INSY 669 - Text Analytics

Group Project

MMA3

Peirou (Emma) Zhang || 260983073

Leying (Dorothy) Zou || 260950477

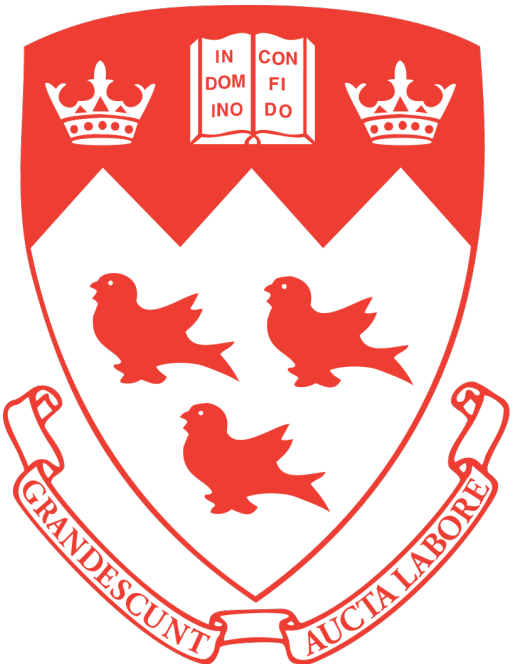
Khaled Al-Masaid || 260623070

Larbi Farihi || 260676506

February 22nd, 2021

Master of Management in Analytics

McGill University



Presented to:

Prof. Changseung Yoo

Summary

List of Figures

A Introduction

Consumers that take the initiative to post reviews play an essential role in the product development process. With the vast amount of user reviews, big companies can directly benefit from customer feedback by gaining valuable insight of consumers' satisfaction or dissatisfaction. Text analytic methods can be used to analyze the reviews and extract information that allows organizations to gain a better understand of the voice of customers. Analyzing the reviews will enable businesses to respond in a more strategic and effective manner to satisfy customer requirements.

B Data description

The data-sets contain customer reviews that were extracted from Twitter by using hashtags specific to a video game. Initially we extracted around 67,900 tweets for five different games including: Cyberpunk 2077, FIFA 21, NBA2K 21, Assassin's Creed Valhalla, Call of Duty-Black Ops Cold War. After pre-processing the data and removing duplicates, we were left with 39,600 tweets. Each review was tokenized by splitting into individual words, then each word is lemmatized and stemmed. Stop words along with punctuations were removed and a function to fix abbreviations was also created. In order to gain insights from the user reviews; Sentiment analysis, K-means clustering and topic modeling with LDA were applied on the extracted reviews.

C Methodology and Analysis

C.1 K-mean clustering

The K-means algorithm was used to create clusters for reviews based on similarity. It is an unsupervised learning algorithm that determines how to label the reviews. A 'For' loop was ran to determine the optimal number of clusters to use between two, three or five clusters. Figure 1 below displays the results of the K-means model on the Valhalla game reviews; in this case the optimal number of clusters was **three**. The Term frequency - Inverse Document Frequency (TF-IDF) matrix was created and fitted into the K-means model; in the figure below each cluster can represent review category (Positive, Negative and Neutral). The results of K-means model can create useful visuals that gives an overview of customers' reaction to a product, it makes it easier to identify users' perception of a video game.

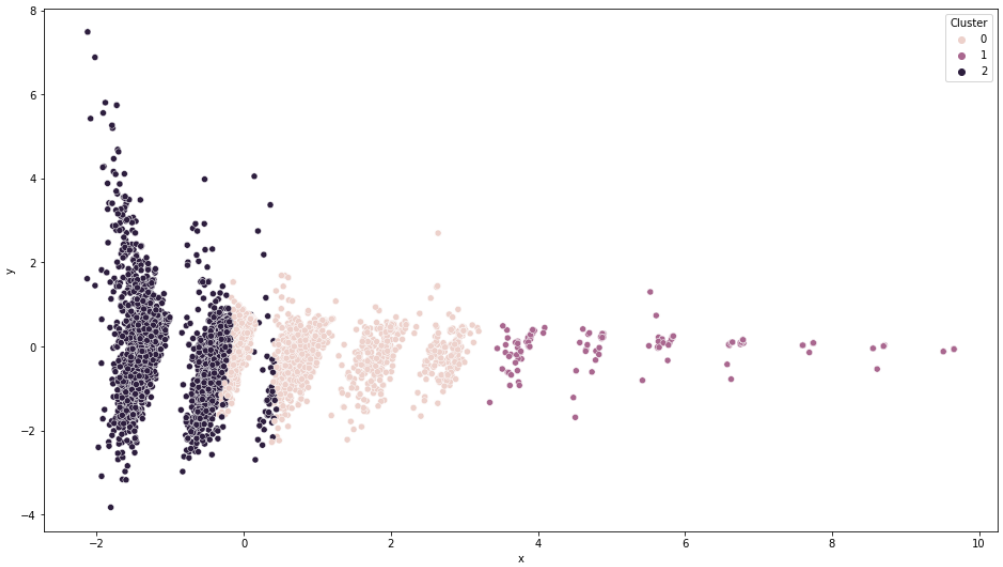


Figure 1: K-mean results for Valhalla game reviews

C.2 Sentiment analysis

The sentiment analysis is processed by VADER, which is a library from Python and NLTK package. It is the advanced method of lexicons, which assigns weights on words with different polarity and calculate the sentiment of a whole sentence. VADER is an unsupervised sentiment analysis model; hence the accuracy of predicting the sentiment polarity is not included in this project. The procedure of doing sentiment analysis can be summarized as three steps. The first step is using VADER model to generate the compound score for each sentiment or review. Next, the scores are classified into three categories: positive, negative and neutral. The threshold of categories can be described by this formula: negative $<$ -0.05 $\leq$ neutral $\leq$ 0.05 $<$ positive. In other words, negative reviews have the scores lower than -0.05, and positive reviews mean they have the scores which are higher than 0.05, while the scores between -0.05 and 0.05 inclusive are neutral. Furthermore, the last step is calculating the sentiment polarity distribution for each game, so it is helpful to gauge public opinions and understand the evaluation of the games without manual reviewing and labeling process. Based on a VADER model and further classification, the result of sentiment analysis are shown in figure 2. As figure 2 demonstrated, Call of Duty: Black Ops Cold War has the highest percentage value of positive reviews, which means it is the most praised between the five games. Also, CyberPunk 2077 and FIFA 21 have the top two and three rankings and the similar results of positive review contributions, while Assassin's Creed Valhalla is the least popular game when comparing the contributions of positive and negative reviews among five games. Moreover, NBA 2k21 have the most neutral reviews contribution, that means 51.08% of Twitter users who commented this game are holding an innocuous attitude. Thus, the ranking of these five games regarding to the sentiment analysis depend on the calculation of overall score. The method is adding 1 score for each percentage of positive reviews

while dropping 1 score for each percentage of negative reviews, and all neutral reviews do not increase or decrease the overall scores. For example, for the game FIFA 21, the overall score equals to 43.07 minus 14.14 which is **28.93**. Table 1 is sorted by overall score in descending order, so the top one is the highest ranked game among five selected games. Consequently, the ranking of games from the top are Call of Duty: Black Ops Cold War, FIFA 21, CyberPunk 2077, NBA 2k21, and Assassin's Creed Valhalla.

Names	Positive	Neutral	Negative	Overall Score
Call of Duty: Black Ops Cold War	46.81%	35.71%	17.48%	29.33
FIFA 21	43.07%	42.79%	14.14%	28.93
CyberPunk 2077	43.14%	35.75%	21.11%	22.03
NBA 2k21	35.31%	51.08%	13.61%	21.7
Assassin's Creed Valhalla	15.72%	10.98%	73.3%	-57.58

Figure 2: Sentiment analysis results

C.3 Topic modeling

Understanding the topic trends and popularity generated by users can motivate the game company's research and development in the market. The topic modeling supported by the Latent Dirichlet Allocation (LDA) assumption can be implied to examine the probabilities of each word appearing in each topic, and different topic percentages within each document. Basically, the algorithm creates dictionaries containing the number of times a word appears, trains LDA model using the bag of words or TF-IDF score and finally visualizing the topics for each document and the corresponding probabilities. The data preprocessing is quite essential when creating the LDA model because it is sensitive to the property of a certain word. Therefore, besides the transformation of lower case and the removal of stopwords and punctuation, the detection of word count frequency is also executed. The top ten words that appears frequently without analytics value are removed from the text for each game. After the modification for original data, the author needs to decide the number of topics K to discover, in this case K=10. Then, the gensim and nltk libraries installed in python are applied to create dictionaries for word counts, run the LDA model and explore the weight of each word in each document, visualize the dominant topic in a nicely formatted output as shown in figure 3 below. The figure demonstrates the ranking of topics in selected documents, among which the higher the score, the more dominant the topic is. The column "Topic_Perc_Contrib"

represents the weight of the topic within the sentence. Therefore, by visualizing the results for the top five games, we can conclude the most popular topics, from which the company is able to gain insights in terms of cooperation, promotion and improvement. For NBA 2k21and Fifa 21, people discuss a lot about the share of highlights recordings on PS5 and X-box. Since they are games designed for sports, users also mention the sports community, league and future stars in their posts. Cyberpunk was a hit when being released in 2020 with people prefer to discuss and use hashtags such as ray tracing –smallstreams, streamer wall and promoting AMGAMERs. The sale promoted in the store and the character duties in the game are the other two hot topics. For Assassin's creed Vahalla there is less information that can be retrieved which may result from either the content itself or the overwhelming appearance of word "Assassin". Instead of a particular topic, users retweet and mention the account called "phollarkemie1" frequently. Furthermore, the distribution of discussions fall into the game-play experience. Lastly, the reviews for call of duty shares a similar content with that of Cyberpunk, while other topics such as the teams in league and seasons update/patch are also quite popular.

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	5.0	0.7000	get, grind, live, 2kcommunity, twitch, gen, ne...	[2kcommunity, basic, 2kcrewfinder]
1	1	5.0	0.7000	get, grind, live, 2kcommunity, twitch, gen, ne...	[grind, check, e3]
2	2	3.0	0.8714	ps5, nba2k21, live, drop, best, new, like, nba...	[pink, shawn, reward, card, marion, evolution,...
3	3	9.0	0.8200	nba, go, ', nba2k21, game, 2k21, ..., check, n...	NaN
4	4	8.0	0.5983	nba, 2k, live, get, come, nba2k, ..., 2k21, wa...	NaN
5	5	3.0	0.8714	ps5, nba2k21, live, drop, best, new, like, nba...	[nba, video, nbatwitter, gamingcreatorsn, yout...
6	6	0.0	0.1000	build, let, nba, xboxshare, 2k21, next, gen, g...	[follower, twitch, yall, follow, know, may]
7	7	4.0	0.5351	xboxshare, game, get, nba2k21, lol, good, firs...	[nba2k21lockercodes, sell, xbox, nba2k21myteam...
8	8	9.0	0.7750	nba, go, ', nba2k21, game, 2k21, ..., check, n...	[house, lmaooooo]
9	9	1.0	0.5499	get, ..., ', nba2k21, myteam, finally, nba2k, ...	[teammed, ready, court, get, full, double, xbo...

Figure 3: Topic modeling results

D Conclusion

Product reviews are essential to the evolution of a product. Companies can use text analytics methods to gain a better understanding of the voice of the customer. Analyzing reviews can make it easier to identify issues with a video game and help users with their decision making by having a better awareness of the general customer satisfaction level. Based on the results of the various analysis conducted in this project, the following recommendations and insights can be made:

D.1 Recommendations

- Next generation games must be an improvement, text analytics can help developers and game companies determine areas of interests and focus to satisfy customer requirements.

- Currently, the rating system of games treated all reviews equally. To improve the reliability of game rating, we can enhance our rating system by adding weighting algorithm that assigns higher weights to verified twitter accounts and data sources (i.e. professional game testers)
- Our algorithms can be re-used to analyze other games. Thus, the further improvement of this project is to create an algorithm to collect textual information on different social media platforms and generate the game rankings based on each platform. Furthermore, we can build another dashboard to compare the game rankings on different

D.2 Insights

- Text analytics on user reviews can make it easier to identify any existing issues with a video game.
- Help users with their decision making by having a better awareness of the general customer satisfaction level.
- Generate the integrated evaluation of games by sentiment analysis helped marketing teams of the game companies to adjust their strategies
- Measure the public feedback of games and the commonly used words that strongly correlated to the games from big data