1. Introduction

1.1. Problem Background

Steam belongs to a company called Valve Corporation also known as Valve Software. It is a video game publisher, developer and digital distribution organization in America. Nowadays, it is convenient to see what other Steam users think about the games before the users buy it. With Steam Reviews, the users can browse for the reviews that others have found useful or even write your own reviews on Steam. Due to the massive dataset which cannot be analyzed manually, the company wishes to use Natural Language Processing (NLP) to analyze the sentiment of the reviews. Hence, an analysis of customer reviews is developed.

1.2. Objectives/Aims

The objectives / aims of this collection is to provide an implementation of NLP Machine Learning methods for a set of problems of realistics value so hopefully appropriate for business applications. Our team has decided to analyze the game's product at the Steam Store, mostly because we know the domain pretty well. Moreover, Steam Store has a good metadata ecosystem for NLP tasks. For instance, reviews have 'helpful' score, 'funny' score, number of hours reviewer played the game.

1.3. Motivation

Sentiment analysis is a branch of text classification, which analyzes subjective words or phrases with the positive, negative and neutral emotional.In addition, the words and phrases spread rapidly to social media as well as negative comments gain as quickly as possible. If the organization does not quickly and respectfully deal with dissatisfied customers, they may share their disappointment to their friends or even public such as post to their social media. Moreover, the organization can use sentiment analysis to analyze what the customer likes and does not like about their services, brands and products which is quite essential to their business. Apart from tracking own online records, the organization could also track their competitors' comments by using sentiment analysis to see how the business can be stacked up. Positive sentiments are able to help you determine the key factor of the success of competitors. While negative sentiments are able to open up opportunities for the organization.

1.4. Timeline/Milestone

Week	Progress
1	Browsing dataset
2	Background studies
3	Selection of algorithm (KNN) and feature engineering (Bag-of-word)
4 & 5	Study NLP for sentiment analysis
6	Research the suitable development tools
7	Starting the coding part and trying to remove noise from dataset ● Convert to lowercase ● Tokenization ● Remove punctuation ● Remove stopwords
8	Lemmatization and remove non-english word from dataset
9	Feature Generation using Bag-of-word
10	Using KNN to train and test the model
11	Trying to improve the accuracy of testing set
12	Visualization using Wordcloud
13	Documentation

2. Research Background

2.1. Background of the applications

Sentiment Analysis is the process of determining whether the sentence or word is neutral, positive, negative or even to multiple sentiment. Moreover, it is also a conceptual text mining process that extracts and identifies subjective information from source material and helps businesses understand the social sentiment of their service, product or brand while monitoring online conversations. Yet, the analysis of social media streams is usually limited to the basic sentiment analysis. Furthermore, a sentiment analysis system combines Natural Language Processing also known as NLP and machine learning techniques to assign the weight for sentiment score to the themes, topic and categories within a sentence.

2.2. Analysis of selected tool with any other relevant tools

Tools comparison	Remark	Scikit-learn	Matplotlib	Google Colab
Type of license and open source license	State all types of license	New BSD License	Matplot License	BSD License
Year founded	When is this tool being introduced?	June 2007	2003	2006
Founding company	Owner	Google Summer of Code projected by David Cournapeau	John D.Hunter	Travis Oliphant
License Pricing	Compare the prices if the license is used for development and business/commercialization	Free	Free	Free
Supported features	What features that it offers?	Clustering, Regression and Classification	Histogram, Bar Chart, Pie Chart, Table, Scatter Plot, Boxplot.	Multi-dimensional array, mathematical operations
Common applications	In what areas this tool is usually used?	Image recognition, stock prices and customer segmentation	Plotting, subplots and images	Mathematical analysis, signal processing, and symbolic computation
Customer support	How the customer support is given, e.g. proprietary, online community, etc.	User questions, Bug tracker, documentation resources	User questions, Bug tracker, documentation resources	User questions, Bug tracker, documentation resources
Limitations	The drawbacks of the software	Does not use any hardware acceleration as making it slow when training the model	R / Matlab is still doing many cutting-edge advanced academic research	Always face array.reshape (1, -1) problem

2.3. Justify why the selected tool is suitable

First of all, sklearn is a free machine learning tool for Python. It contains multiple algorithms such as k-neighbors, random forests, decision trees, support vector machines etc. Moreover, sklearn also supports Python scientific libraries which are Scipy and Numpy. LabelEncoder feature is being used for converting categorical data to integer. Secondly, I have chosen bag-of-word as my module, thus CountVectorizer is being used to perform feature generation for the development. Followed by the train test split and K-nearest Neighbors Classification for model building. Furthermore, Matplotlib library does help us to visualize the bar charts, line charts such as comparing the accuracy of the training set and training based on k value. Lastly, Numpy remains critical for this development such as performing mathematical calculation when doing machine learning.

3. Methodology

3.1. Description of dataset

The dataset is retrieved from Kaggle website which may contain some missing values. It basically contains 434,892 rows and 8 columns in total. Considering some issues occur we have removed the dataset to 31,718 rows. Other than that, the data types consist of 1 boolean type, 3 integer type and 4 object type.

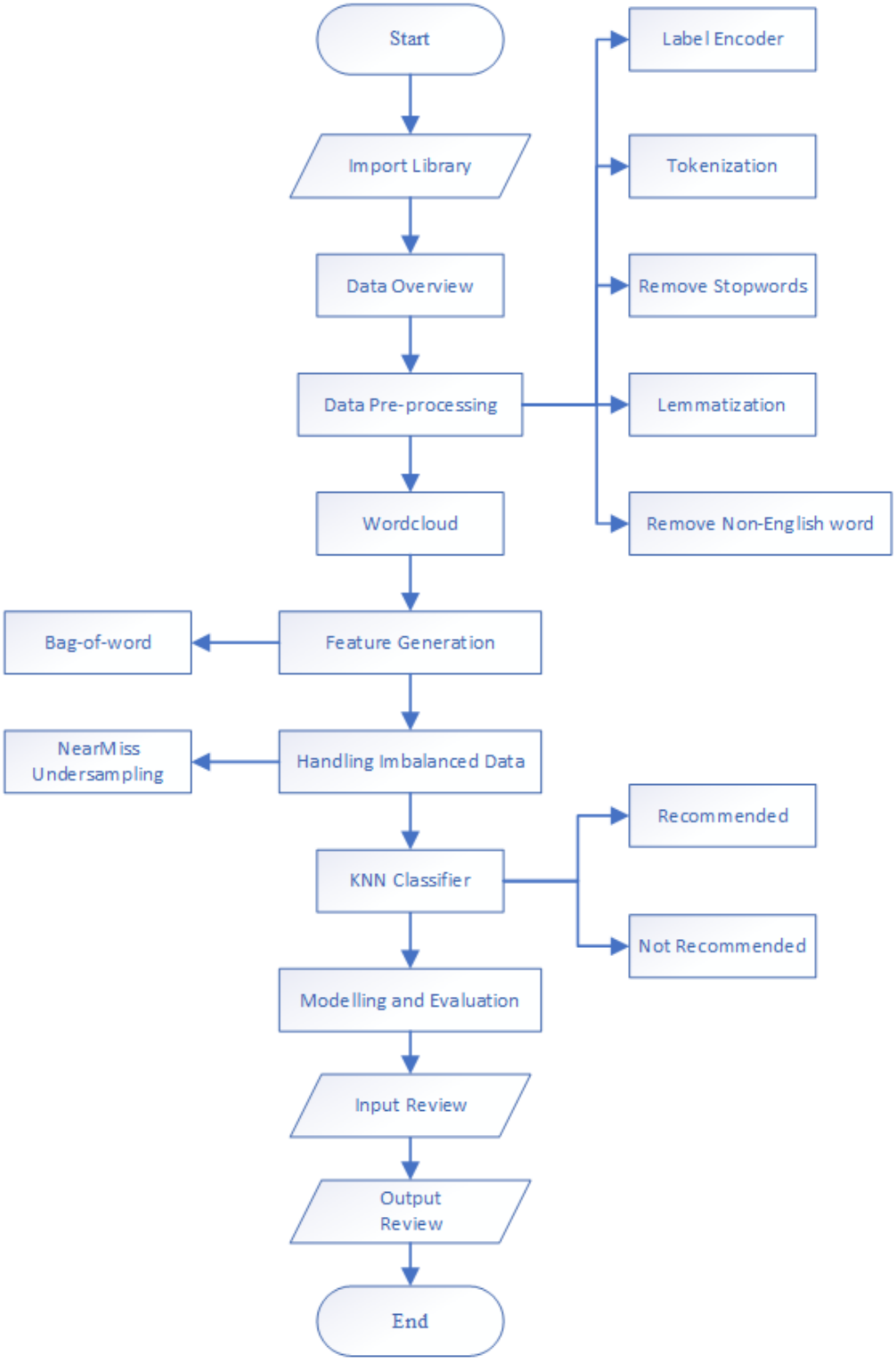
Variables Breakdown:

1. **date_posted**: The date a review is posted
2. **funny**: How many other player think the review is funny
3. **helpful**: How many other player think the review is helpful
4. **hour_played**: How many hour a reviewer play the game before make a review
5. i**s_early_access_review**: Is it a early access based on the review
6. **recommendation**: Whether reviewer recommend the game or not
7. **review**: The text of user review
8. **title**: The game's title that's being reviewed

3.2. Applications of the algorithm(s)

KNN Classification has been selected to perform this supervised machine learning. KNN is used to classify by looking for the K nearest matches in the training set as well as using the closest point to perform prediction. I have implemented class KNeighborsClassifier() with the standard functions “fit” for training set and “predict” for testing set. In addition, we know that k value denotes how many of the nearest neighbors will be used to make predictions. However, we use k = 1 will actually come in an overfitting case, therefore the comparison graph between mean square error (mse) and k value has been constructed for choosing the reasonable k value.

3.3. System flowchart/activity diagram



3.4. Proposed test plan/hypothesis

Considering to develop NLP for this assignment, thus it is necessary to carry out the text cleaning (data preprocessing) process. Convert Lower-cases, remove punctuations and remove special characters is the first step in the text cleaning part. First of all, this function is able to convert all the alphabet to lowercase letters. Secondly, this function is also able to remove all the punctuations and special characters from the dataset by importing Regular Expression (RegEx) library. Followed by the second step which is Tokenization. It is a process of breaking sentences and phrases into a single word which separate by comma. Token represents a single entity that builds for a paragraph and sentence. Dataset will always contain noise which means the word represents meaningless such as am, is are, there, the, an and etc. Therefore, we need to create a list for english stopwords and apply to filter out stopwords from the dataset. Furthermore, Lemmatization considers another type of noise in the text. It is able to convert all the words to become their root word such as "told" will be converted to "tell". The reason for choosing Lemmatization rather than Stemming is because the transformation of Lemmatization is referring to a dictionary. Apart from that, since I am doing this sentiment analysis with english words, the system will not recognize non-english words, so that we can treat it as a noise and try to remove them from the dataset. The KNN algorithm only recognizes numeric, so we need to convert the label (recommendations) from object variable to integer variable where 0 and 1 represent not recommended and recommended respectively. In order to know which word appears the most frequently from the dataset, word clouds are formed so that it is easy to visualize the most frequent words between positive and negative review. Lastly, bag-of-words is selected to perform feature engineering at the following step. Bag-of-Word is a method that is widely used in NLP development. It is able to create a vocabulary for all occurring words from a dataset and count the unique words that appear several times.

4. Result

4.1. Results

Convert to Lower-case, Remove Punctuation and Special Characters

	recommendation	review	new_review
0	Recommended	&gt Played as German Reich&gt Declare war on B...	gt played as german reichgt declare war on bel...
1	Recommended	yes.	yes
2	Recommended	Very good game although a bit overpriced in my...	very good game although a bit overpriced in my...
3	Recommended	Out of all the reviews I wrote This one is pro...	out of all the reviews i wrote this one is pro...
4	Recommended	Disclaimer I survivor main. I play games for f...	disclaimer i survivor main i play games for fu...
5	Recommended	ENGLISH After playing for more than two years ...	english after playing for more than two years ...
6	Recommended	Out of all the reviews I wrote This one is pro...	out of all the reviews i wrote this one is pro...
7	Recommended	I have never been told to kill myself more tha...	i have never been told to kill myself more tha...
8	Recommended	Any longtime Dead by Daylight player knows tha...	any longtime dead by daylight player knows tha...
9	Recommended	if you think cs go is toxic try this game	if you think cs go is toxic try this game

Screenshot 4.1.1 Convert to Lower-case, Remove Punctuation and Special Characters

Tokenization

	recommendation	review	new_review
0	Recommended	&gt Played as German Reich&gt Declare war on B...	[gt, played, as, german, reichgt, declare, war...
1	Recommended	yes.	[yes]
2	Recommended	Very good game although a bit overpriced in my...	[very, good, game, although, a, bit, overprice...
3	Recommended	Out of all the reviews I wrote This one is pro...	[out, of, all, the, reviews, i, wrote, this, o...
4	Recommended	Disclaimer I survivor main. I play games for f...	[disclaimer, i, survivor, main, i, play, games...
5	Recommended	ENGLISH After playing for more than two years ...	[english, after, playing, for, more, than, two...
6	Recommended	Out of all the reviews I wrote This one is pro...	[out, of, all, the, reviews, i, wrote, this, o...
7	Recommended	I have never been told to kill myself more tha...	[i, have, never, been, told, to, kill, myself,...
8	Recommended	Any longtime Dead by Daylight player knows tha...	[any, longtime, dead, by, daylight, player, kn...
9	Recommended	if you think cs go is toxic try this game	[if, you, think, cs, go, is, toxic, try, this,...

Screenshot 4.1.2 Tokenization

Remove Stopwords

	recommendation	review	new_review
0	Recommended	&gt Played as German Reich&gt Declare war on B...	[gt, played, german, reichgt, declare, war, be...
1	Recommended	yes.	[yes]
2	Recommended	Very good game although a bit overpriced in my...	[good, game, although, bit, overpriced, opinio...
3	Recommended	Out of all the reviews I wrote This one is pro...	[reviews, wrote, one, probably, serious, one, ...
4	Recommended	Disclaimer I survivor main. I play games for f...	[disclaimer, survivor, main, play, games, fun,...
5	Recommended	ENGLISH After playing for more than two years ...	[english, playing, two, years, given, task, re...
6	Recommended	Out of all the reviews I wrote This one is pro...	[reviews, wrote, one, probably, serious, one, ...
7	Recommended	I have never been told to kill myself more tha...	[never, told, kill, playing, game]
8	Recommended	Any longtime Dead by Daylight player knows tha...	[longtime, dead, daylight, player, knows, isnt...
9	Recommended	if you think cs go is toxic try this game	[think, cs, go, toxic, try, game]

Screenshot 4.1.3 Remove Stopwords

Lemmatization

	recommendation	review	new_review
0	Recommended	&gt Played as German Reich&gt Declare war on B...	[gt, play, german, reichgt, declare, war, belg...
1	Recommended	yes.	[yes]
2	Recommended	Very good game although a bit overpriced in my...	[good, game, although, bite, overprice, opinio...
3	Recommended	Out of all the reviews I wrote This one is pro...	[review, write, one, probably, serious, one, w...
4	Recommended	Disclaimer I survivor main. I play games for f...	[disclaimer, survivor, main, play, game, fun, ...
5	Recommended	ENGLISH After playing for more than two years ...	[english, play, two, years, give, task, review...
6	Recommended	Out of all the reviews I wrote This one is pro...	[review, write, one, probably, serious, one, w...
7	Recommended	I have never been told to kill myself more tha...	[never, tell, kill, play, game]
8	Recommended	Any longtime Dead by Daylight player knows tha...	[longtime, dead, daylight, player, know, isnt,...
9	Recommended	if you think cs go is toxic try this game	[think, cs, go, toxic, try, game]

Screenshot 4.1.4 Verbs Lemmatization

31704	Recommended	Edit It appears they have fixed the disconnect...	[edit, appear, fix, disconnection, issue, alth...
31705	Recommended	we love the game but the server s#cks big time...	[love, game, server, scks, big, time, price, e...
31706	Recommended	The game in itself is already great but after ...	[game, already, great, connection, fix, become...
31707	Recommended	This recommendation is only positive because o...	[recommendation, positive, effort, modding, co...
31708	Recommended	yeet	[yeet]
31709	Recommended	"s &amp s is good for beginners" Someone who n...	[amp, good, beginner, someone, never, use, amp]
31710	Recommended	To all the players who complained and whined a...	[player, complain, whine, game, lock, monster,...
31711	Recommended	I dont think I have words to describe what cou...	[dont, think, word, describe, could, one, best...
31712	Recommended	I just really really love the game.Thats it.	[really, really, love, gamethats]
31713	Not Recommended	Summary If you have played a Monster Hunter ga...	[summary, play, monster, hunter, game, youll, ...
31714	Recommended	Great game except there's a connection issues ...	[great, game, except, there, connection, issue...
31715	Recommended	Great game really enjoying it so far!	[great, game, really, enjoy, far]
31716	Recommended	The core game experience is a masterpiece.Each...	[core, game, experience, masterpieceeach, map,...
31717	Recommended	Overall really good game and got me hooked to ...	[overall, really, good, game, get, hook, serie...

Screenshot 4.1.5 Nouns Lemmatization

Remove Non-English Words

	recommendation	review	new_review
0	Recommended	&gt Played as German Reich&gt Declare war on B...	[gt, play, declare, war, cant, break, go, capi...
1	Recommended	yes.	[yes]
2	Recommended	Very good game although a bit overpriced in my...	[good, game, although, bite, overprice, opinio...
3	Recommended	Out of all the reviews I wrote This one is pro...	[review, write, one, probably, serious, one, w...
4	Recommended	Disclaimer I survivor main. I play games for f...	[disclaimer, survivor, main, play, game, fun, ...
5	Recommended	ENGLISH After playing for more than two years ...	[play, two, year, give, task, review, game, re...
6	Recommended	Out of all the reviews I wrote This one is pro...	[review, write, one, probably, serious, one, w...
7	Recommended	I have never been told to kill myself more tha...	[never, tell, kill, play, game]
8	Recommended	Any longtime Dead by Daylight player knows tha...	[longtime, dead, daylight, player, know, horro...
9	Recommended	if you think cs go is toxic try this game	[think, c, go, toxic, try, game]

Screenshot 4.1.6 Remove Non-English Words

Label Encoding

	recommendation	new_recommendation
0	Recommended	1
1	Recommended	1
2	Recommended	1
3	Recommended	1
4	Recommended	1

Screenshot 4.1.7 Label Encoding

Wordcloud Visualization

Screenshot 4.1.8 Positive Wordcloud

Screenshot 4.1.9 Negative Wordcloud

Bag-of-Word

	ab	abandon	abandonment	abduct	aberration	abhor	abhorrent	abide	ability	abject	able	abnormal	abnormally	aboard	abominable	abomination	abort	abor
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
31657	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
31658	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
31659	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
31660	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
31661	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Screenshot 4.1.10 Bag-of-Word

ield	yikes	yo	yoga	yoke	yolk	yore	young	younger	youngest	yr	yuck	yummy	yup	zap	zappy	zealous	zen	zero	zest	zinger	zip	zit	zombie	zone	zoo	zoom
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Screenshot 4.1.11 Bag-of-Word

4.2. Discussion/Interpretation

Convert to Lower-case, remove punctuation and special characters

Based on the last part, we can conclude that the system is able to perform text cleaning / analysis well. First of all, the screenshot 4.1.1 illustrates the conversion of all the uppercase to lowercase letters Played→ played, ENGLISH → english and etc. Beside that, all the punctuation and special characters have been removed from the data frame. The regex library is used for this part of text cleaning.

Tokenization

Secondly, the screenshot 4.1.2 shows the result after Tokenization. The Tokenization is able to break out a sentence into pieces such as words and separate them by a comma. As a result, the word_tokenize library is used to perform this part of text cleaning.

Remove Stopwords

In addition, the screenshot 4.1.3 illustrates the elimination of the meaningless word which is also known as stopword. This process is to remove every single stopword from the data frame. E.g. as, a, to, is, of, the and etc has been removed. By the way, the stopwords library is used to perform this part of text cleaning.

Lemmatization

Furthermore, Lemmatization is separated by 2 parts which lemmatize the verbs and nouns and it has been shown in screenshot 4.1.4 and screenshot 4.1.5. The first lemmatization is to convert all the verbs to their root word. E.g. played → play, reviews → review, wrote → write, told → tell and etc. Moreover, second Lemmatization for only verbs is always not enough, therefore nouns should be converted so. E.g. problems → problem, guys → guy, issues → issue, beginners → beginner and etc. Accordingly, the WordNetLemmatizer library is used to perform this part of text cleaning.

Remove Non-English Words

Lastly, removing non-english words remains critical in this NLP assignment since we are focusing on english review, hence the other language such as Italian and French would be meaningless. As a result, the enchant library is used to remove all the non-english words that have been removed from the data frame e.g. Reich&gt and etc.

Label Encoding

LabelEncoder library is normally used to convert categorical variables into numeric. In this case, the recommendation has been converted to numeric which is 0 and 1 as shown at screenshot 4.1.7.

Wordcloud Visualization

Word Cloud is one of the ways to visualize the frequently occurring words from the dataset where the bigger the word size, the more words that appear. Screenshot 4.1.8 and screenshot 4.1.9 clearly show the positive and negative word Cloud respectively. WordCloud library is used to perform this part of visualization.

Bag-of-Word

Screenshot 4.1.8 and screenshot 4.1.9 show that all the unique words have been vectorized and stored in a data frame. All the vectors in this process will be used in the KNN algorithm for prediction and classification of sentiment analysis. CountVectorizer library is used to perform this feature engineering.

5. Discussion and Conclusion

5.1. Achievements

In this assignment, I have learned multiple methods to perform text cleaning / analysis. The basics of this operation is using NLTK tools such as Tokenization, Stopwords, Normalization, Lemmatization, Stemming, POS Tagging. Although I did not include every tool in the assignment which was listed in the previous sentence, I did study on every tool and choose the suitable tool. As we know that sentiment analysis is widely used in business areas especially the social media field. As a result, I have also learned one ability to survive in the market although my development is not perfect but I will spare no effort to enhance it in the future. Conclusively, I would like to mention that I have fulfilled the objectives of this assignment.

5.2. Limitations and Future Works

One of the important issues I have to mention is the dataset is quite huge, therefore I have to reduce the dataset until the supportable range. It is because with the huge dataset I will always get the run time crashed in Google Colab especially the feature generation part. While sentiment analysis is a powerful application of NLP, yet the computer programs will always have difficulties in understanding the phrases and words such as sarcasm and jokes because as a human we would have a little trouble to understand so. If we do not take action or ignore this kind of case, at the end the findings might be distorted. For example, “Recommend” may be classified as a positive class of sentiment analysis, however with the sentence “I would not recommend” it should be classified as a negative class. As a result, by doing sentiment analysis

we have to be careful on every single word, as an example before, although “not” is considered as a stopword but it represents an important role in this sentence then the findings and results will be totally different in this case. After finishing this assignment, NLP became the foundation for my future NLP subject as well as I am able to integrate faster than others.

Reference & Source

Steam Reviews Dataset | Kaggle. Available at: <https://www.kaggle.com/luthfim/steam-reviews-dataset> (Accessed: 3 September 2020).

Natural Language Toolkit — NLTK 3.5 documentation. Available at: <https://www.nltk.org/> (Accessed: 3 September 2020).

scikit-learn: machine learning in Python — scikit-learn 0.23.2 documentation. Available at: <https://scikit-learn.org/stable/> (Accessed: 3 September 2020).

Matplotlib: Python plotting — Matplotlib 3.3.1 documentation. Available at: <https://matplotlib.org/> (Accessed: 3 September 2020).

numpy - Google Search. Available at: <https://www.google.com/search?rlz=1C1CHBF_enMY885MY885&sxsrf=ALeKk01xD7OoppijA4basHNPdIzuvjhx2Q%3A1599067527411&ei=h9VPX93aGK2Z4-EP1tGkoAU&q=numpy&oq=numpy&gs_lcp=CgZwc3ktYWIQAzIECCMQJzIECCMQJzIECCMQJzIFCAAQkQIyBAgAEEMyBAgAEEMyBAgAEEMyBAgAEEMyBAgAEEM6BwgAEEcQsANQpklYpklgnktoAHAAeACAAVmIAZEBkgEBMpgBAKABAaoBB2d3cy13aXrAAQE&sclient=psy-ab&ved=0ahUKEwidusXf_srrAhWtzDgGHdYoCVQQ4dUDCA0&uact=5> (Accessed: 3 September 2020).

seaborn: statistical data visualization — seaborn 0.10.1 documentation. Available at: <https://seaborn.pydata.org/> (Accessed: 3 September 2020).

(Tutorial) Text ANALYTICS for Beginners using NLTK - DataCamp. Available at: <https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk> (Accessed: 3 September 2020).

An introduction to Bag of Words and how to code it in Python for NLP. Available at: <https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/> (Accessed: 4 September 2020).

k-nearest neighbor algorithm in Python - GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/> (Accessed: 4 September 2020).

KNN Classification using Scikit-learn - DataCamp. Available at: <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn> (Accessed: 4 September 2020).

Python NLTK sentiment analysis | Kaggle. Available at: <https://www.kaggle.com/ngyptr/python-nltk-sentiment-analysis> (Accessed: 8 September 2020).

k-nearest neighbor algorithm in Python - GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/> (Accessed: 4 September 2020).