16)

770 posts  
(0.12)

798 posts  
(0.12)

1353 posts  
(0.21)

Our baseline is 0.21 – our dominant class. If we can build a model that can score higher than this, then we can reject the null hypothesis and concede that there are genuine differences in the content produced by our brands

Unsurprisingly, our branded content was full of...branding



This makes things far too easy for a model to learn from



We need to do some cleaning up



# How?



Extensive &  
bespoke  
'stop word' lists

Lots and  
lots of  
regular  
expressions

Remove all 'hard' branding cues : specific mentions of a brand



# Remove all 'soft' branding cues : celebrity endorsements



# Remove all 'soft' branding cues : hashtags



All we want left is the narrative



# All we want left is the narrative

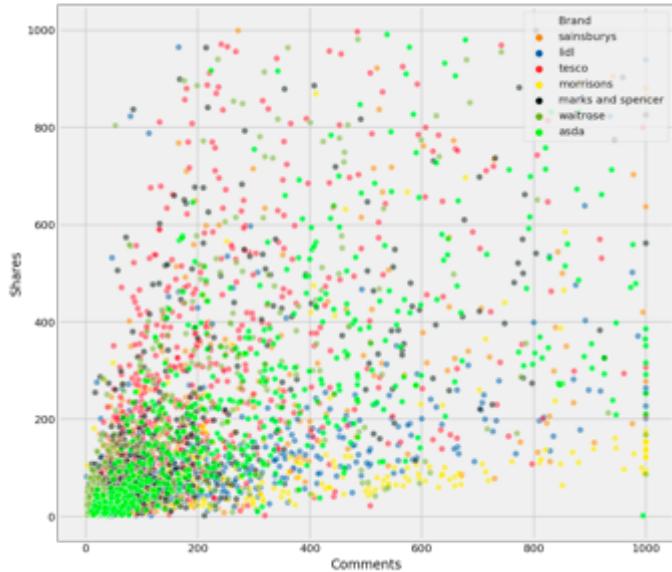




# EDA & MODELLING

# Engagement metrics aren't correlated strongly

$r=0.36$

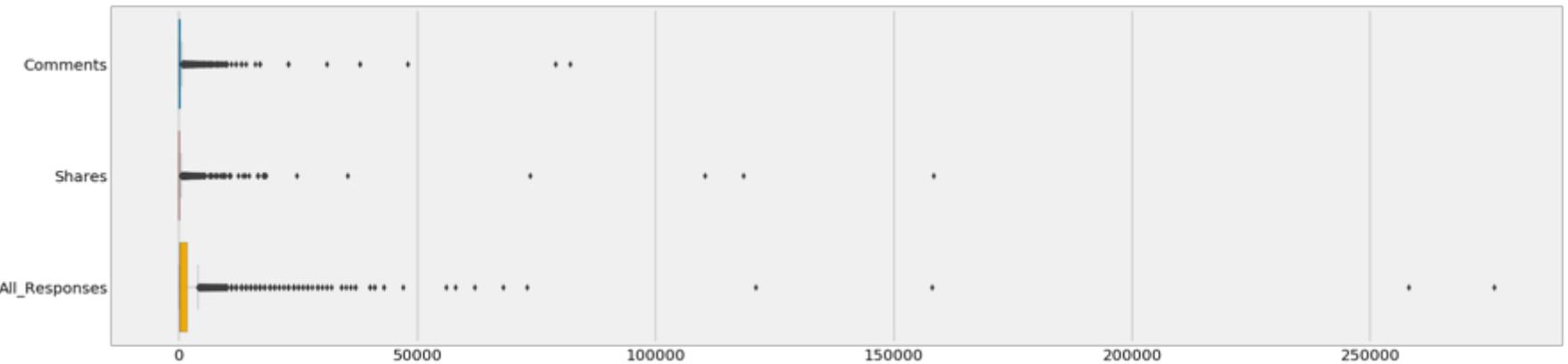


$r=0.38$



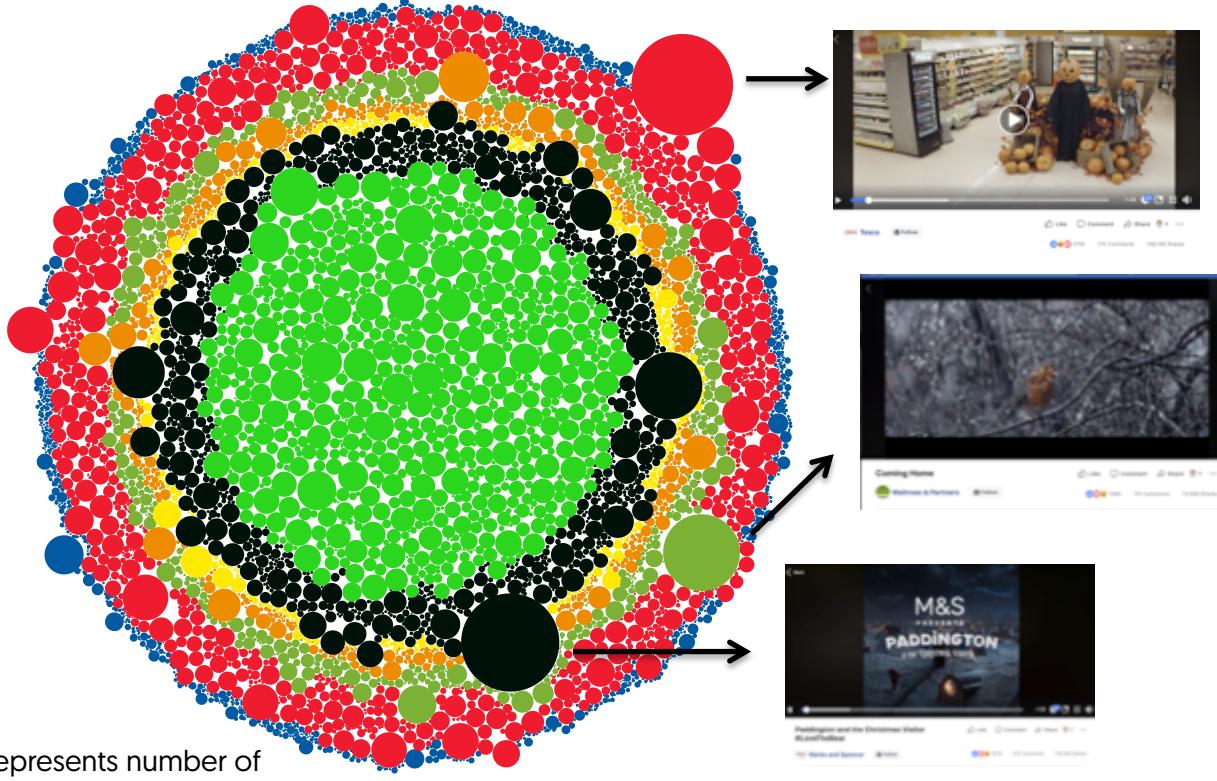
i.e. if a post gets a lot of comments, it doesn't necessarily get a lot of shares

# Engagement metrics are prone to outliers\*

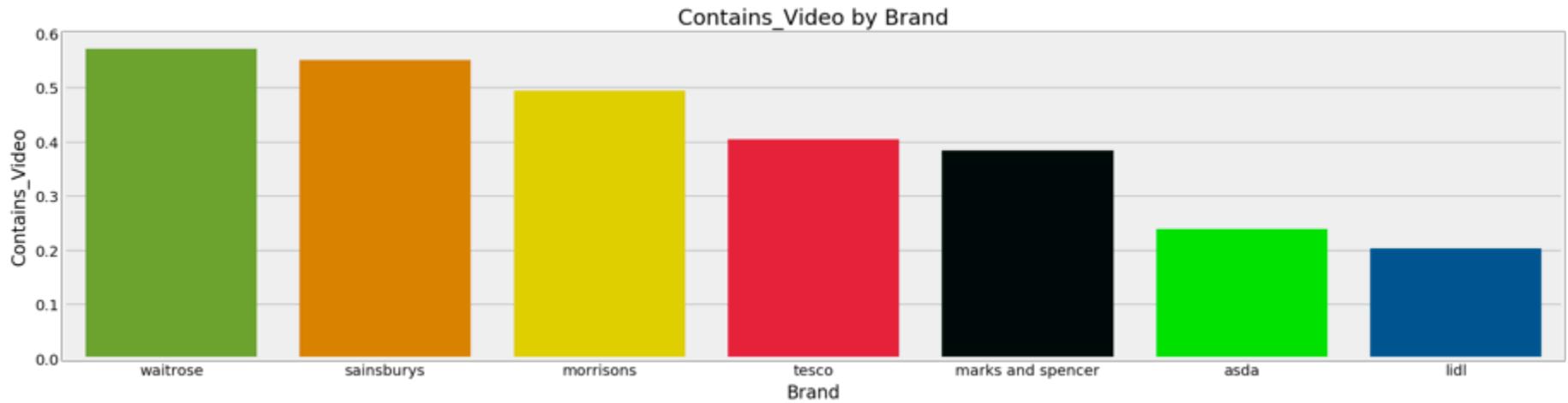


\*posts that were very popular, receiving an unusually high number of shares, comments and responses

# Those outliers are usually videos that are entertaining or that carry an emotional resonance of some kind

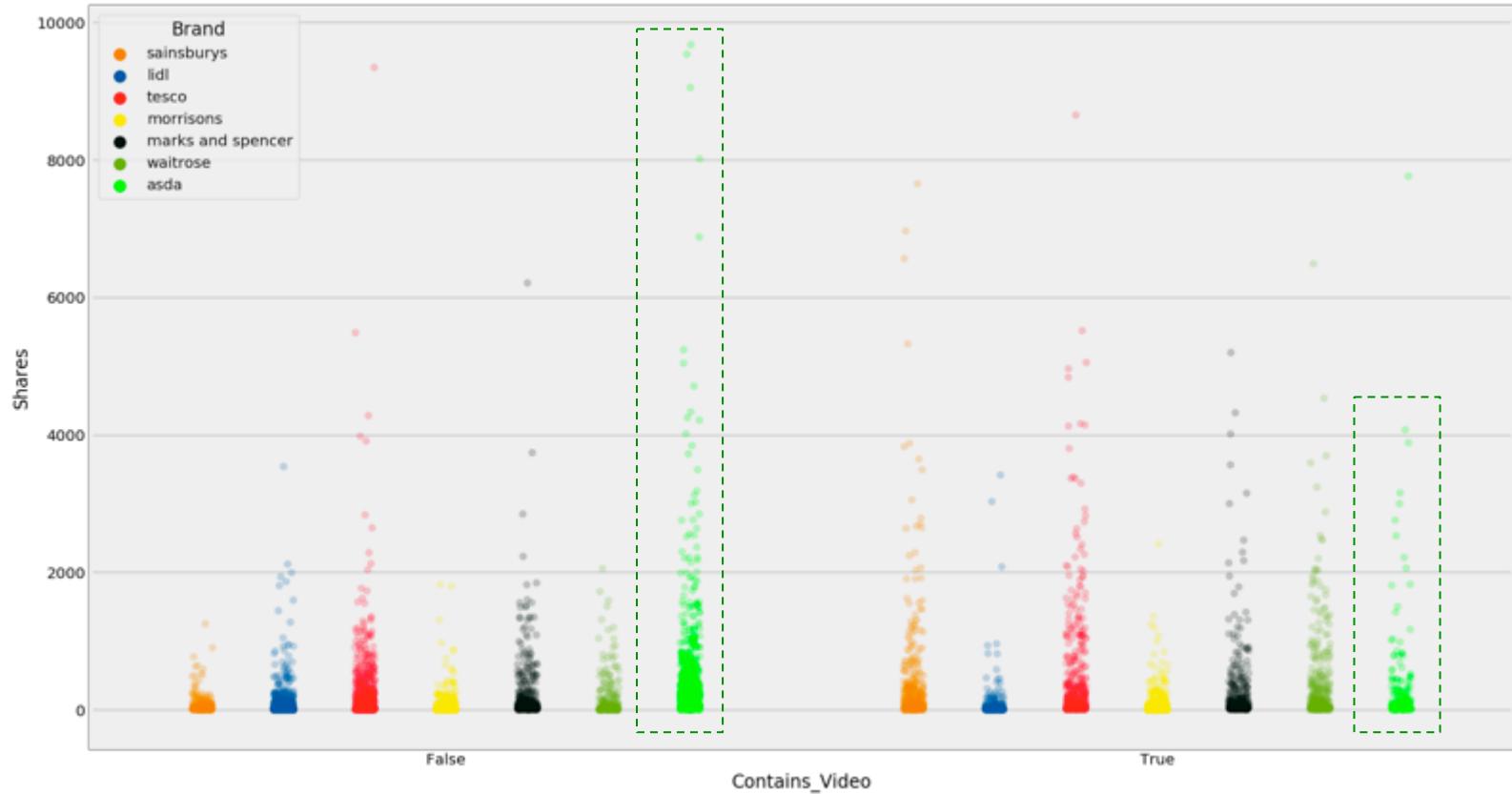


# Some brands are far more likely to post videos than others



Does this mean they also get a lot of shares?

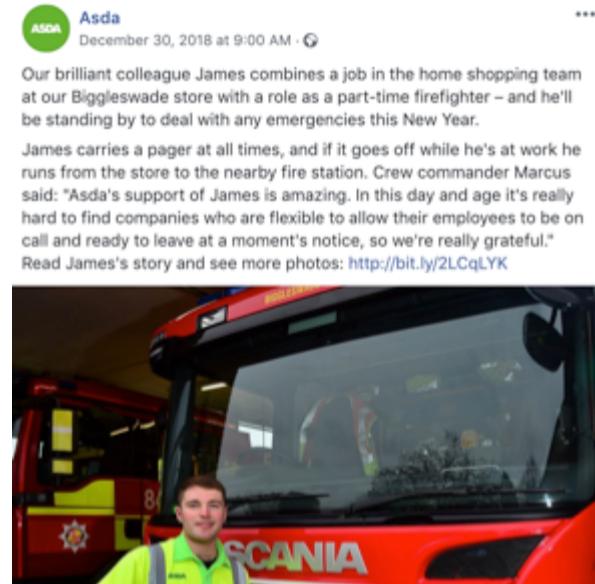
Not in all cases and particularly with ASDA – whose posts without videos consistently get shared more



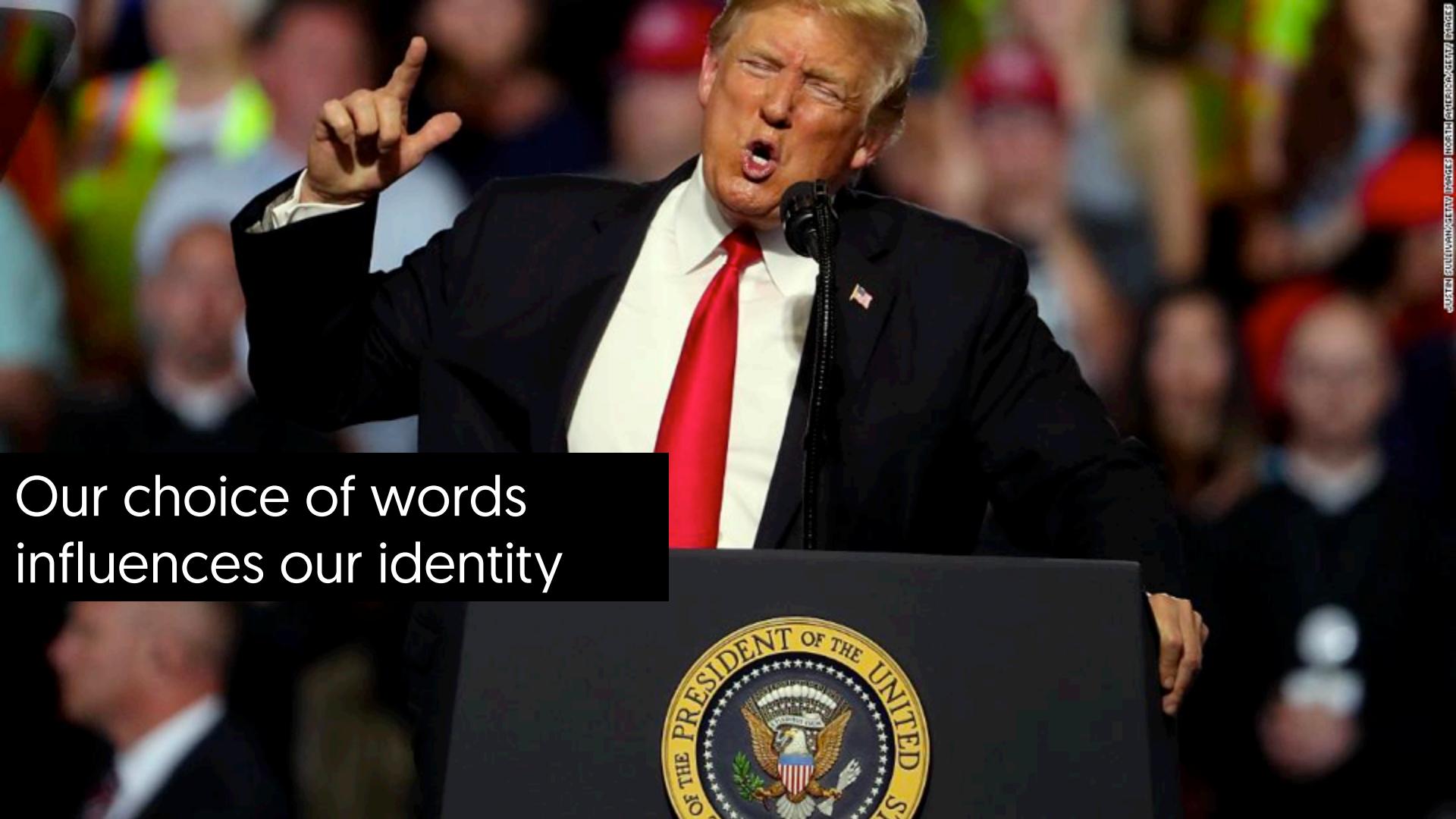
# Qualitatively, it feels there is a big difference in what supermarket brands talk about on Facebook



Short. To the point.  
Food. Recipes.



Wordier.  
People. Causes



Our choice of words  
influences our identity



Sainsbury's talk about their magazine/recipes.....less so for ASDA who focus on people and communities



Weighted term frequencies by brand : [TF-IDF] Vectorisation

Lidl talk about price and stock availability  
whereas Waitrose focus on recipes



price	just	today	deluxe	tomorrow
stocks	thursday	week	time	day
	sunday	weekend	got	pick
correct	christmas	make	range	

# Waitrose

recipe	watch	perfect	chocolate	christmas	comments
	try				
delicious	recipes	favourite	simple	make	free
		tell			
read	win	dinner	simply		
		enjoy	ultimate		

Weighted term frequencies by brand : [TF-IDF] Vectorisation

# M&S talk about new things to shop for, Morrisons (a bit like Asda), avoid talking about food

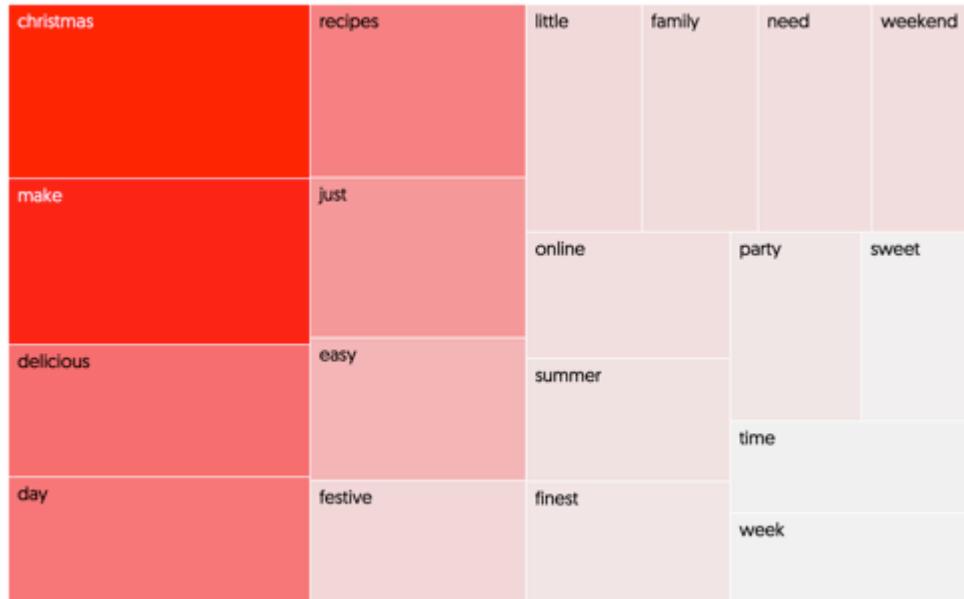
M&S

EST. 1884



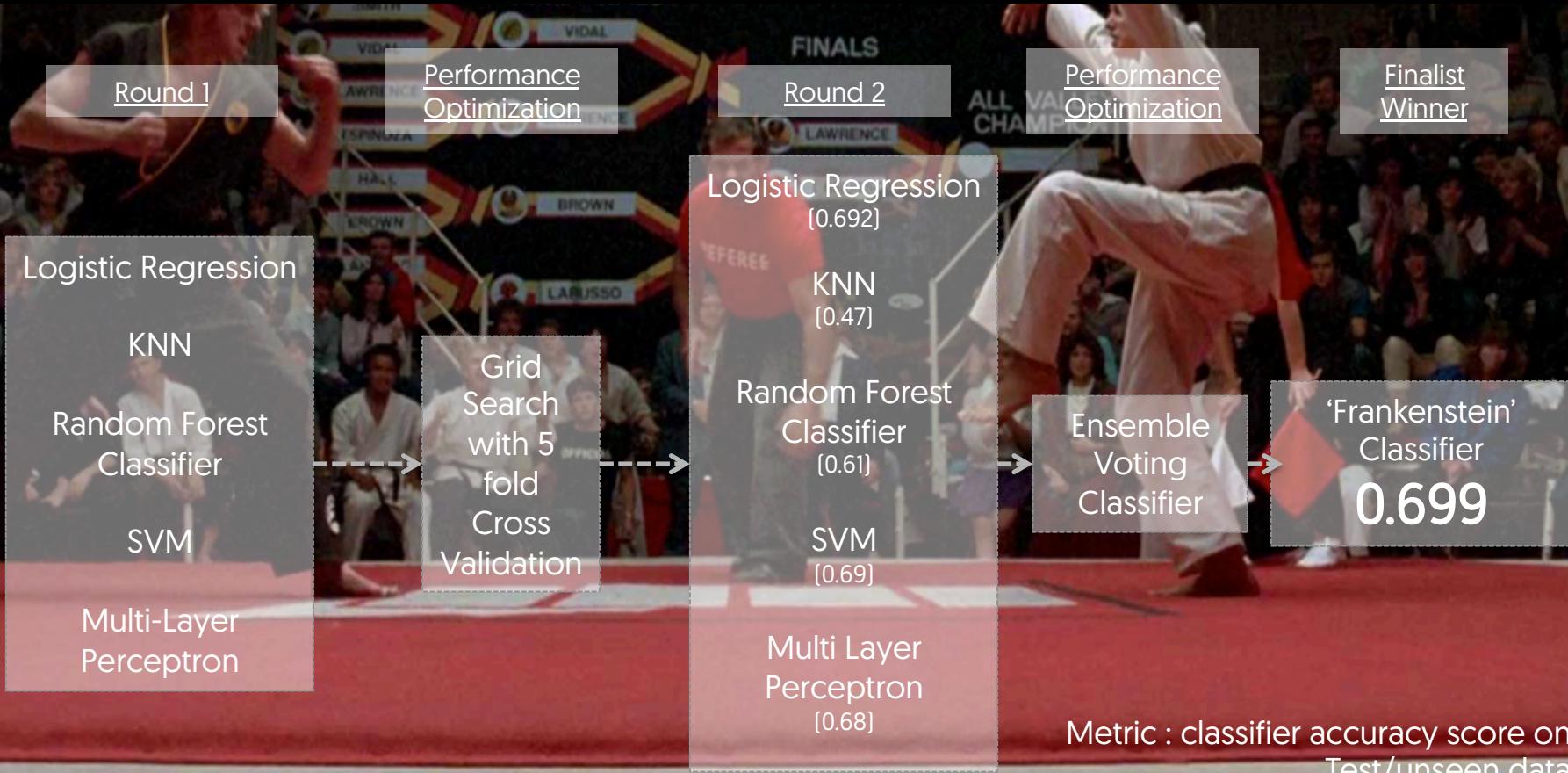
Weighted term frequencies by brand : [TF-IDF] Vectorisation

# Tesco owns Christmas



Weighted term frequencies by brand : [TF-IDF] Vectorisation

# Modelling Approach



# Classifier Evaluation : Confusion matrix

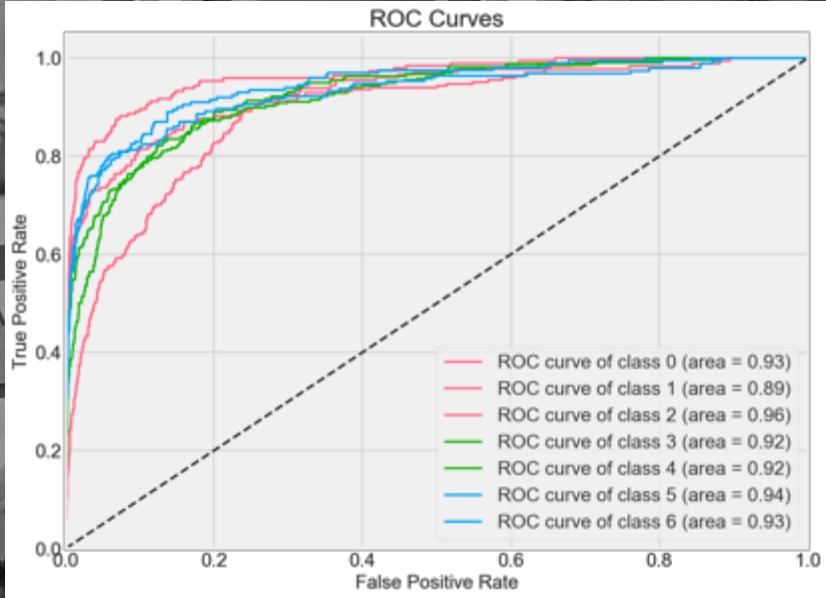
High precision with Waitrose, Sainsbury's and ASDA i.e. when our model predicted these brands over 80% of the time it was correct

	p_Sainsbury's	p_Tesco	p_Waitrose	p_Lidl	p_M&S	p_Morrisons	p_ASDA
Sainsbury's	111	19	5	23	11	8	1
Tesco	5	189	7	30	17	13	6
Waitrose	0	22	147	11	7	6	1
Lidl	9	42	4	246	23	6	8
M&S	3	33	3	23	135	19	2
Morrisons	1	20	2	10	8	153	6
ASDA	3	27	2	8	9	14	130

	precision	recall	f1-score	support
Sainsbury's	0.84	0.62	0.72	178.0
Tesco	0.54	0.71	0.61	267.0
Waitrose	0.86	0.76	0.81	194.0
Lidl	0.70	0.73	0.71	338.0
M&S	0.64	0.62	0.63	218.0
Morrisons	0.70	0.76	0.73	200.0
ASDA	0.84	0.67	0.75	193.0

Poorer performance with M&S and Tesco

# Classifier Evaluation : ROC-AUC



Area under ROC curve [ROC-AUC]:

Sainsbury's: 0.93

Tesco: 0.89

Waitrose: 0.96

Lidl: 0.92

M&S: 0.92

Morrisons: 0.94

ASDA: 0.93

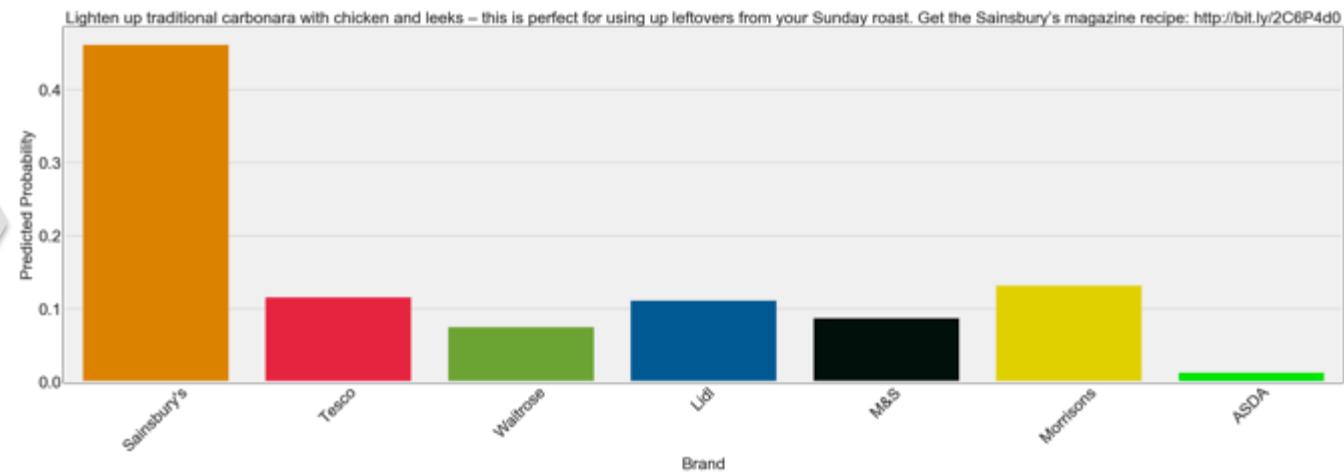
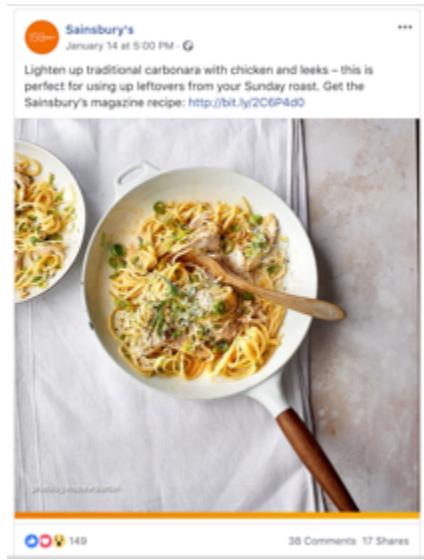
## ROC-AUC Curve Interpretation

All classes (brands) have high ROC-AUC scores which implies that for most of our brands, our model has been able to provide strong separability between **true positives** for that brand (i.e. predicting 'Waitrose' and it being brand 'Waitrose') and true negatives (i.e. correctly predicting it as something else other than Waitrose).

EMILY

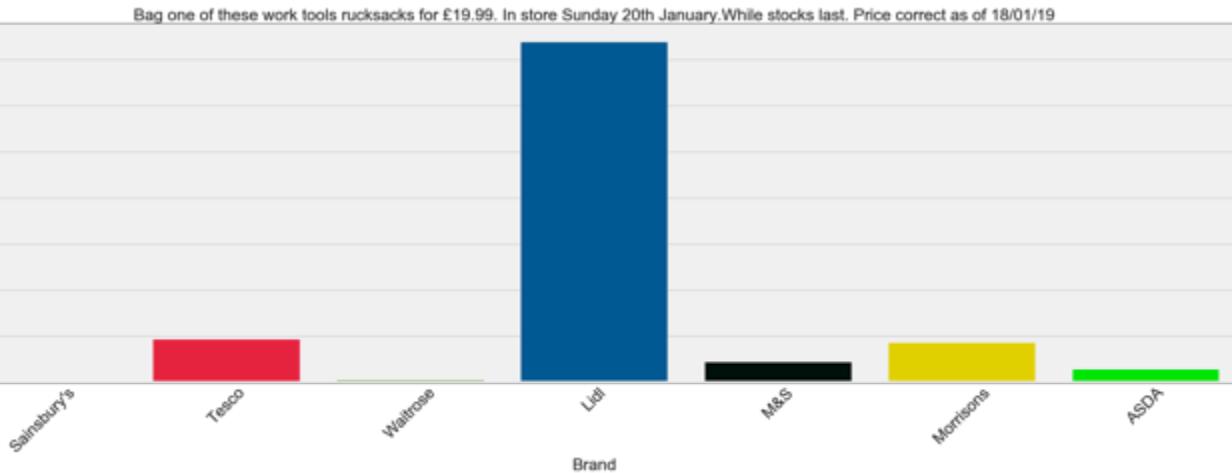
The only brand that has slightly weaker AUC scores is Tesco - this implies that the proportion of false positives and false negatives for Tesco is higher i.e. that our model sometimes incorrectly classed a post as Tesco when it was Morrisons (False Positive) and should have classed a post as Tesco, when it classed it as something else (e.g. Morrisons)

# Testing the model on new data

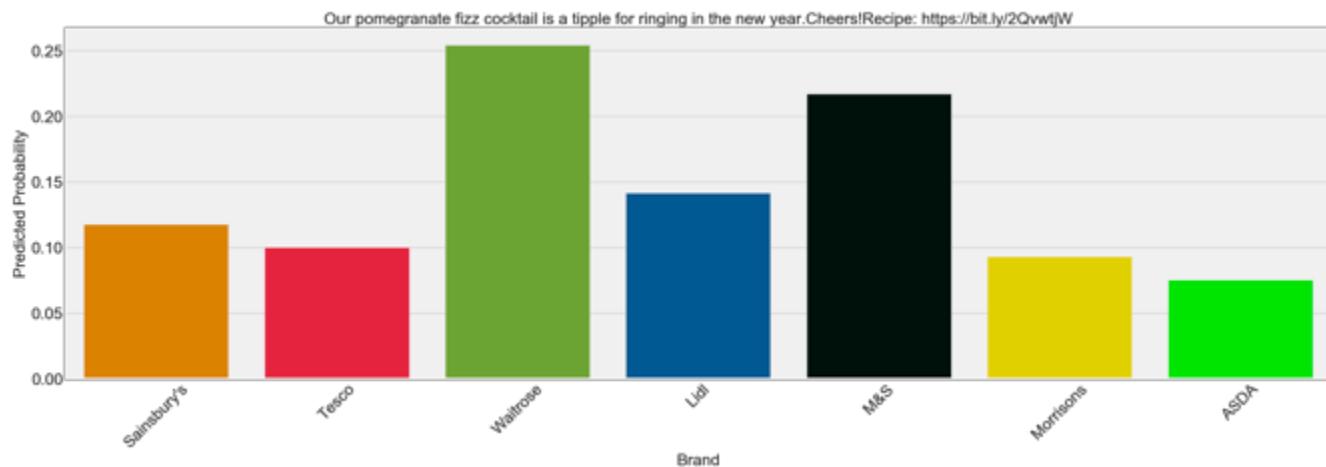


Note – there is a pre-processor in the pipeline that removes all branding cues before the model makes any predictions !

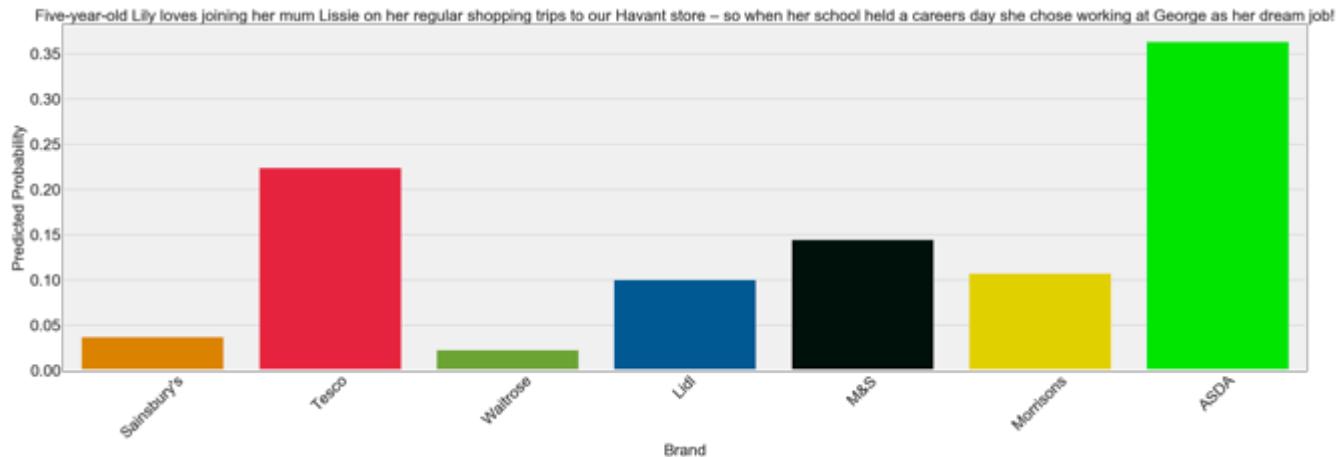
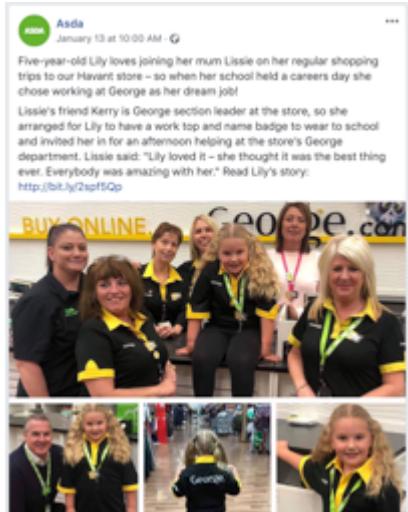
# Testing the model on new data



# Testing the model on new data



# Testing the model on new data



# IMPLICATIONS



# Sense checking your social content



**StubHub**  
@StubHub



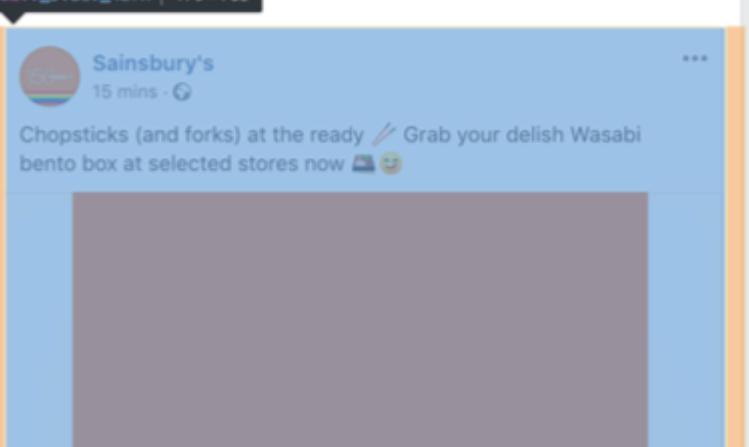
Thank [REDACTED] it's Friday! Can't wait  
to get out of this stubsucking hell  
hole.

I have no doubt that most social media  
managers know what they're doing, but a little  
sense check never hurt anyone



```
><div id="u_fetchstream_6_n">
><div class="1xnd">
```

# Transferability to other categories



Facebook pages are (mostly) built with consistent html.

Furthermore, the social infrastructure across brands (likes, comments, shares) is also consistent for all brands.

So in terms of acquiring new data for other categories and brands, it would be fairly straightforward to replicate this project again for anything else you can think of

```
::before
><div class="4-u3 _5dwa _5dwb _3v6c">..</div>
><div class="5val _427x">..</div>
><div class="5val _427x">..</div>
```

#facebook body div#u\_0\_j\_ li div#globalContainer.uiContextualLayerParent

Styles Event Listeners DOM Breakpoints Properties Accessibility

Filter :hover .cls +

element.style {  
}

#facebook .\_\_kb div { ls-WIYBBWFA.css:33  
font-family: inherit;  
}

.\_\_zyq #globalContainer { nvJ6KSC iVp.css:38  
width: 1012px !important;  
}

#globalContainer { ls-WIYBBWFA.css:35  
}

position 0  
margin -  
border -  
padding -  
1012 x 3401 -127 0  
0 -  
0 -

Filter Show all  
color  
direction  
display  
font-family  
font-size  
.css:12  
ltr  
block  
system...  
12px

DOM node highlight that node

Console and select

Store as global variable

# Risks & Limitations

## It's (not) been emotional

Word counts - even TF-IDF –  
don't capture  
sentiment very well.

Future iterations of content  
analysis could look at sentiment  
and see if emotion is a useful  
predictor?

## Neglected Features

Many features were acquired,  
that although were helpful  
for EDA, weren't used in any  
modelling.

Could we look at regression models  
and see what kinds of content  
predict social 'success'

## Neglected Channels

There's more to social  
media  
than Facebook.

Could we integrate data  
from Twitter, Instagram,  
Snapchat?

# QUESTIONS?

Mr. Mark Zuckerberg