Sentiment Analysis and User Behavior Prediction in Social Networks

LIU MINGJIE

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 4

INTRODUCTION

4.1 Introduction

This chapter presents the exploratory data analysis that was done to understand the main dataset and get an initial feel related to the research. EDA is among the most important steps of any data science project because it helps in identifying patterns, detecting anomalies, forming hypotheses, and validating assumptions by summary statistics and visual representations. This chapter will first present the EDA process, followed by the details of the visualizations and descriptive statistics used in the exploration of data. Initial insights from the EDA and feature engineering processes are also presented. Finally, the chapter concludes with a summary emphasizing how important these initial findings are to drive the subsequent phases of the research.

4.2 Exploratory Data Analysis (EDA)

EDA is a major step to comprehend the underlying structure of the data, find patterns, identify anomalies, and create hypotheses. The following subsections present the visualizations, descriptive statistics, preliminary insights, and feature engineering that were conducted during the EDA phase.

4.2.1 Visualizations and Descriptive Statistics

To understand the data, various visualizations and descriptive statistics were created. Important visualizations include:

Word Clouds: To visualize most frequent words in user posts and comments.

Bar Charts: distribution of sentiment-whether positive, negative, or neutral-across time or user groups.

Heatmap: to understand the correlation between different features, such as sentiment vs. user activity.

Histograms: frequency distribution of different user activities, like posts, comments, and likes, among many others.

Box Plot: to show the distribution and variability in user activity metrics across various user groups, such as age or gender.

Descriptive statistics mean, median, standard deviation, and quartiles were calculated in view of central tendency, spread, and distribution.

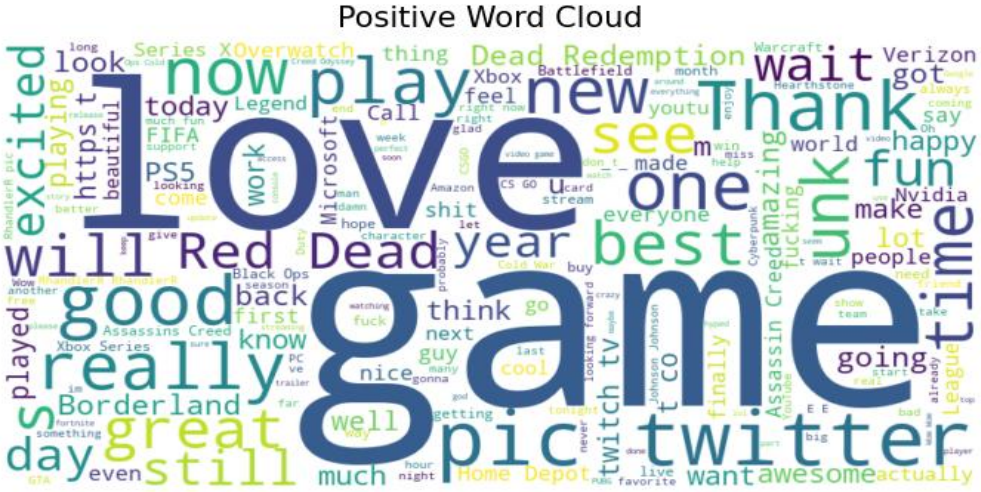

Figure 4.1: Word Cloud of Positive Sentiment

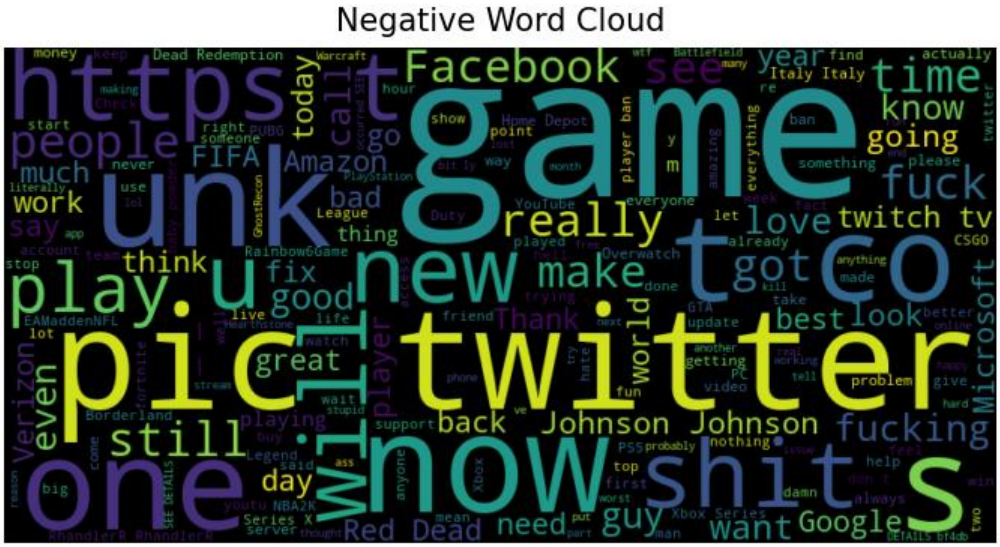

Figure 4.2: Word Cloud of Negative Sentiment

4.3 Preliminary Findings by EDA

Following is the list of initial insights extracted from EDA:

Identify any trend or patterns in user sentiment over time or among various user groups.

Recognize patterns in user activities or engagement metrics of users in a given timeframe.

Correlations: Analyze the relations between different features, for example, sentiment and user activity, or user characteristics and behavior.

Anomalies and Outliers: Identify the outliers or anomalies in the data that might indicate extreme events or data quality issues.

4.4 Feature Engineering

Feature engineering is a process of creating additional variables from the available data to improve the predictive power of the models. Major steps involved in feature engineering include:

Linguistic Features: These include features extracted using pre-trained models, such as sentiment scores and lexicon-based approaches; n-gram features to show the sequence of words; and POS tags, which reveal grammatical structure. The user behavior features would include information about user activity, represented by the number of posts, comments, and likes, and engagement, given by the average number of interactions per post. Demographic User Features: It would include age, gender, and location to understand their influence on user behavior.

Interaction Terms: Interaction between different features to get combined effects. For example, the interaction of sentiment and user activities.

Temporal Features: These will include features such as day of the week, time of day, and season of the year, which will allow capturing temporal patterns in user behavior.

4.5 Expected Outcome

A deep understanding of the patterns and trends that take place in user sentiment and behavior on social networks.

Linguistic features, user activity, and meta-information about the users are some of the key factors that drive user behavior. Designing various predictive models representing user sentiment and behavior that are accurate and reliable. Provide an insight into how data from social networks can be utilized in order to understand and predict user behavior, with many potential applications in marketing, public health, and social science research. 4.6 Future Work Future work could take the following directions:

Research on the impact caused by different social network platforms to the behaviors of users.

Investigating the impact of social influence and network structure in influencing user behavior.

Building a personalized recommendation system by the prediction of user behaviors.

Ethical discussions regarding the use of social network data for prediction purposes.

4.6 Summary

First, this chapter introduced the main source of data: a dataset that included social network user data. Then, in EDA, it created visualizations and descriptive statistics to explore the data and gain a first insight into the data. Feature engineering steps are provided in detail, with variable generation done to help in enhancing the predictive accuracy of the models. The expected outcome and future works are also described in this chapter.