TEXT MINING AMAZON REVIEWS

FATHIMA AYOOB

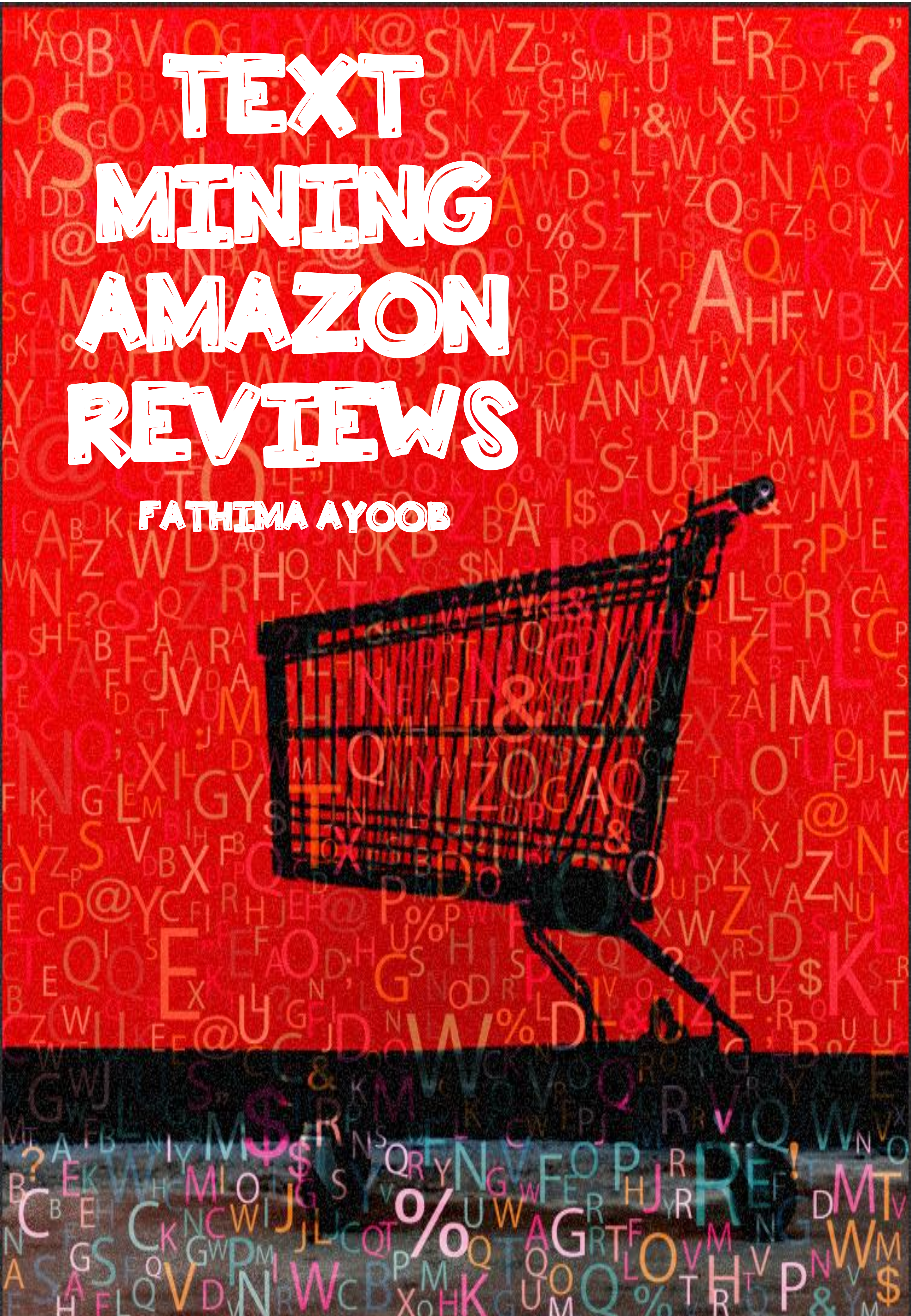


TABLE OF CONTENTS

INTRODUCTION

A review is an evaluation of a publication, service, or company such as a movie, video game, musical composition, book; a piece of hardware like a car, home appliance, or computer; or an event or performance.

Online product reviews are a great source of information for consumers. From the sellers' point of view, online reviews can be used to gauge the consumers' feedback on the products or services they are selling. However, since these online reviews are quite often overwhelming in terms of numbers and information, an intelligent system, capable of finding key insights (topics) from these reviews, will be of great help for both the consumers and the sellers. This system will serve two purposes:

- Enable consumers to quickly extract the key topics covered by the reviews without having to go through all of them
- Help the sellers/retailers get consumer feedback in the form of topics (extracted from the consumer reviews)

It has been found that for nearly 9 in 10 consumers, an online review is as important as a personal recommendation.

Customers are likely to spend 31% more on a business with "excellent" reviews.

72% say that positive reviews make them trust a local business more.

92% of users will use a local business if it has at least a 4-star rating.

72% of consumers will take action only after reading a positive review.

ABOUT THE STUDY

OBJECTIVE:
This project is on analysing the reviews of the amazon product, "All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 16 & 32 GB - Includes Special Offers, Magenta"

The analysis helps us understand whether majority of the consumers are satisfied with the product or not. It also helps us to find the reasons why the product must have been unfavourable to some. Thereby helping the company to make improvements to increase profits.

ABOUT THE DATA:
The data was scraped from the amazon website using R.
The scraped data was then converted to csv format with the following columns:

reviews.doRecommend: whether the customer will recommend the product or not.

reviews.rating: rating given to the product by the customer.

reviews.text: review given by the customer.

METHODOLOGY

TEXT MINING:

Text mining is the process of examining large collections of text and converting the unstructured text data into structured data for further analysis like visualization and model building.

In terms of text mining approaches, there are 2 broad categories:

Semantic Parsing

Bag of words

Here the word sequence, word usage as noun or verb, hierarchical word structure etc. matters

Here all the words are analysed as a single token and order does not matter.

IMPORTANT TERMS AND DEFINITION USED IN TEXT MINING:

CORPUS:

A corpus is defined as “a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject”.

CLEANING DATA:

Once the data is loaded into the work space, it is time to clean this data .The goal is to create independent terms(words) from the data file before we can start counting how frequently they appear.

First convert the entire text to lowercase to avoid considering same words like "write" and "Write" differently.

Then remove : URLs , emojis, non-english words,punctuations,numbers,whitespace and stop words.

STOP WORDS:

The commonly used English words like "a"," is ","the" in the tm package are referred to as stop words. These words have to be eliminated so as to render the results more accurate. It is also possible to create your own custom stop words.

TOKENISATION:

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens.

STEMMING:

Stemming is a process of reducing words to its root form even if the root has no dictionary meaning. For eg: **beautiful** and **beautifully** will be stemmed to **beauti** which has no meaning in English dictionary.

LEMMETIZATION:

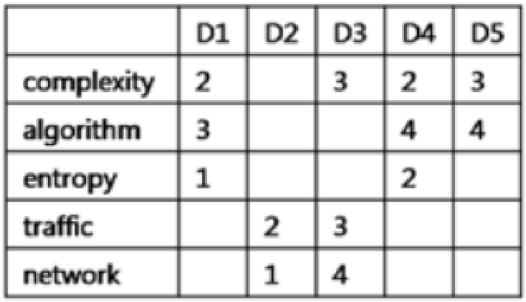
Lemmatisation is a process of reducing words into their lemma or dictionary. It takes into account the meaning of the word in the sentence.

For eg: **beautiful** and **beautifully** are lemmatised to **beautiful** and **beautifully** respectively without changing the meaning of the words.

But, **good**, **better** and **best** are lemmatised to **good** since all the words have similar meaning.

TERM DOCUMENT MATRIX:

After the cleaning process ,we are left with independent terms that exist throughout the document. These are stored in a matrix that shows each of their occurrence. This matrix logs the number of times the term appears in our clean data set thus being called a term matrix.



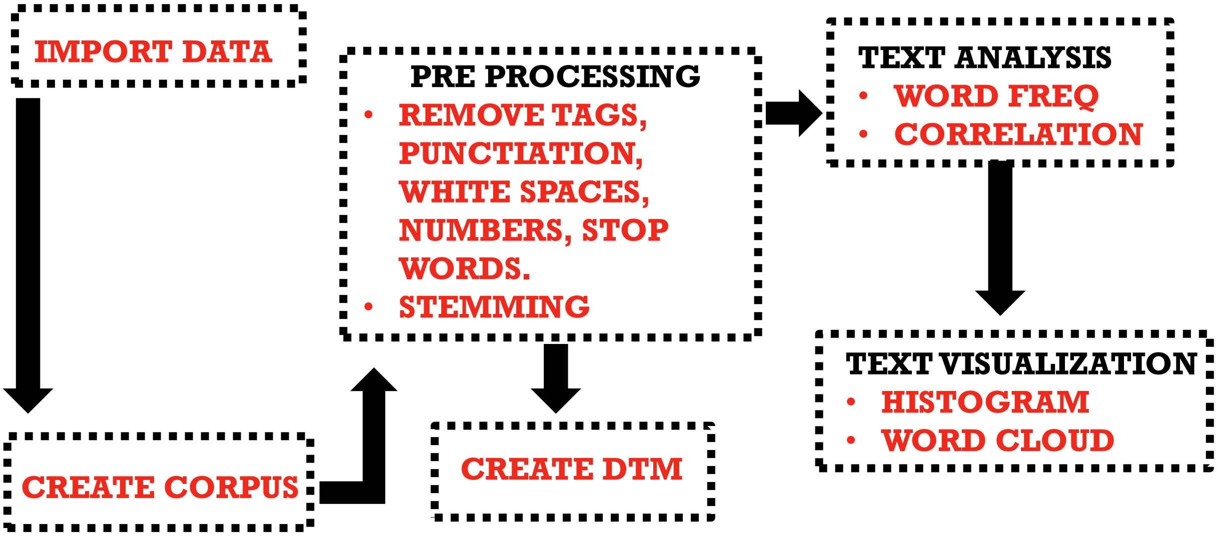
	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

WORD FREQUENCIES:

These are the number of times words appear in data set. Word frequencies will indicate to us from the most frequently used words in the data set to the least used using the compilation of occurrences from the term matrix.

TEXT MINING IN R:



IMPORT DATA

CREATE CORPUS

PRE PROCESSING
- REMOVE TAGS, PUNCTIATION, WHITE SPACES, NUMBERS, STOP WORDS.
- STEMMING

CREATE DTM

TEXT ANALYSIS
- WORD FREQ
- CORRELATION

TEXT VISUALIZATION
- HISTOGRAM
- WORD CLOUD

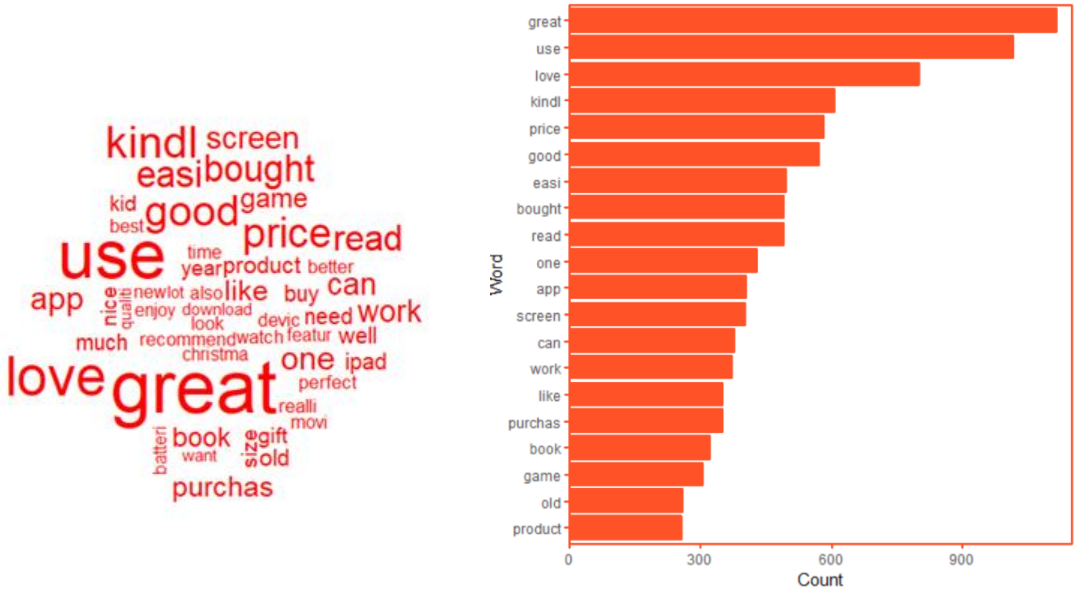
N-GRAMS:

An N-gram means a sequence of N words. For example, "Medium blog" is a 2-gram (a bigram), "A Medium blog post" is a 4-gram, and "Write on Medium" is a 3-gram (trigram).

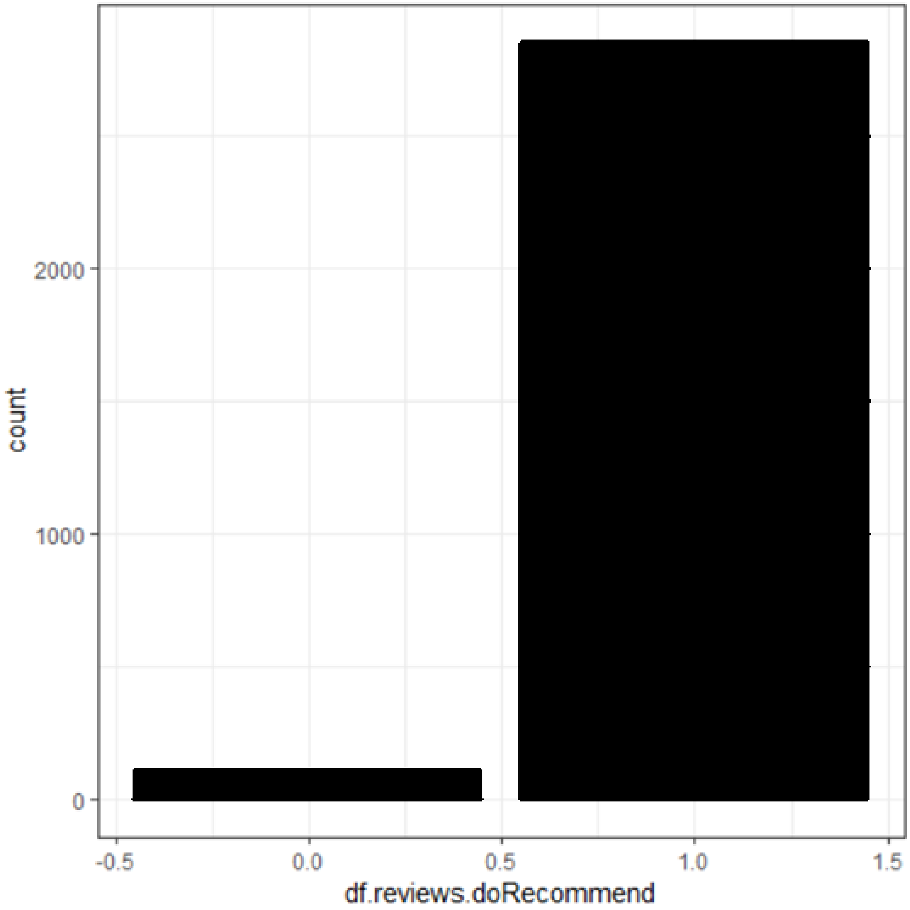
ANALYSIS

MOST FREQUENT WORDS USED

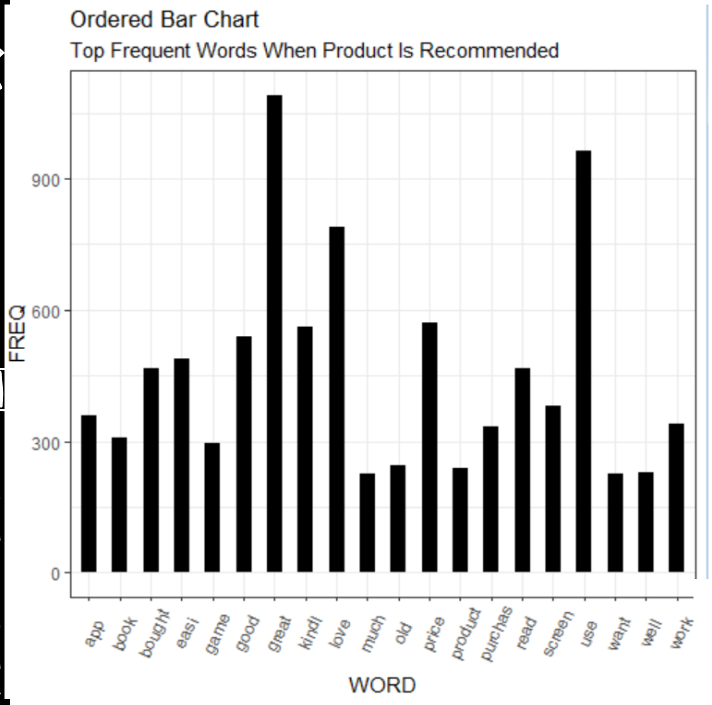
Fig 1

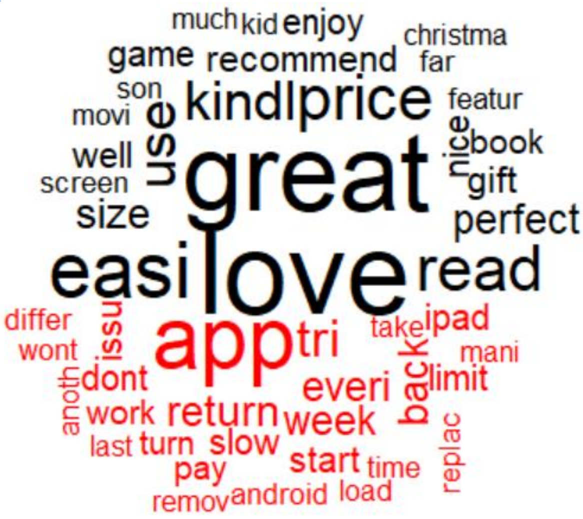


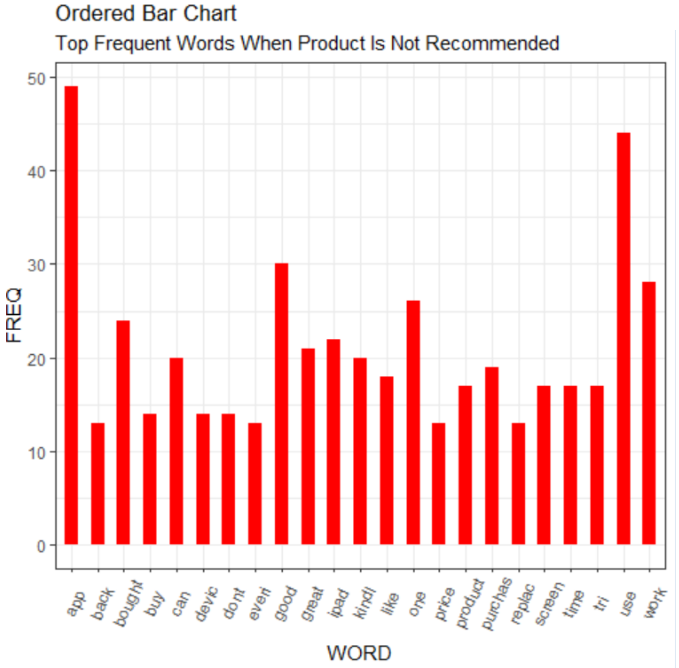
COUNT OF CUSTOMERS WHO DID AND DID NOT RECOMMEND THE PRODUCT



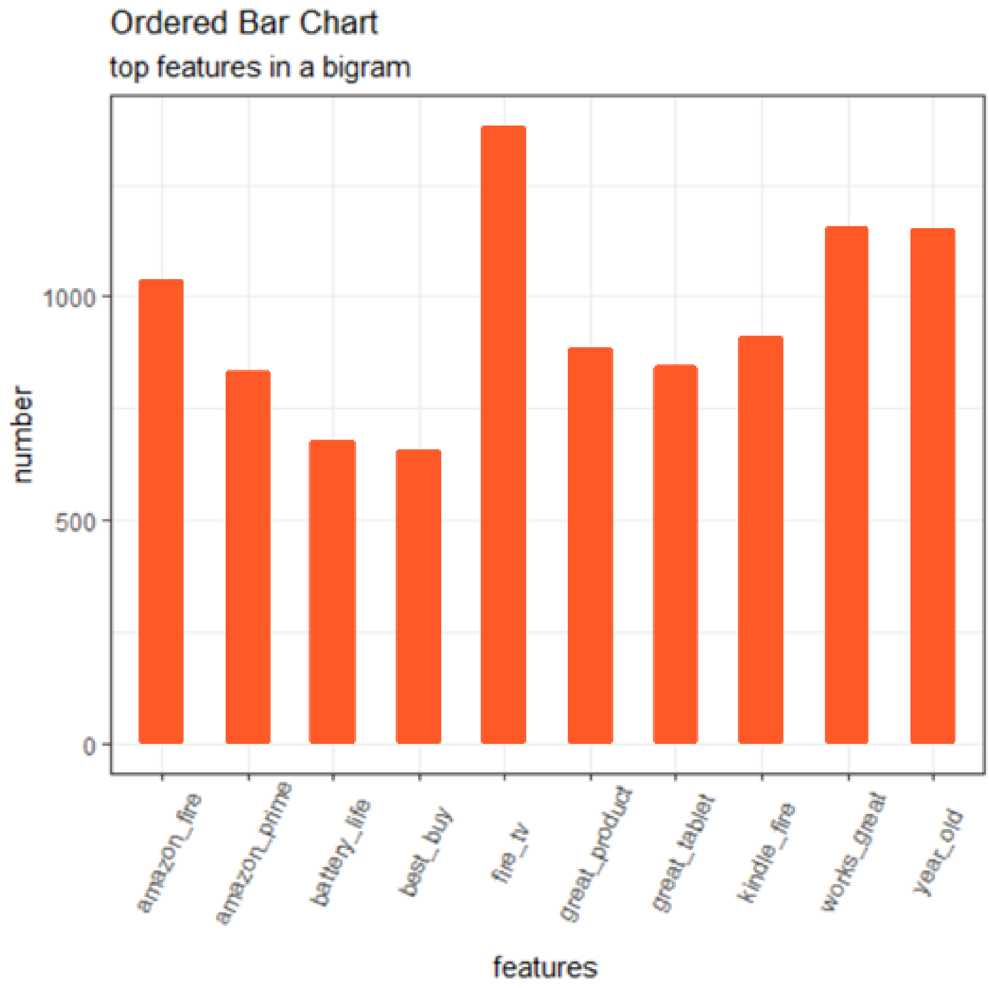
MOST FREQUENT WORDS USED FOR RECOMMENDED AND NOT RECOMMENDED PRODUCTS (Fig 2)

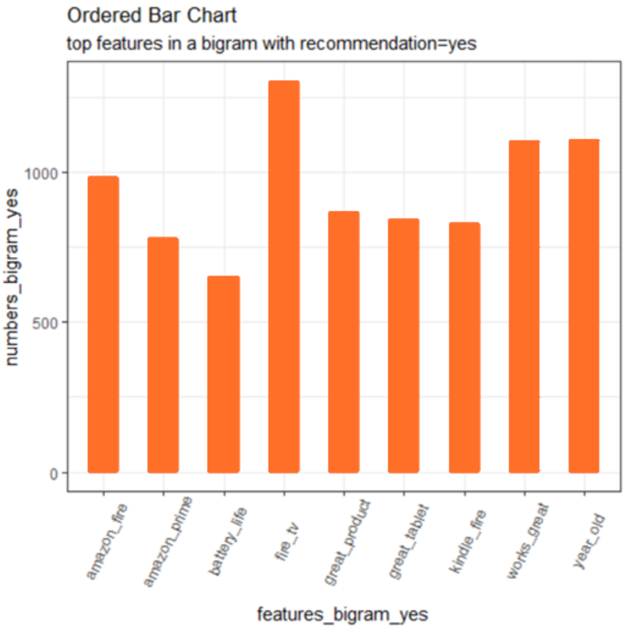


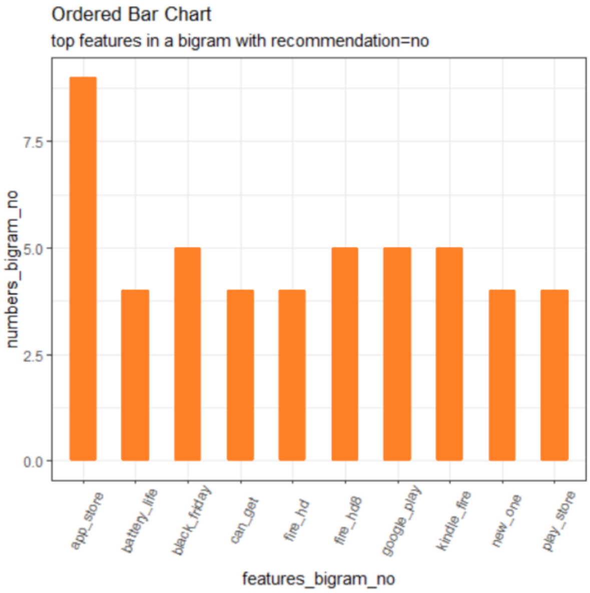




MOST FREQUENT BI-GRAM WORDS USED FOR RECOMMENDED AND NOT RECOMMENDED PRODUCTS



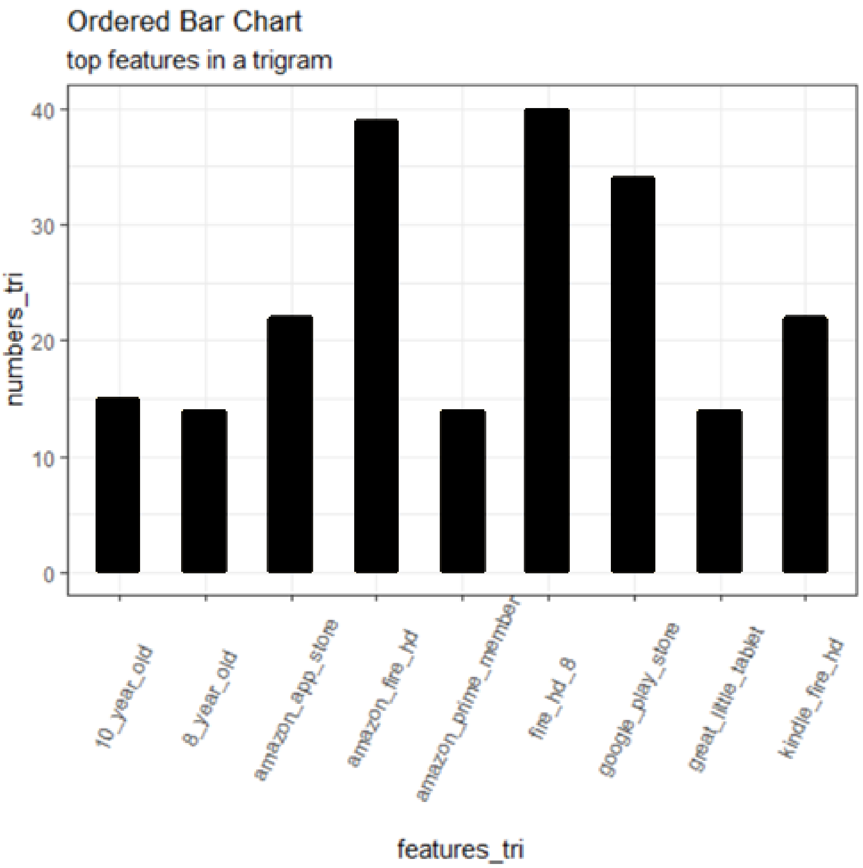


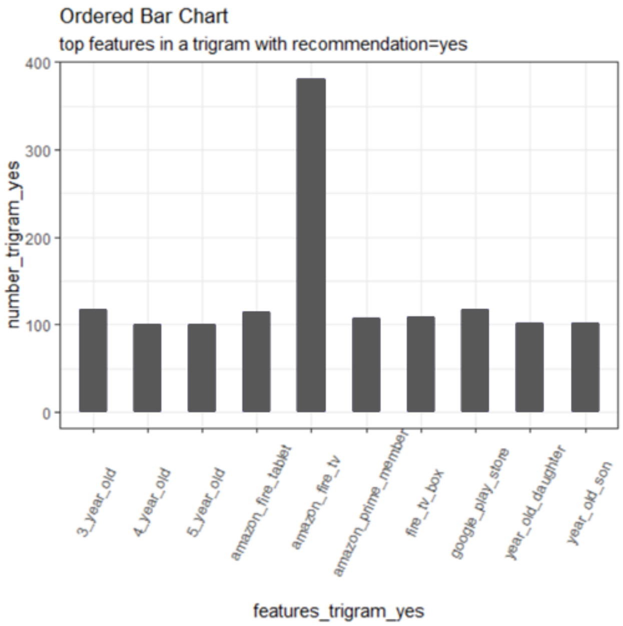


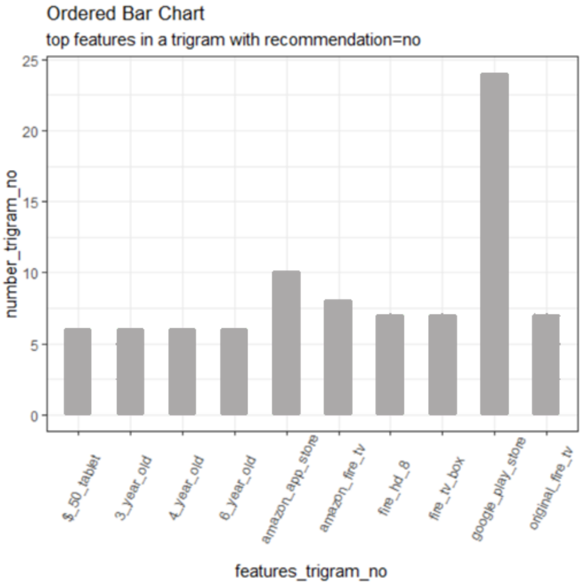
TEXT MINING

AMAZON REVIEWS

MOST FREQUENT TRI-GRAM WORDS USED FOR RECOMMENDED AND NOT RECOMMENDED PRODUCTS







CONCLUSION

From Fig 1 we saw the words which frequently occurred in the reviews and its corresponding word cloud.

Since “great” is the most frequently used word it is safe to assume that most of the customers are happy with the product.

Then in Fig 2 we saw the words commonly occuring in reviews where the product was recommended and not recommended.

We can see that majority of the people would recommend the product and the one of the issues of the product could be that of its battery power.

From the bi-grams and tri-grams we can see that the product is bought for children and is also given as gifts during Christmas.

It would be good idea to update the current product in such a way that it is child friendly and increase sales by providing offers or gift vouchers for the purchase of the product during Holidays.

REFERENCES

https://bizmapllc.com/why-reviews-are-important/#:~:text=Some%20of%20the%20reasons%20reviews,you%20in%20front%20of%20consumers

https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/

https://www.google.com/search?q=review+meaning&oq=revie&aqs=chrome.2.69i59j69i57j0l3j69i60l3.3461j0j7&sourceid=chrome&ie=UTF-8

https://www.amazon.com/Fire-Tablet-Alexa-Display-Magenta/dp/B01AHB9CN2/ref=cm_cr_arp_d_product_top?ie=UTF8

https://towardsdatascience.com/understanding-and-writing-your-first-text-mining-script-with-r-c74a7efbe30f#:~:text=A%20corpus%20is%20defined%20as,text%20file%20from%20local%20computer

https://www.quora.com/What-is-difference-between-stemming-and-lemmatization

https://www.google.com/search?q=document+term+matrix+&tbm=isch&ved=2ahUKEwiV_Zi2jOrpAhUGlksFHXxvDS4Q2-cCegQIABAA&oq=document+term+matrix+&gs_lcp=CgNpbWcQDDIECCMQJzICCAAyBAgAEB4yBggAEAgQHjIGCAAQCBAeMgQIABAYMgQIABAYMgQIABAYMgQIABAYMgQIABAYUPAzWJpOYIlsaABwAHgAgAF2iAGqDpIBBDEuMTaYAQCgAQGqAQtnd3Mtd2l6LWltZw&sclient=img&ei=4ezZXtWIKIasrtoP_N618AI&bih=642&biw=1422&safe=strict#imgrc=nO05THqJZwi5IM