MGT 6203 Group Project Final Report

Predicting consumer engagement affected by social media message characteristics

Team #: 48

GitHub Project Repository: Link

Dataset: Link

Team Members: Ju Yeun Hong, Sunwook Kim, Sumin Shin, Taewoo Kim, Yujeong Lozalee

1. Introduction

1.1. Background

In marketing, there are some important frameworks that underscore the crucial role of brand awareness and customer engagement. One of those frameworks is the AIDA model [1], which highlights the critical stages of attention, interest, desire, and action. Without the first two steps of 'attention' and 'interest,' a consumer's purchase is not likely to happen. Similarly, the Hierarchy of Effects model [2] outlines stages a consumer undergoes, emphasizing that purchases usually follow stages like awareness, knowledge, liking, etc.

In line with these principles, social media marketing has become an essential component to enhance brand awareness and customer engagement in marketing strategies.[3, 4, 5] For instance, businesses utilize social media platforms to amplify brand awareness and customer engagement, effectively actualizing the 'awareness', 'knowledge', and 'liking' stages and building a more robust connection with consumers.

1.2. Business Justification

Our primary aim is to find a way to boost consumer engagement on social media, particularly Twitter (now rebranded as X), by analyzing user reactions to posts. For instance, if data reveals that tweets with text and images lead to a 20% increase in positive interactions, we could recommend that businesses prioritize incorporating images in future posts. We will quantify engagement using numeric predictors like positive emotion score, text volume, and visuals, tracking post likes for impact assessment. Aligned with AIDA and Hierarchy of Effects theories, enhancing brand awareness and customer engagement on Twitter can serve as the initial step to drive sales through strategic advertising. Optimizing engagement strategy can not only enhance brand visibility, but also lay the foundation for potential sales growth.

1.3. Problem Statement

In the dynamic landscape of digital marketing, understanding how to optimize brand visibility and customer engagement on social media platforms has become imperative. Twitter, with its massive user base and real-time communication capabilities, stands out as a critical platform for businesses aiming to bolster their brand presence. However, crafting messages that foster engagement is a challenge. The intricacies of message characteristics, such as emotional tone, text volume, and the inclusion of visuals, can play a significant role in how consumers interact with brand content. Considering this, our team is committed to exploring the unfolding dynamics of engagement and interaction on social media. Specifically focusing on Twitter, we aim to uncover the profound impact it has on brand visibility and customer interaction. Our investigation will scrutinize how Twitter messages shape consumer engagement, providing valuable insights to assist businesses in refining their digital marketing strategies.

Primary Research Question:
What is the relationship between social media message characteristics and consumer engagement?

Supporting Research Questions:
- H1: The more positive the emotion of a social media message, the more likes the message will generate.
- H2: The amount of textual information in a social media message will decrease likes of the message.
- H3: The presence of a visual attachment in a social media message will increase likes of the message.
- H4a: The presence of business-related words in a social media post will increase the number of likes to the post.

· H4b: The presence of environment-related words in a social media post will increase the number of likes to the post.

2. Data Overview

2-1. Data Sources and description

A group member collected Twitter metadata on February 3, 2023. With his permission, we used the data for this project. The group member gathered all tweets (n = 147,857) from the timelines of 79 sustainable fashion brands, using the Twitter API and the R package rtweet. The dataset comprises 44 variables, and the tweets range from the oldest on February 7, 2009, to the newest on February 3, 2023.

2-2. Data Cleaning Process

1) Focusing on original tweets: We are solely interested in original tweets, excluding retweets and replies for the purpose of this project. As a result, 55,605 organic tweets created by the brands remained.

2) Creating a new dataframe and defining key variables: After filtering out most unnecessary variables, we established a new dataframe, 'data,' and configured some key variables required for our future modeling.

3) Checking for missing values and creating transformed variables: After ensuring that our key variables do not contain missing values, we created log-transformed and square root-transformed variables for our convenience in future modeling.

4) Establishing filtered dataframes : To account for the potential impact of each Twitter account's follower count on the response variable, we introduced an additional variable, 'norm_fav,' calculated as the ratio of 'favorite_count' to 'followers_count' (the brand's number of followers). As 'followers_count' was obtained on February 3, 2023, considering 'norm_fav' in older tweets might not be reasonable. Therefore, we created a filtered dataframe, 'recent_data,' containing data only from 2022 to 2023. Additionally, anticipating a substantial number of 0 values in the response variable, we generated other filtered dataframes, 'data_nonzero' and 'recent_data_nonzero,' excluding instances with a response value of 0.

5) Discovering keywords related to business and sustainability: Assuming that the top 500 liked tweets in 'recent_data' contain specific business and sustainability keywords, we examined the top 500 liked tweets to identify those keywords. Binary variables were created to indicate whether the tweet includes those specific business and sustainability keywords.

2-3. Key Variables

Attribute Name	Role	Type	Description	N/A value
favorite_count	Response	Continuous	An engagement metric indicating how much the followers and public are interested in the tweet.	False
norm_fav	Response	Continuous	‘favorite_count’ divided by ‘followers_count’(the number of followers of the brand). It is to control the effects of the exposures of a tweet on ‘favorite_count’.	False
scaled_norm_fav	Response	Continuous	‘norm_fav’ multiplied by 1,000 (for ZIP and ZINB models).	False
fav_thres	Response	Categorical	Certain percentile of 'favorite_count' as a threshold for logistic regression.	False
norm_fav_thres	Response	Categorical	Certain percentile of ‘norm_fav’ as a threshold for logistic regression.	False
word_count	Predictor	Continuous	The number of characters in the text.	False

syuzhet	Predictor	Continuous	A sentiment score based on emotional tone or attitude expressed in each tweet. Lexicon-based sentiment analysis R package was used for this.	False
visual_presence	Predictor	Categorical	If a tweet contains any image, assign 1; otherwise, assign 0.	False
biz_presence	Predictor	Categorical	If a tweet contains top 10 biz keywords, assign 1; otherwise, assign 0. This only exists in recent_data'	False
eco_presence	Predictor	Categorical	If a tweet contains top 10 eco keywords, assign 1; otherwise, assign 0. This only exists in recent_data'	False

2-4. Exploratory Data Analysis

Exploratory Data Analysis(EDA) was conducted in four different parts: Distribution, Linearity & Boxplot, Outlier, and Multicollinearity. Three different dataframes ('data,' 'data_nonzero,' 'recent_data') were used for each part.

Through a comprehensive EDA, our goal was to identify and understand relationships among variables. The insights gained from this analysis enabled us to understand potential challenges, the need for transformed data, the appropriate model to use, etc.

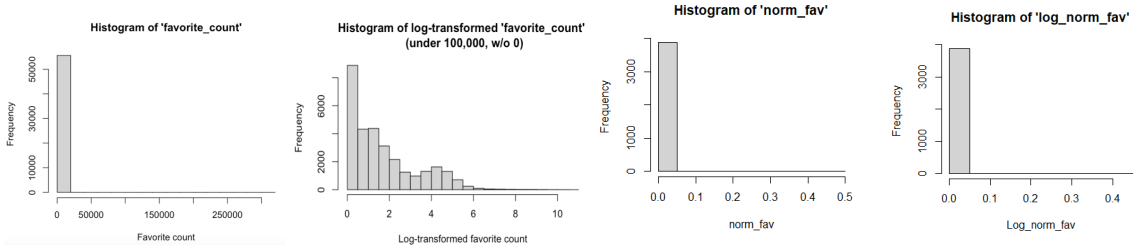
Following the EDA, we decided to concentrate on the ‘recent_data’ data frame for our modeling analysis, taking into account the potential impact of follower count.

2-4-1. Distribution

The distributions of predictor variables appear satisfactory, but the distribution of the response variable 'favorite_count' is extremely right-skewed. 'favorite_count' spans a wide range from 0 to 313,114. Remarkably, 75% of the data points have values under 4, and 45% of all observations are exactly 0.

After log-transformation, removing potential outliers and 0 values from the response variable, the distribution remains right-skewed. Normalizing the response variable ('favorite_count' divided by 'followers_count') was expected to alter the distribution, but the check on the normalized response variable, 'norm_fav,' showed minimal change in distribution.

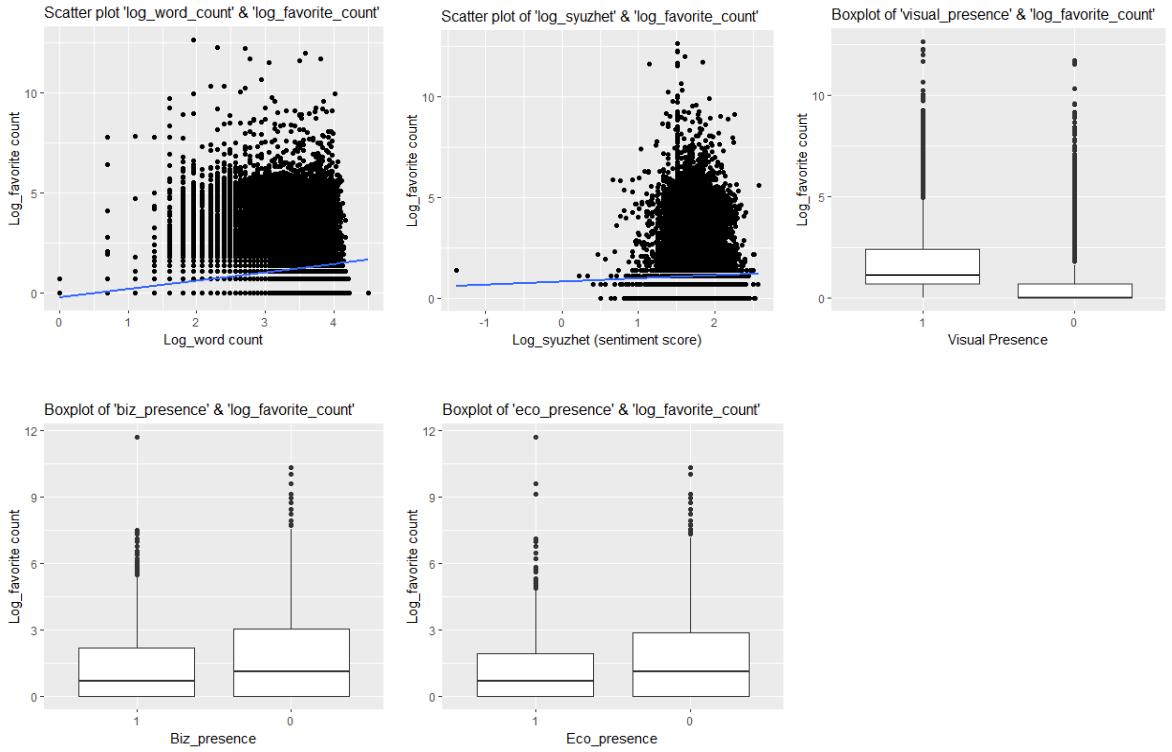
Therefore, although we might have to look at ‘Residuals vs Fitted’ plot after modeling, we can anticipate that this skewed distribution of the response variable might impact Linear Regression. Consequently, we may need to consider other models such as Logistic Regression, Zero-Inflated Poisson (ZIP), and Zero-Inflated Negative Binomial (ZINB).



2-4-2. Linearity & Boxplot

When checking for linearity in three different dataframes, the results were similar. Continuous predictor variables do not show much of a linear relationship with the response variable, but they might have a nonlinear relationship that can be captured by other models. Regarding categorical predictor variables, 'visual_presence' seems to indicate that including an image might affect the response variable. However, examining 'biz_presence' and 'eco_presence' in 'recent_data,' the results seem to indicate that the presence of business and economic

keywords did not significantly influence the response variable; in fact, they showed an almost slightly opposite effect.

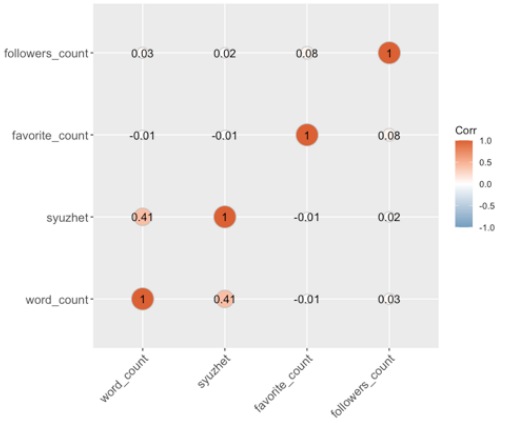


2-4-3. Outlier

We examined potential outliers in the response variable, hoping that excluding them would result in a better distribution. There are 8 data points with extreme values exceeding 100,000. However, excluding them did not significantly alter the distribution, likely due to the presence of too many 0 values.

2-4-4. Correlation

We checked correlation scores and assessed multicollinearity for continuous variables using ggcorrplot. The ‘syuzhet’ and ‘word_count’ indicate a moderate positive relationship, but variables show no signs of multicollinearity overall. We will confirm this later with VIF after modeling, including all categorical variables.



< Correlation matrix for continuous predictor variables in ‘data’ >

3. Methodology

We chose four different modeling methods for this project. The first two are the linear regression model and the logistic regression model:

- **Linear Regression:** Linear regression is used to predict the quantitative value of a response variable based on the predictor variable(s). We employed this method for its simplicity and effectiveness in predicting numeric variables, such as the like count.
- **Logistic Regression**: Logistic regression is used for binary classification problems. It estimates the probability of a discrete outcome given an input variable. In case a linear relationship between the predictor and response variables was unclear, we considered transforming the response variable into a binary format, subsequently applying logistic regression.

Based on the variables' distributions observed in the EDA phase, we explored various transformations on variables and tried different variable selections when building both linear and logistic regression models. However, considering the prevalence of 0 values in the response variable, we were concerned with potential limitations of simple model techniques like linear and logistic regression. We further employed Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) regressions because they are statistical models tailored for count data exhibiting an excess number of zero outcomes[6].

- **Zero-Inflated Poisson Regression**: In ZIP regression, the model a blend of a Poisson count model and a logistic regression model. The logistic part predicts the likelihood of an observation being a 'true zero' or an 'excess zero,' while the Poisson part models the count data assuming a Poisson distribution.
- **Zero-Inflated Negative Binomial Regression**: ZINB regression is used when count data not only have excess zeros but also show over-dispersion. Like ZIP, ZINB combines a logistic model for zero-inflation with a count model, but the count part is based on negative binomial distribution to accommodate over-dispersion.

4. Modeling

The data was randomly split into a training set (n = 44484) and a test set (n = 11121) with an 8:2 ratio to assess the model's ability to generalize to new data and to prevent overfitting. The training set was used to fit the models, while the test set was then used to evaluate how well the models generalize to new real-world data.

4-1. Linear Regression

We assessed the performance of various linear regression models on recent data. In this analysis, we performed linear regression modeling focusing on the prediction of the scaled_norm_fav variable. We ran the models as input and evaluates their performance by calculating several metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared values, Adjusted R-squared values, and Akaike Information Criterion (AIC) to understand the factors influencing these variables and to identify the best-fitting model.

Eleven different linear regression models have been constructed predicting scaled_norm_fav with independent variables: word_count, syuzhet, visual_presence, biz_presence, eco_presence. The models are Linear-Linear (Model 1), Linear-Linear (Model 2; the predictor variable is only visual_presence), Linear-Log (Model 3), Linear-Square Root (Model 4), Log-Linear (Model 5), Log-Linear (Model 6; the predictor variable is only visual_presence), Log-Log (Model 7), Log-Square Root (Model 8), Square Root-Linear (Model 9), Square Root-Log (Model 10), and Square Root-Square root (Model 11). Each model's equation is described in Table 1 in the Appendix.

We calculated the performance of each model to assess various aspects of its effectiveness. The results of this evaluation process are summarized in Table 2 located in the Appendix.

Among the evaluated linear regression models, Model 7 emerges as the most favorable choice, boasting the highest Adjusted R-Squared, which accounts for the number of predictors, reinforces the model's superiority. Moreover, Model 7 demonstrates competitive performance across other metrics. It achieves relatively lower values for both RMSE and MAE, indicating less deviation between predicted and actual values. Additionally, Model 7 exhibits a lower Akaike Information Criterion (AIC), signifying a good balance between model fit and complexity.

While these metrics point to the favorable performance of Model 7, it's important to acknowledge that the overall predictive power remains limited across all models. Further exploration of features and consideration of alternative modeling approaches may be valuable for refining predictions and uncovering more intricate patterns within the data.

The regression model was fit using the following formula: log_scaled_norm_fav = log_word_count + log_syuzhet + visual_presence + biz_presence + eco_presence

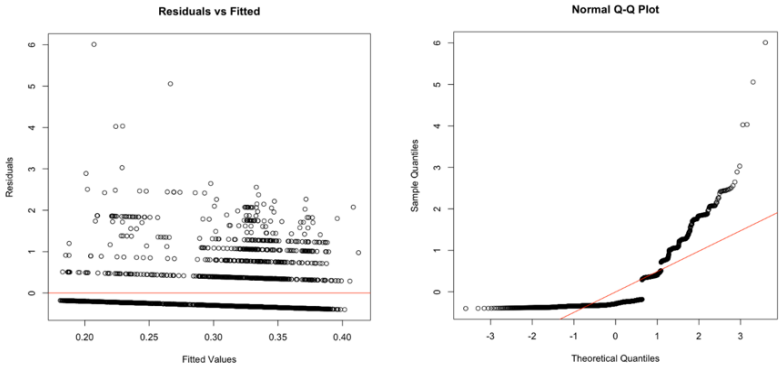
The predictors explained 6.76% of the variance (R^2 = .068, Adjusted R^2 = .0602). The model was statistically significant, $F(5, 3109) = 4.774$, p = .000238. The intercept, b = 0.482, $t(3109)$ = 4.237, p < .001, indicated that when all predictors are held at zero, the expected log of scaled normalized favorites is about 0.482. Visual_presence was found to be a significant negative predictor of log_scaled_norm_fav, b = -0.108, $t(3109)$ = -4.395, p < .001, suggesting that one unit increase in visual_presence is associated with a decrease of 0.108 in the log of scaled normalized favorites. None of the other variables were found to be significant predictors. Table 3 shows the details of the model.

Table 3. Linear Regression Model Results for log_scaled_norm_fav.

Variable	*b*	*SE*	*t*
Intercept	0.482***	0.114	4.237
log_word_count	-0.026	0.023	-1.133
log_syuzhet	-0.005	0.059	-0.077
visual_presence	-0.108***	0.025	-4.395
biz_presence	-0.037	0.024	-1.561
eco_presence	-0.043	0.025	-1.729

Notes: R^2 = 0.008, adj. R^2 = 0.006, F = 4.774 (p < .001). p<.05, **p<.01, ***p<.001.

The variables' VIF values are low, with log_word_count at 1.42, log_syuzhet at 1.32, visual_presence at 1.06, biz_presence at 1.19, and eco_presence at 1.07, indicating no significant multicollinearity concerns. However, the graphs below suggest that the linear regression assumptions, specifically the normality of residuals and homoscedasticity, may not be fully met.



4-2. Logistic regression

We applied logistic regression to ‘recent_data’ to predict the likelihood of ‘fav_thres’—a binary marker for 'favorite_count' meeting or exceeding the top 25%—was likely, based on various predictors. We fitted ten logistic regression models (glm_mod1 to glm_mod10) with different sets of predictors. We evaluated the performance of each model using various classification metrics at different probability thresholds (0.2, 0.3, 0.4, 0.5, 0.6). The metrics included Accuracy, Sensitivity, Specificity, Precision, F1_Score, and AUC. With a threshold above 0.6, many models predominantly predicted most data points as 0. This led to anomalies in some metrics (Precision, F1_Score, and AUC), resulting in undefined values for several models. Due to this observation, we decided not to proceed with higher thresholds, as the absence of these metrics could affect the comprehensive evaluation of model performance. Based on the provided tables for different probability thresholds (0.2, 0.3, 0.4, 0.5, 0.6), Table 4 presents the key findings.

Table 4. Performance at probability thresholds (0.2, 0.3, 0.4, 0.5, 0.6).

Threshold	Best Model	Accuracy	Sensitivity	Precision	F1 Score	AUC	AIC	Residual Deviance	Key Observations
0.2	glm_mod5	0.6547	0.9739	0.4030	0.5785	0.6582	3997.7	3985.7	High Sensitivity; Low Precision; Potential False Positives
0.3	glm_mod6	0.6650	0.9739	0.4450	0.6192	0.6579	3998.1	3988.1	Improved Precision; Maintains High Sensitivity
0.4	glm_mod1	0.6175	0.3420	0.5122	0.4375	0.6621	4010.4	3998.4	Increased Accuracy and Precision; Balanced Performance.
0.5	glm_mod6	0.6650	0.3485	0.6182	0.4505	0.6579	3998.1	3988.1	Highest Accuracy and Sensitivity; Relatively High Precision.
0.6	glm_mod7	0.6020	0.0130	0.3636	0.0251	0.6371	4023.5	4017.5	High Accuracy and Low Sensitivity; Modest Precision

As we want a balance between sensitivity and precision, we prioritized the 0.4 or 0.5 threshold models and think they are more suitable than others. In these thresholds, models 1 and 2 seem to perform well in terms of overall accuracy, specificity, and precision. Details of the results for predictors at the thresholds of 0.4 and 0.5 are provided in Appendix Tables 5 and 6.

When we examined logistic regressions for model 1 (fav_thres ~ word_count + syuzhet + visual_presence + biz_presence + eco_presence), all predictors, except for syuzhet, demonstrated statistical significance. The results are presented in Table 7 in Appendix.

When we executed Model 2 (fav_thres ~ word_count + visual_presence + biz_presence + eco_presence), the results in Table 8 suggest the effectiveness of the model, with all predictors being statistically significant.

Table 8. Logistic Regression Model Results for Model 2.

Variable	*b*	*SE*	*z*
Intercept	-0.255	0.149	-1.709
word_count	-0.015***	0.003	-4.972
visual_presence0	-0.818***	0.089	-9.182
biz_presence0	0.300***	0.080	3.737
eco_presence0	0.438***	0.090	4.856

Note: *p<.05, **p<.01, ***p<.001.

4-3. Zero-inflated models

Zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models were employed to analyze the influence of word_count, syuzhet, visual_presence, biz_presence, and eco_presence on scaled_norm_fav. These two models accommodated the overdispersion and excess zeros in the data.

ZIP Model: The model exhibited significant relationships for coefficients for word_count, syuzhet, visual_presence0, and biz_presence0. The zero-inflation part of the model (predicting excess zeros) was also significant. The model showed an Akaike Information Criterion (AIC) of 11858.06 and a Bayesian Information Criterion (BIC) of 11900.36. The log-likelihood was -5922.03 with 7 degrees of freedom. Pseudo R-squared was 0.1041, indicating a modest improvement in fit over a null model with only an intercept. The model achieved an RMSE of 1.9734 on test data, reflecting its predictive accuracy.

ZINB Model: Significant effects were found for word_count, syuzhet, and visual_presence0. The zero-inflation component was not significant. The model had an AIC of 6682.994, a BIC of 6731.336, and a log-likelihood of -3333.497 with 8 degrees of freedom. Pseudo R-squared was 0.0090, suggesting a slight improvement over the null model. The model recorded an RMSE of 1.7086, indicating better predictive accuracy compared to the ZIP model.

The ZINB model demonstrated a better fit than the ZIP model, as evidenced by lower AIC and BIC values and a lower RMSE, indicating higher predictive accuracy. The pseudo-R-squared values for both models were modest, with the ZINB model showing a slight improvement over its null model. Given the lower AIC, BIC, and RMSE, the ZINB model is a more appropriate choice for modeling scaled_norm_fav in this context.

Table 9. Zero-Inflated Negative Binomial Model Results for Predicting scaled_norm_fav

Variable	b	SE	Z
Count Model			
Intercept	-.503*	.209	-2.410
word_count	.011*	.005	2.435
syuzhet	-.096	.049	-1.941
visual_presence	.763***	.113	6.760
biz_presence	.112	.114	.983
eco_presence	-.071	.121	-.585
Log (theta)	-1.915***	.048	-39.755
Zero-Inflated Model			
Intercept	-9.039	23.270	-.388

Notes: Theta = .147; AIC = 6780.921; BIC = 6829.273; Log-likelihood = -3382.461; Pseudo R-squared = .0008; RMSE = 1.838. *p<.05, **p<.01, ***p<.001.

4-4. Model selection

In our data, a significant overdispersion is present, as indicated by the very high maximum Pearson residual. The presence of overdispersion is a clear sign that the assumptions of the linear regression model are violated, as linear regression assumes a constant variance of errors. While appropriate for binary outcomes, the logistic regression model may not be as suitable for count data that are not strictly binary or where the counts are not just zero and one but can take on a wider range of values. The zero-inflated negative binomial model is particularly suitable for count data that have an excess of zero counts, which are not well modeled by regression and standard Poisson models due to overdispersion and the presence of two different data-generating processes: one that generates only zeros, and another that generates count outcomes, including zeros. Therefore, among the models, a Log-Log regression model, logistic regression models, and zero-inflated negative binomial model discussed above, we chose the zero-inflated negative binomial model. The insignificant variables (biz_presence and eco_presence) were excluded from the chosen model, and our final model only included word_count, syuzhet, and visual_presence, to predict scaled_norm_fav. Table 10 shows the detailed results.

Table 10. Finalized Zero-Inflated Negative Binomial Model Results for Predicting scaled_norm_fav

Coefficient	b	SE	Z
Count Model			
Intercept	-.475***	.209	-2.410
word_count	.010*	.005	2.435
syuzhet	-.101*	.049	-1.941

visual_presence	.772***	.113	6.760
Log (theta)	-1.916***	.048	-39.874
Zero-Inflated Model			
Intercept	-10.49	48.02	-.218

Notes: Theta = .147; AIC = 6678.247; BIC = 6814.511; Log-likelihood = -3383.124; Pseudo R-squared = .0007; RMSE = 1.834. *p<.05, **p<.01, ***p<.001.

4-5. Hypothesis Verification

The first hypothesis expected the positive association of the sentiment of a social media message with the number of likes toward the message. Our final model indicated that message sentiment significantly predicts the number of likes (b = -.101, p < .05), but the influence direction was opposite to our initial expectation. That means the more negative the emotion of a social media message, the more likes the message generates. Thus, H1 was not supported.

The second hypothesis was about the negative association of the length of a social media message with the number of likes toward the message. The results showed that word count increases the number of likes (b = .010, p < .05), which is the opposite direction from the hypothesis. The amount of textual information in a social media message increases likes toward the message. Thus, H2 was not supported.

The third hypothesis argued that the presence of a visual attachment in a social media post increases the likes toward the message. The results revealed that placing an image or video with a social media post significantly increases the number of likes toward the message (b = .772, p < .001). Thus, H3 was supported.

The fourth hypothesis assumed the positive impact of the presence of business-related or environment-related words in a social media message on the number of likes toward the message. The results indicated the presence of the words did not make a significant difference in the number of likes. Thus, H4a and H4b were not supported.

5. Conclusion

In this project, we employed four different modeling techniques: linear regression, logistic regression, zero-inflated Poisson regression, and zero-inflated negative binomial regression. A significant challenge we encountered was the highly skewed nature of our dataset, characterized by a large number of zero like counts. Despite our efforts in variable transformations, the methods we learned in class, such as linear regression and logistic regression, faced limitations in addressing this skewness. This challenge led us to venture beyond the scope of our coursework and explore zero-inflated models. Our analysis, grounded in the zero-inflated Poisson regression model, led us to conclude that while hypotheses 1, 2, and 4 were not supported, hypothesis 3 was validated. This was part of our investigation into how sentiment, message length, visual elements, and business or environmental-related words in social media posts influence the number of likes received. Despite the majority of hypotheses yielding results contrary to our expectations, three variables—sentiment, message length, and visual elements—demonstrated statistical significance at a 5% level. Notably, the role of visual elements was especially significant, achieving a significance level of 0.1%.

Based on these findings, businesses can effectively increase the number of likes on their social media posts by incorporating visual elements such as images or videos. This is evidenced by the 'visual_presence' variable, which shows that transitioning from 0 (no visual content) to 1 (presence of visual content) can lead to an impressive 116% increase in likes. This underscores the powerful impact of visual content on social media engagement. In addition, our study reveals that a one-unit increase in 'syuzhet', indicating a shift towards more positive sentiment, correlates with a 9.6% decrease in likes. This suggests that while positive sentiment is essential, overly positive sentiments might not resonate strongly with the audience. Moreover, extending the word count of a post by one word results in an estimated 1% increase in likes. These

insights provide valuable guidelines for businesses to effectively tailor their social media content for maximum engagement.

Limitations and Future Research Ideas:

- **Content of Visual Attachments:** The inclusion of visual attachments emerged as the most significant variable in our study. To further understand what types of visual content generates more consumer engagement, additional research focusing on the characteristics of these attachments is necessary.
- **Varying Social Media Platforms:** Different social media platforms may have users with varying behaviors and preferences, which could limit the generalizability of our findings across all platforms. Brands that predominantly use social media platforms other than Twitter (X) might benefit from conducting similar research using datasets specific to their primary platforms.
- **Follower Count Influence:** Follower count is a crucial factor in assessing consumer engagement. Unfortunately, our project did not explore the extent to which social media activities influence follower growth. Future studies could focus on this aspect to provide a more comprehensive understanding of engagement dynamics.
- **Impact on Sales:** Access to a brand's sales data, along with metrics such as like counts and follower growth, could offer insights into how social media activities contribute to overall sales performance. This would be particularly valuable for testing models like AIDA, provided other variables like the impact of significant discount events are well controlled. Such analysis could help in understanding the direct correlation between social media strategies and sales outcomes.

6. Reference

[1] Shahizan Hassan, Siti Zaleha Ahmad Nadzim, Norshuhada Shiratuddin (2015), Strategic Use of Social Media for Small Business Based on the AIDA Model, Procedia - Social and Behavioral Sciences Volume 172, Pages 262-269

[2] Bambang Sukma Wijaya (2011), The Development of Hierarchy of Effects Model in Advertising, International research journal of business studies pages 73-85

[3] Hassan, S., Nadzim, S. Z. A., & Shiratuddin, N. (2015). Strategic use of social media for small business based on the AIDA model. Procedia-Social and Behavioral Sciences, 172, 262-269.

[4] Dr.C.Kathiravan (2017), Effectiveness SNS (Social Network Sites) Advertisements on Brand image, International Journal of Research Volume 04 Issue 09

[5] Hutter, K., Hautz, J., Dennhardt, S., & Füller, J. (2013). The impact of user interactions in social media on brand awareness and purchase intention: the case of MINI on Facebook. Journal of product & brand management, 22(5/6), 342-351.

[6] Washington, S., Karlaftis, M. G., Mannering, F., & Anastasopoulos, P. (2020). Statistical and econometric methods for transportation data analysis. CRC press.

7. Appendix

4-1. Linear Regression

Table 1. Linear Regression Model (Model 1 - Model 11) with Variables - Predicting scaled_norm_fav

	Predictor Variables
Model 1 (Linear:Lineal)	word_count, syuzhet, visual_presence, biz_presence, eco_presence
Model 2 (Linear:Linear)	visual_presence
Model 3 (Linear:Log)	log_word_count, log_syuzhet, visual_presence, biz_presence, eco_presence
Model 4 (Linear:Square Root)	Predictor Variables: sqrt_word_count, sqrt_syuzhet, visual_presence, biz_presence, eco_presence
Model 5 (Log:Linear)	Predictor Variables: word_count, syuzhet, visual_presence, biz_presence, eco_presence
Model 6 (Log:Linear)	visual_presence
Model 7 (Log:Log)	log_word_count, log_syuzhet, visual_presence, biz_presence, eco_presence
Model 8 (Log:Square Root)	sqrt_word_count, sqrt_syuzhet, visual_presence, biz_presence, eco_presence
Model 9 (Square Root: Linear)	word_count, syuzhet, visual_presence, biz_presence, eco_presence
Model 10 (Square Root:Log)	log_word_count, log_syuzhet, visual_presence, biz_presence, eco_presence
Model 11 (Square Root:Square Root)	sqrt_word_count, sqrt_syuzhet, visual_presence, biz_presence, eco_presence

Table 2. Performance of Model1 - Model 11

	FMSE	MAE	R_Squared	Adjusted_R^2	AIC	Log_Likelihood
recent_lm_mod1	1.8374	1.2280	0.0016	-2.6307	23191.6	-11588.8
recent_lm_mod2	1.8372	1.2264	0.0013	1.0196	23184.4	-11598.2
recent_lm_mod3	1.8411	1.2320	0.0015	-7.6481	23191.8	-11588.9
recent_lm_mod4	1.8395	1.2302	0.0015	-4.4955	23191.7	-11588.8
recent_lm_mod5	1.8261	0.8760	0.0072	5.6386	5695.63	-2840.82
recent_lm_mod6	1.8269	0.8763	0.0059	5.6014	5691.76	-2842.88
recent_lm_mod7	1.8267	0.8755	0.0076	6.0239	5694.43	-2840.21
recent_lm_mod8	1,8266	0.8758	0.0073	5.7408	5695.31	-2840.65
recent_lm_mod9	1.7977	0.9580	0.0011	-4.2913	6470.54	-3228.27
recent_lm_mod10	1.7993	0.9578	0.0009	-6.9770	6471.38	-3228.69
recent_lm_mod11	1.7988	0.9580	0.0009	-6.3748	6471.19	-3228.59

4-2. Logistic regression

Table 5. Results for Predictors at the threshold of 0.4

	AIC	Residual Deviance	Accuracy	Sensitivity	Specificity	Precision	F1 score	AUC	Log_Likelihood	Pseudo_R2	Adjusted_R2
glm_mod1	4010	3998.39	0.6174	0.6156	0.6186	0.5121	0.4375	0.6621	-1999.1	0.0417	0.0355
glm_mod2	4010	4000.01	0.6110	0.5895	0.6250	0.5055	0.4430	0.6610	-2000.0	0.0413	0.0364
glm_mod3	4044	4035.94	0.5994	0.5863	0.6080	0.4931	0.4033	0.6476	-2017.9	0.0327	0.0290
glm_mod4	4042	4036.01	0.5969	0.5700	0.6144	0.4901	0.4033	0.6475	-2018.0	0.0327	0.0302
glm_mod5	3998	3985.73	0.6174	0.6221	0.6144	0.5120	0.4312	0.6581	-1992.8	0.0447	0.0386
glm_mod6	3998	3988.13	0.6071	0.5732	0.6292	0.5014	0.4505	0.6579	-1994.0	0.0442	0.0392
glm_mod7	4024	4017.57	0.5892	0.5504	0.6144	0.4814	0.3747	0.6371	-2008.7	0.0371	0.0346
glm_mod8	4003	3991.34	0.6161	0.6221	0.6122	0.5106	0.4322	0.6597	-1995.6	0.0434	0.0372
glm_mod9	4004	3993.57	0.6059	0.5798	0.6228	0.5000	0.4500	0.6596	-1996.7	0.0429	0.0379
glm_mod10	4031	4025.31	0.5879	0.5537	0.6101	0.4802	0.3991	0.6421	-2012.6	0.0353	0.0328

Table 6. Results for Predictors at the threshold of 0.5

	AIC	Residual_Deviance	Accuracy	Sensitivity	Specificity	Precision	F1 score	AUC	Log_Likelihood	Pseudo_R2	Adjusted_R2
glm_mod1	4010	3998.39	0.6534	0.3420	0.8559	0.6069	0.4375	0.6621	-1999.1	0.0417	0.0355
glm_mod2	4010	4000.01	0.6546	0.3485	0.8538	0.6079	0.4430	0.6610	-2000.0	0.0413	0.0364
glm_mod3	4044	4035.94	0.6392	0.3094	0.8538	0.5792	0.4033	0.6476	-2017.9	0.0327	0.0290
glm_mod4	4042	4036.01	0.6392	0.3094	0.8538	0.5792	0.4033	0.6475	-2018.0	0.0327	0.0302
glm_mod5	3998	3985.73	0.6546	0.3322	0.8644	0.6144	0.4312	0.6581	-1992.8	0.0447	0.0386
glm_mod6	3999	3988.13	0.6649	0.3485	0.8707	0.6369	0.4505	0.6579	-1994.0	0.0442	0.0392
glm_mod7	4024	4017.57	0.6444	0.2703	0.8877	0.6102	0.3747	0.6371	-2008.7	0.0371	0.0346
glm_mod8	4003	3991.34	0.6559	0.3322	0.8665	0.6181	0.4322	0.6597	-1995.6	0.0434	0.0372
glm_mod9	4004	3993.57	0.6611	0.3517	0.8622	0.6242	0.4500	0.6596	-1996.7	0.0429	0.0379
glm_mod10	4031	4025.31	0.6482	0.2964	0.8771	0.6107	0.3991	0.6421	-2012.6	0.0353	0.0328

Table 7. Logistic Regressions Results for Model 1 (fav_thres ~ word_count + syuzhet + visual_presence + biz_presence + eco_presence)

Variable	*b*	*SE*	*z*
Intercept	-0.270	0.150	-1.805
word_count	-0.017***	0.003	-5.068
syuzhet	0.046	0.036	1.274
visual_presence	-0.834***	0.090	-9.266
biz_presence	0.321***	0.082	3.916
eco_presence	0.445***	0.090	4.922

Note: *p<.05, **p<.01, ***p<.001.