Sentiment Analysis and User Behavior Prediction in Social Networks

LIU MINGJIE

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 4

INTRODUCTION

4.1 Introduction

According to experts, social media data has turned into a rich mine of data, which reveals customer sentiment as well as behavior. The purpose of this chapter is geared towards bringing into focus the models building technique that is both accurate to predictions and comprehensive in results on sentiment analysis and user behavior prediction. Firstly, the study will be focused on two pairs of problems: the dependence of image modality on emotional polarity and the link that connects emotional polarity to particular brands. The answers to these two questions are the two essential components that form part of the qualitative and quantitative parameters of user behavior that are critical in developing the ultimate social media strategy.

4.2 Exploratory Data Analysis (EDA)

As an initial step in the exploration of the data, EDA was the technique used. The end goal of Exploratory Data Analysis (EDA) is to produce statistical graphics that provide intuitive ability to describe the key population characteristics, anomalies, patterns underlying the mass of the data, which will guide subsequent modeling and hypothesis testing.

4.2.1 Dataset Description

1.The relationship between image modality and affective polarity

An Emotion dataset: This archive consists of a vast variety of Tweets, with different modalities, such as images and text. Also, the polarity labels are declared, indicating the sentiment of the given Tweet. With the assistance of this dataset, analyzing the impact of the different modes of visual ways on emotional polarity would be relayed out to us. There is a gamut of datasets, as we sample them from various sources in our efforts to experiment and verify our hypotheses.

	Twitter ID	Tweet content	Image	Label the	modality	polarity	emotion
2	1483637809449771	kate and toby's smoker heating up to destroy their marriage #ThisIsUs https://t.co/C2W	http://10.112.67.227:8666/img/download/1	3	image	negative	["anger"]
3	1498163459749711	@VIVIZ_staff All the best sinb, eunha, umji #VIVIZ #Queendom2 https://t.co/jEbTMa	http://10.112.67.227:8666/img/download/1	3	image	positive	["joy"]
4	1499194611511791	does your heart ever go : ♡ ♡ ♡ ♡ ♡ ♡ ♡ ♡ joshu	http://10.112.67.227:8666/img/download/1	3	image	neutral	["anyway"]
5	1499196380400781	@MrNiceGuy513 @Seedalicious @Great_Katzby @POCculture @GailSimone @SuperSuit	http://10.112.67.227:8666/img/download/1	3	image	neutral	["confused"]
6	1499212575128671	"Dont stare at your dmen after every goal against"Mrazek: https://t.co/JvKKa4WpBE	http://10.112.67.227:8666/img/download/1	3	image	negative	["anger","surprise"]
7	1499215363220951	going back and forth on Campbell and Mrazek https://t.co/uBpCZ6NZnJ	http://10.112.67.227:8666/img/download/1	3	image	negative	["anger","disgust"]
8	1499767722355431	@LangmanVince Awww. U sad unemployment is 3.8% and almost 700k jobs added. http	419b-9234-8352cc5fb406.gif	3	image	neutral	["confused"]
9	1499831024141931	Our Little Cat #SUGADAY https://t.co/s4dcnZFOB1	http://10.112.67.227:8666/img/download/1	3	image	negative	["fear","disgust"]
10	1499918578014821	@DrOz 7.4 million jobs created and 3.8 unemployment rate. That's a comeback https://t	http://10.112.67.227:8666/img/download/1	3	image	negative	["malice"]
11	1499928896241741	@donwinslow @RonnaRono Republicans woke up EXTRA early this morning, eagerly ant	http://10.112.67.227:8666/img/download/1	3	image	negative	["sadness"]
12	1500084459047361	@Joshua_M_Hump I do wanna https://t.co/EZNvsiN2wp	http://10.112.67.227:8666/img/download/1	3	image	neutral	["neutral"]
13	1500222764229301	@KEEMSTAR Anthony Joshua watching the fight https://t.co/XLJ4453UGr	http://10.112.67.227:8666/img/download/1	3	image	positive	["joy"]
14	1500233387361141	@PierrePoilievre You included Charest in this tweet? Are you getting paranoid Skippy Pol	http://10.112.67.227:8666/img/download/1	3	image	positive	["joy"]
15	1500528052417271	Ella really had the AUDICITY to start crying and say "I feel really bad now" after telling Jo	http://10.112.67.227:8666/img/download/1	3	image	negative	["fear","disgust"]
16	1500537646950591	#DeFi created for the people, is it already Time for communities to drive DeFi in every pc	http://10.112.67.227:8666/img/download/1	3	image	neutral	["neutral"]
17	1500652195804501	Yo! Mike's friend is the fucking man. We all need friends who is blunt and honest. #90Da	http://10.112.67.227:8666/img/download/1	3	image	positive	["joy"]
18	1500655275732251	Buying the Lakers for $67 million in the 80s. Probably seemed steep back then, but what	http://10.112.67.227:8666/img/download/1	3	image	neutral	["neutral"]
19	1500662828025401	No Ben, the person you are speaking with does not a relationship with you. She wanted	http://10.112.67.227:8666/img/download/1	3	image	negative	["fear"]
20	1500663336819641	I'm open to reviewing #WinningTime — figured I just put it out there. https://t.co/GhjFD	http://10.112.67.227:8666/img/download/1	3	image	positive	["joy"]
21	1500665242447221	Lord pls don't let me be 50+ crying over a wannabe rapper on TV. #90dayfiancebeforeth	http://10.112.67.227:8666/img/download/1	3	image	neutral	["neutral"]
22	1500665670735831	Can someone explain to me how Michigan is the country and why Norm Nixon, a man fr	http://10.112.67.227:8666/img/download/1	3	image	negative	["confused"]
23	1500666907611311	Jerry West every time he's on camera #WinningTime https://t.co/h6voxz6Ird	http://10.112.67.227:8666/img/download/1	3	image	negative	["sadness","anger"]
24	1500668225931121	Ximena's face when she realized the gravy train may be over #90dayfiancebeforethe90d	http://10.112.67.227:8666/img/download/1	3	image	negative	["anger","surprise",
25	1500668337373821	Jerry West in every shot #WinningTime https://t.co/vWQ8Web8Vz	http://10.112.67.227:8666/img/download/1	3	image	negative	["anger","surprise"]
26	1500668548863201	Ximena: Mike you're disgusting, I'm not in love with you, stay away from meAlso Ximena	http://10.112.67.227:8666/img/download/1	3	image	neutral	["confused"]
27	1500670417597391	Don't even feel sorry for delusional women like Kimberly anymore. Look in the damn mir	http://10.112.67.227:8666/img/download/1	3	image	neutral	["surprise"]
28	1500672194157891	Jerry West whenever he saw his Finals MVP trophy #WinningTime https://t.co/VpF2Yl	http://10.112.67.227:8666/img/download/1	3	image	positive	["joy"]
29	1500673118154601	Me after watching #WinningTime https://t.co/lvKpC0rHmN	http://10.112.67.227:8666/img/download/1	3	image	neutral	["surprise"]

Figure 4.1 Anote Dataset

Figure 4.1 Anote Dataset is the first 28 rows of data in the Anote dataset

Furthermore, we have a pre-processing phase when analyzing data that includes the mentioned steps. Delete the image-containing and text-containing Tweets for the ease of specific analysis. Clean data to eliminate noise and other irrelevant information for good quality of the sanitized data. The polarity of emotions is then granted to indulge in tweets for sake of effects and the remaining part of the analysis.

Here are the following methods which are used:

(1)How a chi-square test would be applied to compare the distribution of sentiment polarity of images as opposed to plain text and show the results.

a) Construct the Contingency Table

Construct the Contingency Table First, form a frequency table according to the sentiment polarity. In particular, let M indicate Modalities (Image, Composite, Text) and let S represent Sentiments (Negative, Neutral, Positive):

	Negative	neutral	positive	Row Total
image	1801	1666	1326	4793
synthesis	1301	250	752	2303
text	897	609	1060	2566
Column Total	4000	2525	3138	10663

b)Calculate the Expected Frequencies

Under the assumption of independence, we expect to observe frequency values in each cell. The formula below is for the expected frequency:

$$E_{ij} = \frac{(Row_i\ Total) \times (Column_j\ Total)}{Total\ Sample\ Size}$$

c)Calculate the Chi-Square Statistic

Calculate the observed and expected frequencies which are used in the calculations of Chi-Square. The formula is:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

In this formula, O_{ij} represents the observed frequency and E_{ij} represents the expected frequency. We do the calculations for each piece of data and then add up all the results.

d)As can be seen from the results of the calculation, the chi-square statistic is 613.04 and the p-value is 2.34×10^-131. This illustrates that, assuming that modal polarity and affective polarity are independent, the likelihood of observing a contingency table structure than this or more outrageous is very low. So, we reject the null hypothesis that there are significant differences in the distribution of sentiment polarity between tweets representing different modalities.

(2)Creating bar charts to bring visualization to results while helping to the readers process information faster.

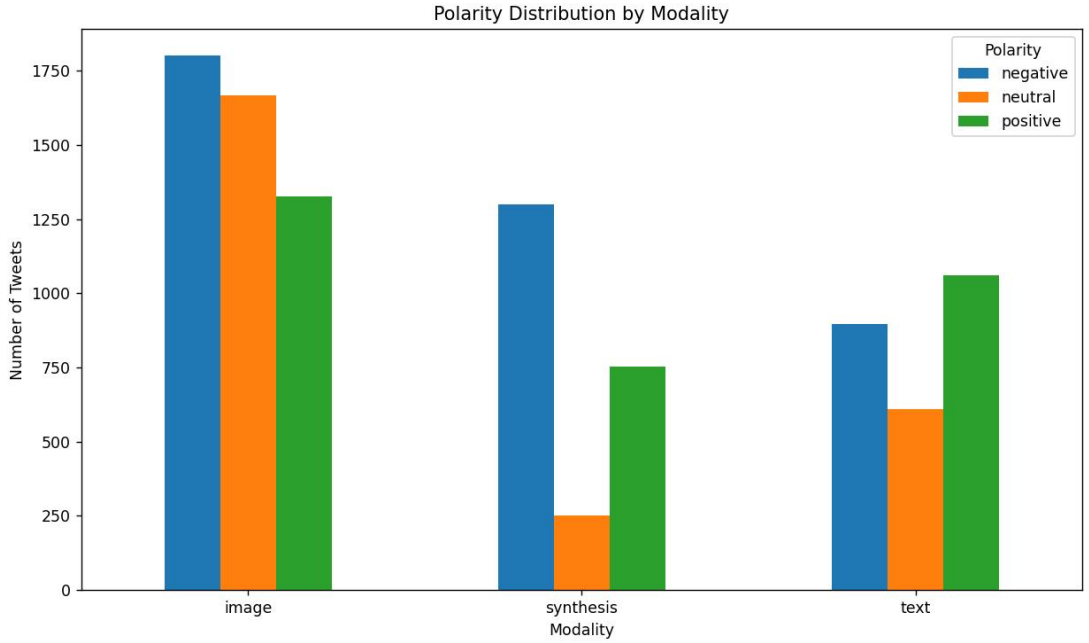

Figure 4.2 Polarity Distribution by Modality

Figure 4.2: Polarity Distribution by Modality shows the spread of emotional polarity by mode (image, composite, text) varies. While the number of tweets with negative sentiment in the imagery form is top with that of text form coming next, this is self-evident from this trend graph. It is the synthetic modality that scores in the top position in terms of the neutral sentiment tweets. The trends of positive and negative tweets across the modalities show that, except for the synthetic modality, the positive

sentiment(omit) numbers are higher when compared to the negative sentiment numbers across the twitter feed. This conclusion can be drawn that people prefer to express positivity in visuals and text.

(3)Evidences from analysis suggests that, actually, there is significant difference in sentiments of emotion between image and plain-text styled tweets. Classic text tweets, for example, show a slightly higher positive mood, which confirms our assumption of local insights. Visual representation also demonstrates this very point, capturing the unique role played by visuals in helping us to express feelings. Such reports add evidence on the significance of image content in social media sentiment analysis through experimentalist inference. Picture serves as a particular brand feature which is associated with either positive or negative emotions of customers as well.

2. Emotional polarity relates to a particular brand

Twitter_company1 dataset: This data offers Tweets from several brands, where we will hold a sentiment analysis based on social media data analysis. While this data enables us to gain real-life market orientation, it contributes to the progress of the analysis and the model.

	A	B	C	D	E	F	G
1	Twitter II	Content Theme			Tweet content	modality	polarity
2	2401	Borderlands	I am coming to the borders and I will kill you all,			text	Positive
3	2401	Borderlands	im getting on borderlands and i will kill you all,			text	Positive
4	2401	Borderlands	im coming on borderlands and i will murder you all,			text	Positive
5	2401	Borderlands	im getting on borderlands 2 and i will murder you me all,			text	Positive
6	2401	Borderlands	im getting into borderlands and i can murder you all,			text	Positive
7	2402	Borderlands	So I spent a few hours making something for fun. . . If you don’t know I am a HUGE @Borderl			text	Positive
8	2402	Borderlands	So I spent a couple of hours doing something for fun... If you don’t know that I’m a huge @			text	Positive
9	2402	Borderlands	So I spent a few hours doing something for fun... If you don’t know I’m a HUGE @ Borderland			text	Positive
10	2402	Borderlands	So I spent a few hours making something for fun. . . If you don’t know I am a HUGE Rhandler			text	Positive
11	2402	Borderlands	2010 So I spent a few hours making something for fun. . . If you don’t know I am a HUGE Rha			text	Positive
12	2402	Borderlands	was			text	Positive
13	2403	Borderlands	Rock-Hard La Varlope, RARE & POWERFUL, HANDSOME JACKPOT, Borderlands 3 (Xbox) dlvr.it/RMTrg			text	Neutral
14	2403	Borderlands	Rock-Hard La Varlope, RARE & POWERFUL, HANDSOME JACKPOT, Borderlands 3 (Xbox) dlvr.it / RMT			text	Neutral
15	2403	Borderlands	Rock-Hard La Varlope, RARE & POWERFUL, HANDSOME JACKPOT, Borderlands 3 (Xbox) dfr.it / RMTr			text	Neutral
16	2403	Borderlands	Rock-Hard La Vita, RARE BUT POWERFUL, HANDSOME JACKPOT, Borderlands 1 (Xbox) dlvr.it/RMTrgF			text	Neutral
17	2403	Borderlands	Live Rock - Hard music La la Varlope, RARE & the POWERFUL, Live HANDSOME i JACKPOT, Borderl			text	Neutral
18	2403	Borderlands	I-Hard like me, RARE LONDON DE, HANDSOME 2011, Borderlands 3 (Xbox) dlvr.it/RMTrgF			text	Neutral
19	2404	Borderlands	that was the first borderlands session in a long time where i actually had a really satisfy			text	Positive
20	2404	Borderlands	this was the first Borderlands session in a long time where i actually had a really satisfy			text	Positive
21	2404	Borderlands	that was the first borderlands session in a long time where i actually had a really satisfy			text	Positive
22	2404	Borderlands	that was the first borderlands session in a long time where i actually enjoyed a really sat			text	Positive
23	2404	Borderlands	that I was the first real borderlands session in a nice long wait time where i actually had			text	Positive
24	2404	Borderlands	that was the first borderlands session in a hot row where i actually had a really bad comba			text	Positive
25	2405	Borderlands	the biggest dissappoinment in my life came out a year ago fuck borderlands 3			text	Negative
26	2405	Borderlands	The biggest disappointment of my life came a year ago.			text	Negative
27	2405	Borderlands	The biggest disappointment of my life came a year ago.			text	Negative
28	2405	Borderlands	the biggest dissappoinment in my life coming out a year ago fuck borderlands 3			text	Negative
29	2405	Borderlands	For the biggest male dissappoinment in my life came hanging out a year time ago fuck border			text	Negative

Figure 4.3 Twitter_company1 Dataset

Figure 4.3 Twitter_company1 Dataset is a few dozen rows of data

Data preprocessing

The data preprocessing steps include:

Pick out my brand's tweets from the included datasets (Amazon, Microsoft, Borderlands, Google) to ensure the focus of my analysis.

The following analysis methods have been used:

(1) Sentiment Score Analysis

Sentiment Score Analysis In order to calculate the sentiment score for each Tweet, we determine an emotion score for each one so that the tweet contains a sentiment score that is included in the dataset.

This average sentiment score that will be computed as well as standard deviation in order to measure the sentiment polarity of brand tweets.

a)First, we should denote a numerical value that will correspond to an affective polarity. Positive:1, 0 neutral, -1 negative We then chose four brand. We have chosen Amazon, Microsoft, Borderlands, and Google. Calculate the average sentiment score, as well as the standard deviation for each brand. Based on the values that have been calculated above, create a box plot, as in Figure 4-4 below:

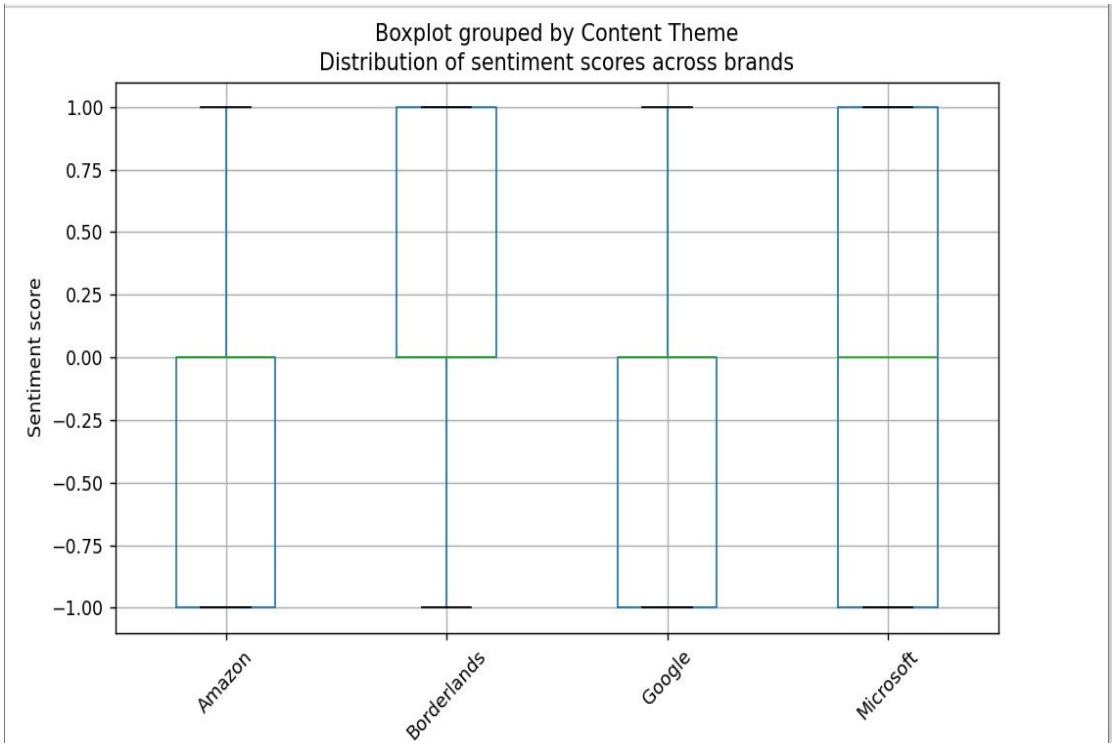

Figure 4.4 Distribution of sentiment scores across brands

b)Brand’ s summed-up sentiment score is a quick indicator of the sentiment that is inherited by the user whenever he or she interacts with it on social media. As exhibited by Borderlands' sentiment score of positive affirmation, users mostly have good opinions on Borderlands; while Amazon and Google's negative affirmation score shows that it's not populated with users, having the number of negative opinions. Microsoft’s overall sentiment tends to be close to neutral yet slightly to the lower side of neutral, which means more branching into the negative direction compared to the other brands. Here, the standard deviation of sentiment reflects fast varying of user sentiment score, which is stability of the emotion or in other words. Standard deviation of cyclicality is low for Amazon and Google, signaling that consumed emotional bias toward these brands is more stable. Such variations largely due to their relative young age as well as business is affected only by major product launches and marketing push versus the established players whose sentiment variabilities are less because, in contrast, they are affected by small-scale local news as well. These readings not only show to what extent positive or ^negative emotions are attributed to the brand’ s image on social media but also serve as an avenue to explore the essential inputs which would be used for further hypothesis testing on future customer behavior predictions.

(2)Emotional polarity distribution analysis

Descriptive Statistics: By calculating the frequency and percentage of positive, negative, neutral tweets for a particular brand.

Visualization: Use bar charts to show the distribution of different sentiment polarities.

a)The number and percentage of sentiment polarity (being positive, negative, neutral, irrelevant) inside brand Tweets were also found. The sentiment polarity is used to know the frequency and the nature of sentiment of the peoples towards the product. The following is a bar chart of the emotional polarity distribution of each brand in every hour of a day at the busiest traffic time of the day In Figure 4-5:

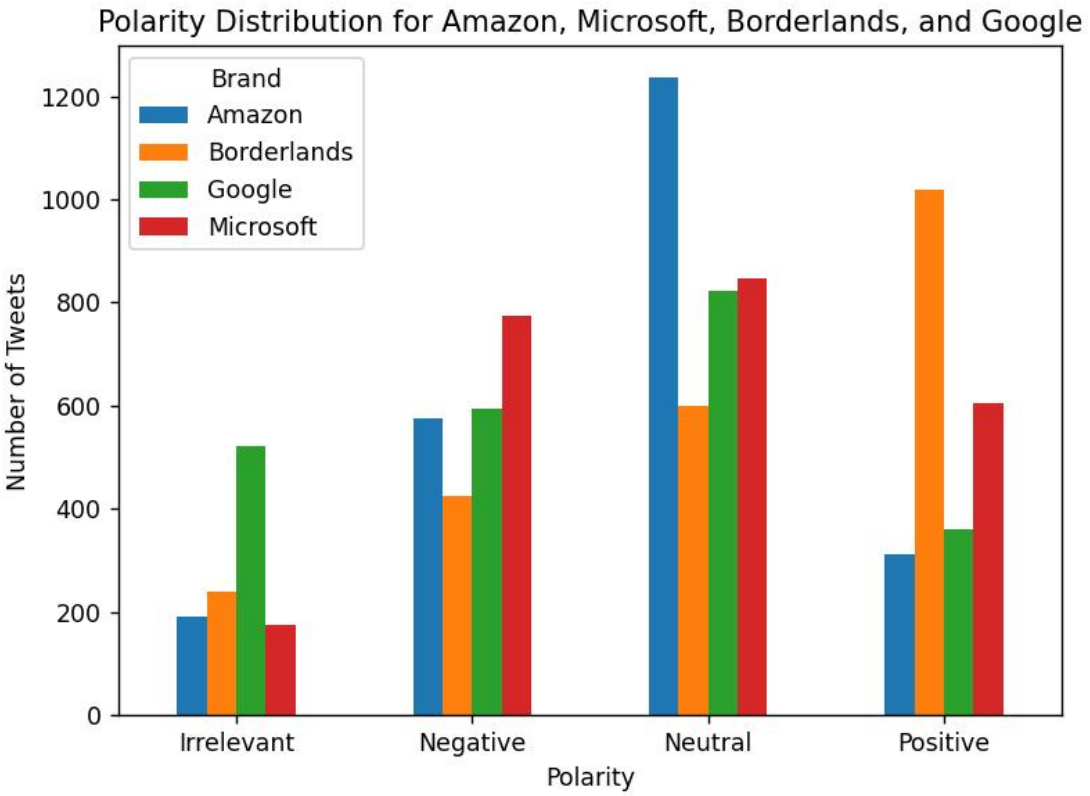

Figure 4.5 Polarity Distribution for Amazon, Microsoft, Borderlands, and Google

From the data in Figure 4-5, the graph illustrates the polarization of emotions between brands on social media, which makes it easier for us to see the difference. Among them, Borderland has the highest proportion of positive emotional polarity, indicating that users have a very high favorability towards it. However, Google's negative sentiment polarity ratio is higher than that of Borderland, indicating that users have more negative perceptions of them. Microsoft's sentiment polarity distribution is more balanced, but negative sentiment polarity still accounts for a large percentage.

4.3 Sentiment Analysis and User Behavior Prediction

4.3.1 Research Hypothesis

The sentiment analysis indicated the effect of sentiment on this behavior, and thus a significant correlation was found between sentiment analysis and user behavior. More concretely, we expect widespread fun spreading to positively challenge users to be more social through more comments, likes, and retweeting. At the same time, the ones without a positive emotion can hinder the social democracy operations by, in essence, churning good customers.

4.3.2 Feature Engineering

For the results of sentiment analysis, I built some features for user behavior prediction. These include the following:Average sentiment score, Standard deviation of sentiment scores, Frequency and percentage of different affective polarities (positive, negative, neutral, uncorrelated).

4.3.3 Predictive Models

We utilized the models that include (tuned hyperparameters for) logistic regression, random forests, and other ML techniques in their development (e.g., logistic regression, random forests) for the prediction of user behavior. Model variables are the key inputs required.

4.4 Model Development

Argentina Modeling Dependent on the results of the research hypothesis and the characteristics constructed from sentiment analysis, we will proceed with the creation of forecasts models for user behavior. Next, we are going to explain the next steps in more detail:

Model Selection: Our aim is to investigate multiple machine learning models that are knowledgeable and capable of predicting human online behavior, such as Logistic Regression: It is one of the linear models we utilized. As well, it is a model used for binary classifications making sure the predictions are binary. For instance, could the user retweet our tweet or don't? Random Forests: It will involve the joint cross-learning of multiple decision trees to raise the accuracy and robustness of a model.

Support Vector Machines (SVM): This is a comprehensive classification algorithm with the facility to solve complex relationships and deal with non-linear data.

Deep Learning Models: Curved neural networks lengthen to the types of structure of recurrent neural networks (RNNs) or convolutional neural networks (CNNs), which capture sophisticated readers' mood for text figuring out and user data. The model is hosted on top of those fine-tuned data.

Model Training: We will create rehearsal and verification datasets. The subject data should be run on the training group and latterly on the verification set for controlling.

Hyperparameter Tuning: It is the task of setting the hyperparameters of the model properly which significantly affects the efficacy or the opposite. It's here we perform a hyperparameter tuning as to find the optimum settings that will lead to the best model performance.

Model Validation: Check out the trained model's performance using various metrics, such as accuracy, precision, recall, F1, ROC AUC, among others. Thereby, it

helps us to select a suitable model from other existing models to be used in respect to the task.

Model Interpretability: To maintain the applicability of our models, the technologies, for instance, Shap values or feature importance matrix could be utilized to figure out their influence on the model′s estimates.

4.5 Model Evaluation

All the developed models evaluation is crucial when it comes to conferring the same and generalizability of the models. There will be a use of different evaluation measures depending on posing the behavioral phenomenon.

Binary Classification: If the task of the retweeting or the liking is to be predicted, then the precision will be the indicators selected.

Multi-Class Classification: When complex issues with multiple groups of behavior are treated, we will also look at frequency rates as well, in addition to accuracy, precision, recall, and F1-score and confusion matrix.

Regression: Metrics like MSE, MAE, and explained variation (R^2) are judged to predict the number of retweets or likes over the retweets or likes. Moreover, we will also add cross-validation to counter the risk of the models overfitting and check their robustness to avoid overfitting. Varca is the method of splitting the data into multiple subsets and testing it on different combinations of these subsets.

4.6 Expected Outcome

Knowing importantly the different patterns and trends that get represented as user's feelings and emotions during social networks. However, some linguistic features, user activity, and user metadata are the primary elements shaping such users′ behavioral activity. Design different predictive models that accurately bring the

impression of the sentiment and behavior of the user, as well as they are very reliable. Data from around the social networks can be exploited to a better understanding and informing the social networks, thus requiring many of the available applications in marketing, public health management, and social science research.

4.7 Summary

In the investigation in this chapter, the relations between image modality and affective polarity as well as the brand affect polarity are explored. These discoveries let us expand our knowledge horizon about social media data analysis, but they also provide guiding lights for future research. Even though with few restrictive points, the present research paper has given an outline of what to do to filter valuable feeling and emotions from the content of social media. The conclusion section will briefly sum up the main contributions of the research and suggest some possible research areas for the future.