Questions:

1. Why is engagement with social media content important for brands?

People use languages to convey their beliefs and thoughts towards all aspects in their life, and those words are informative enough and contain tremendous psychological value. And as the social media user base increases at an incredible speed, it has already become an essential platform for marketing teams of various brands to collect and explore market information. By analyzing the texts on social media and people's corresponding reactions, the companies can convert those psychological values into strategies, and finally generate real gains in the long run.

However, the accuracy of merely using natural languages to predict psychological states are limited, but the information in the engagement with social media content is relatively more straightforward. For instance, people are likely to use some buzzwords or ironic expressions in their daily postings which makes text analytics methods hard to explore useful information. On the contrary, the ways of engagement on social media are more flexible, including likes, shares and comments. By analyzing the textual contents in comments of particular posts, brands can narrow down topics and can better identify the target group; and by observing the number of likes, brands can more intuitively understand the market feedback and expectations. And shares are even more indicative in identifying customers' thoughts towards certain contents in the posts.

Most common marketing problem for brands can be targeting and attracting the wrong customer. By analyzing the engagement with social media contents, brands can increase the accuracy of conveying the market information and can improve the marketing strategy. And for those people who show interest on social media but are not ready to purchase, by analyzing the hints on social media engagement, companies can also adjust their strategy and solve the pain points. By analyzing people's interaction towards relevant posts on websites, brands can also identify whether the current marketing efforts are working or not and gather market expectations.

2. How accurately can engagement (likes + shares + comments) with digital content be predicted? What kinds of variables are most useful for making these predictions?

Methodology:

To better measure the prediction of engagement and digital contents, we firstly processed the dataset about relative comments parsed from Facebook posts, and did necessary data cleaning for further analysis. We aimed to set the total number of interactions of each post to be our target variable, where interactions are the sum of the numbers of likes, shares, comments, etc. And we converted several important textual variables into categories, such as the page type and posting daytime, for further analysis.

To better analyze the digital contents in text, we used the processed textual contents from comments and aforementioned categorical variables to construct the model. And in order to

gain a brief idea of what variables are most useful for making relative predictions, we constructed a linear regression model to predict the total number of interactions. We first put all the predictors into the model, in hope of utilizing the statistical summary to select the features that affect the target variable most. However, the R-square of the model is really low, and most of the features are marked as significant. To improve the model performance, we utilized other machine learning algorithms such as Principal Component Analysis to reduce the dimensionality of datasets. We kept the most important 45 PC components to further construct the prediction model. But according to the result, the mean squared error is too high to be a reasonable model. Thus, we moved on to another approach.

In our second try, other than using the total interactions as the target variable, we utilized engagement which is the sum of the number of shares, likes and comments. And instead of using only textual contents and categorical variables as predictors, we included other digital contents such as page followers and total likes of each page. In addition, according to our observation, the distribution of engagement is highly skewed which makes prediction hard to achieve. To solve this, we processed log transformation to the target variable before constructing the model. In this case, the linear regression model gives an MSE of about 1.59, making it a feasible prediction model to rely on. Due to the fact that the random forest algorithm tends to provide higher accuracy compared with other methods, we measured feature importance with random forest to find out the most useful variables for making predictions.

To observe the feature importance, we set the threshold of feature importance score at 0.01 and we get 18 features. Variables that are most useful for the prediction of engagement are shown in the table below. According to the results shown in the table, the top 5 most influential features to the engagement are followers at posting, type live video complete, likes at posting, time difference, and prep. The result indicates that followers at the posting page appear to have the largest effect on the prediction of engagement amount. The number of followers at posting is useful for the prediction since there might be some relationship between followers at posting and engagement. In general, the more followers a page has, the more engagement it tends to have, but further investigation is required for validation. The format of the post is also important to the prediction. In particular, whether a post includes a complete live video or not might indicate the engagement amount. Similarly, likes at posting and the time difference between a post and the latest post in the dataset play important roles in prediction too. Among the 18 useful features, 13 of them are content-related. As a result, the actual content the posting contains also matters to the prediction of engagement.

predictor	eature importanc
Followers.at.Posting	0.214493
type_Live Video Complete	0.108307
Likes.at.Posting	0.0674495
time_difference	0.0394111
prep	0.0212749
WC	0.0171971
focuspast	0.0148407
number	0.0148016
OtherP	0.0113596
Sixltr	0.0113369
article	0.0111962
fnctn	0.0110718
page_category_NEWS_SITE	0.0109006
WPS	0.0106651
Period	0.0105058
Comma	0.0102949
motion	0.0101939
we	0.0101452
time	0.0100547

3. Drawing from your analyses, and other qualitative considerations, what do these data say about the factors that cause people to engage with social posts? Is engagement mostly a function of factors extrinsic to the post, like who posted it or when they did so, or mostly due to factors intrinsic to the post, like the language it uses?

predictor	feature importance
Likes.at.Posting	0.1252984
Followers.at.Posting	0.051052866
type_Live Video Complete	0.043028365
netspeak	0.029075741
cause	0.024179004
prep	0.024086767
article	0.023567294
WC	0.020077843
space	0.018493972
Sixltr	0.016369391

Feature Importance for: Regression model

We used the random forest to conduct regression models based on our clean dataset with the target variable being the sum of number of shares, likes and comments, which we call total interaction. We also considered the time slot variable based on our understanding that we calculate the time difference between when it posts and current time, and assign the time slot. Adding the new variable makes an improvement into our model, and then we did the feature selection process to see the importance of each predictor. As the result presents the top 3 important factors are followers_at_posting, likes_at_posting (total number of likes at the page) and whether the post is in live video.

As we can learn from the feature importance, although the top 3 can be grouped into a function of factors extrinsic, which relates to the popularity of people who post it, the rest of the feature importance contains a lot of factors which are intrinsic. Predictors like WC means word count, WPS means words or sentence, Sixltr means words greater than 6 letters and articles, those important factors can be contributed into language and linguistic aspects. Therefore, it's hard to conclude which factors cause the most for people to interact with social posts, but we can make an assumption for the cause based on chronological order. Since the extrinsic group has the highest score in feature importance, we assume when people see a post, they will firstly pay attention to who posts it and the time when it post, if the person they are interested or familiar with, then they will take a good look at the details of the post. Now it is the time for intrinsic factors to determine whether the people will leave any interaction like comment or share if they are really impressed by the content of the post.

```
lm(formula = data$target ~ ., data = attri)

Residuals:
    Min      1Q  Median      3Q     Max
-145.40  -40.62  -25.38   -4.20  901.20

Coefficients: (2 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                8.440e+01  5.083e+01   1.660 0.096895 .
Followers.at.Posting      -3.821e-05  1.379e-05  -2.771 0.005616 **
type_Live.Video.Complete   5.540e+01  4.331e+00  12.793  < 2e-16 ***
Likes.at.Posting           4.671e-05  1.447e-05   3.228 0.001256 **
time_difference            1.625e-07  5.673e-08   2.865 0.004188 **
prep                      -1.249e+00  3.187e-01  -3.918 9.04e-05 ***
WC                         1.413e-01  4.956e-02   2.850 0.004390 **
focuspast                  1.660e+00  7.267e-01   2.285 0.022365 *
number                    -1.121e+00  2.840e-01  -3.947 8.01e-05 ***
OtherP                     3.539e-01  2.546e-01   1.390 0.164656
Sixltr                    -8.461e-02  1.858e-01  -0.455 0.648913
article                    8.147e-01  4.164e-01   1.956 0.050477 .
fnctn                      2.502e-01  2.186e-01   1.145 0.252436
page_category_NEWS_SITE   -1.174e+01  5.723e+00  -2.051 0.040271 *
WPS                        5.217e-02  1.941e-01   0.269 0.788070
Period                    -8.949e-01  3.967e-01  -2.256 0.024107 *
Comma                     -1.892e+00  4.469e-01  -4.234 2.34e-05 ***
motion                     1.081e-01  5.866e-01   0.184 0.853831
we                        -2.985e+00  8.270e-01  -3.610 0.000309 ***
time                      -8.367e-01  2.503e-01  -3.343 0.000834 ***
sec_15                    -8.501e+01  1.109e+02  -0.766 0.443465
sec_30                     1.141e+02  1.219e+02   0.936 0.349100
sec_60                            NA         NA      NA       NA
sec_120                           NA         NA      NA       NA
sec_half_day              -1.096e+02  7.972e+01  -1.375 0.169285
sec_day                    5.797e+01  4.192e+01   1.383 0.166751
sec_7_days                -1.717e+01  1.932e+01  -0.888 0.374399
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.92 on 4975 degrees of freedom
Multiple R-squared:  0.06169,   Adjusted R-squared:  0.05717
F-statistic: 13.63 on 24 and 4975 DF,  p-value: < 2.2e-16
```

Regression results

We also have analyzed the coefficients of each predictor to understand how they affect the overall interaction. Here we only analyze predictors with high significance. For extrinsic factors, posts with live videos have positive coefficients, indicating more interactions with users. This explains the increasing popularity of live streaming as users feel more involved. As for the intrinsic factors, posts written in first person plural (we), containing time, numbers or too many commas will likely lower the interactions. This can be explained as well since posts stating “we” seem less engaging and distant compared to posts using third person, whereas too many time and number phrases or commas will lead to difficulty in reading, discouraging users to finish reading the post and interact afterwards.

4. What are the implications of these results for social media marketing managers? In addition to addressing this question in general, select one of the brands represented in the dataset and recommend a strategy they could use to boost engagement with their posts.

Based on our model, we believe social media engagement is more about accumulating influence in the long run since followers and likes for the page are playing a more important role than other factors. It is also worth considering for the social media marketing managers to leverage the social platform recommendation system, such as posting content that fits the trending topic, purchasing the advertisement position, or giving readers incentive to engage with the post

Strategy recommendation for ‘ABC10’:

First, based on the general model we’ve developed using linear regression, we see that post of live video will positively affect the interaction. Meanwhile, using too many words describing number, time and 1st pers plural may result in a lower engagement rate. Therefore, the post should avoid using too many of these phrases.

Second, ABC10 can consider developing a similar prediction model like we’ve done for the general posting dataset, but with only the data in the TV-channel category so it will obtain the model that fits the category better considering people who view different categories social accounts may have a different preference and adjusting its strategy accordingly.

Besides the regression model, we think it is also reasonable to approach the prediction through binary prediction. Due to the different sizes of the company, different industries, and different lengths of running the social media account, it is unfair to compare the engagement with all the accounts on a platform. With binary prediction, the social media marketing manager of ABC10 can conduct research beforehand with in the TV- Channel to set a threshold that is used to measure whether the engagement reaches expectation or not for the campaign, then feed the model with all kinds of predictors and target threshold to see whether the campaign is going to

work well. Right now, with 75% quantile of interaction column as a threshold to conduct the prediction, we are able to achieve around 80% accuracy.

Last, to boost engagement, ABC10 must bring more followers to increase the engagement rate in the long run. ABC10 can conduct some campaign that specifically targets the increase of followers. For example, adding some interaction point in the post like a lucky draw or similar activity