جامعة أم القرى
UMM AL-QURA UNIVERSITY

Summarize Talabat Company Reviews

Group Members:

Revan Mohammed Alqahmi **441001223**
s441001223@st.uqu.edu.sa

Atheer khalid Allogamani **441007686**
s441007686@st.uqu.edu.sa

Ekram Feras Jamous **441015984**
s441015984@st.uqu.edu.sa

Shahad Aboukozzana **441003855**
s441003855@st.uqu.edu.sa

Maarib Abdullah Alsulimani **441005504**
s441005504@st.uqu.edu.sa

Lab Section: 1

Supervisor: Dr.Ashwag Magrhaby

Chapter 1

Introduction
Dataset Details
Load Dataset

Introduction

In today's business environment, understanding customers and their needs is crucial for companies to improve. Especially for delivery companies such as Talabat due to their importance and frequency of use. One way to achieve this is by reading customer reviews. However, with the huge number of reviews, performing them manually can be a time-consuming and expensive process, leading to delays in implementing improvements and customer dissatisfaction.

Using Natural Language Processing (NLP), we can quickly and efficiently develop software for classifying customer reviews about a company. This software will identify issues by classifying negative and positive reviews to help retain satisfied customers. Then, the company can gain valuable insights into its customers and improve its services.

The result will be software, based on NLP, that classify customer reviews of Talabat Company and offer a solution to the problem of efficiently processing large volumes of customer reviews.

Dataset Details

The dataset is a collection of customer reviews of 12 companies in the Arabic Language. It is stored in .csv file, and consists of 40K+ rows and 4 columns, which are index, review_description, rating, and company. Our software will summarize one of these companies which is Talabat, an Online Food Ordering Company. So, we will extract its reviews to have 32,072 rows each rated in numerical order: (-1= Negative, 0= Neutral, 1= Positive).

The data was collected in the period between 2015 and 2020. Most data was collected using web scraping and the rest was manual. The sources are Google Play store, Appstore, Google Maps reviews and Facebook reviews.

Authors names: Fahd Seddik, Marwan Salah, Abanoub Samir, Jack Mahfouz, Mohamed Galal, Abdelrahman Aymen, Youssef Amr, George Bassem, Mohamed Khairy, Ahmed taha, Laila Hamdy, Ahmed Mahmoud, Mahmoud Hesham, Mohamed Ali, Lojain Wail, Bassel Essam and Fouad Ahmed.

The link:
kaggle.com/datasets/fahdseddik/arabic-company-reviews

Load & View Dataset

using Pandas Library, we load the dataset, extract which rows we need and explore what the data contains.

Load Data

```
[3]  1
     2 # read talbat reviews only the first  32073 rows.
     3 # .iloc to delete the first and last column
     4 reviews_df = pd.read_csv("https://drive.google.com/u/0/uc?id=1DrtPc0NLOv7hvIxlyCuFiQZIudSeeFy4&export=download", nrows=32073).iloc[: , 1:-1]
     5
     6 # rename the columns
     7 reviews_df.columns = ['description','rating']
```

View Data and Explor it

```
[4]  1 reviews_df.shape
```

```
(32073, 2)
```

```
[5]  1 reviews_df.head(20)
```

	description	rating
0	رائع	1
1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	1
2	??التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكة.مع انه النت عندي تمام شو الحل	-1

```
[7]  1 reviews_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32073 entries, 0 to 32072
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   description  32072 non-null  object
 1   rating       32073 non-null  int64
dtypes: int64(1), object(1)
memory usage: 501.3+ KB
```

```
[8]  1 reviews_df.isnull().value_counts()
```

```
description  rating
False        False     32072
True         False         1
dtype: int64
```

Chapter 2

Cleaning Dataset by:
Remove Nan Values
Removing punctuation, Standardization
Tokenization using: regex and NLTK
Removing Stop- words using NLTK
Stemming
Lemmatization

Remove Nan Values

In the dataset, using isnull( ) method, we explore that there was a null value. We should remove it because it will cause errors later. The dropna() method was used to do this task.

Remove Nan Values

```
[12]  1 print("---Befor Removaing Null values---")
      2 print(f"The shapse is: {reviews_df.shape}")
      3 print(f"The number of null Values is:\n{reviews_df.isnull().value_counts()}")
```

```
---Befor Removaing Null values---
The shapse is: (32073, 2)
The number of null Values is:
description  rating
False        False     32072
True         False         1
dtype: int64
```

```
[13]  1 # Remove Null values
      2 reviews_df.dropna(inplace=True)
```

```
[14]  1 print("---After Removaing Null values---")
      2 print(f"The shapse is: {reviews_df.shape}")
      3 print(f"The number of null Values is:\n{reviews_df.isnull().value_counts()}")
```

```
---After Removaing Null values---
The shapse is: (32072, 2)
The number of null Values is:
description  rating
False        False     32072
dtype: int64
```

Remove Punctuation & Standardization

We have removed punctuation from our dataset to clean it, because Arabic has a lot of different punctuations that are not included in the punctuation library we use another technique to remove all Arabic and English punctuation by determining exactly what we want to remove.

First should import re library, then create "clean_txt_punc" method, and use (sub) from re library to define a group of special characters and replace it with spaces.

Remove punctuation

```
[ ]  #Removing punctuation
     def clean_txt_punc(input_str):
         input_str = re.sub(r'[؟،.><♡=♥/"؛\~_£-+€*&%$#@،,٪!؟):؛}{÷%("\']', '', input_str)
         return input_str

     reviews_df['no_punc']= reviews_df['description'].apply(lambda x: clean_txt_punc(x))
     reviews_df.head()

     #Reference https://2u.pw/fDYkvq
```

	description	rating	no_punc
0	رائع	1	رائع
1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع
2	??التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكةمع انه النت عندي تمام شو الحل	-1	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكةمع انه النت عندي تمام شو الحل
3	لماذا لا يمكننا طلب من ماكدونالدز؟	-1	لماذا لا يمكننا طلب من ماكدونالدز
4	البرنامج بيظهر كل المطاعم و مغلقه مع انها بتكون فاتحه بقاله كده اكثر من شهر	-1	البرنامج بيظهر كل المطاعم و مغلقه مع انها بتكون فاتحه بقاله كده اكثر من شهر

Remove Emoji

Our data have emojis so we remove them also to have clean data since they are useless. Each emoji has a specific Unicode, we removed it through define ranges of all known emojis in " emoji_pattern" and use it in "clean_txt_emoji" method.
First should import .re library, then create a method, with use defined "emoji_pattern". We remove the emojis with use (sub) from .re library and replace them with spaces.

Removing emojis

```
[ ] # define the emojies regex
    emoji_pattern = re.compile("["
            u"\U0001F600-\U0001F64F"  # emoticons
            u"\U0001F300-\U0001F5FF"  # symbols & pictographs
            u"\U0001F680-\U0001F6FF"  # transport & map symbols
            u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
            u"\U00002702-\U000027B0"
            u"\U000024C2-\U0001F251"
            u"\U000024C2-\U0001F251"
            u"\U0001f926-\U0001f937"
            u"\U00010000-\U0010ffff"
            u"\u2640-\u2642"
            u"\u2600-\u2B55"
            u"\u200d"
            u"\u23cf"
            u"\u23e9"
            u"\u231a"
            u"\ufe0f"  # dingbats
            u"\u3030"
            "]+", flags=re.UNICODE)
```

```
#Removing emojis
def clean_txt_emoji(input_str):
    input_str = (emoji_pattern.sub('', input_str))
    return input_str

reviews_df['no_emoji'] = reviews_df['no_punc'].apply(lambda x: clean_txt_emoji(x))
reviews_df.head()

#Reference https://2u.pw/v2EyOk
```

	description	rating	no_punc	no_emoji
0	رائع	1	رائع	رائع
1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع
2	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكة..مع انه النت عندي تمام شو الحل??	-1	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكةمع انه النت عندي تمام شو الحل	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكةمع انه النت عندي تمام شو الحل
3	لماذا لا يمكننا طلب من ماكدونالدز؟	-1	لماذا لا يمكننا طلب من ماكدونالدز	لماذا لا يمكننا طلب من ماكدونالدز
4	البرنامج يبظهر كل المطاعم و مطقه مع انها بتكون فاتحه بقاله كده اكتر من شهر	-1	البرنامج يبظهر كل المطاعم و مطقه مع انها بتكون فاتحه بقاله كده اكتر من شهر	البرنامج يبظهر كل المطاعم و مطقه مع انها بتكون فاتحه بقاله كده اكتر من شهر

Tokenization

To tokenize the dataset, we have used 2 method one by using Regler expression -RE- and here we used [^ء-ي] which means it will split whenever it sees any non- Arabic word character so it will split on white spaces and second by NLTK, we used word_tokenize function from nltk to tokenize the dataset .

Regex

```
# Tokenization using: 1- regex and split   2- NLTK

    #  1- regex and split
    # The regex ([^\x00-\x7F ]+) in Arabic = (\w+) in English
    # We use this unicode [\u0621-\u064A]+ or [ء-ي]+ to match only Arabic letters in Regex


def regex(text):
    text = re.split('[^ء-ي]+', text)
    return text



reviews_df ['token_regex'] = reviews_df['no_emoji'].apply(lambda x: regex(x))
```

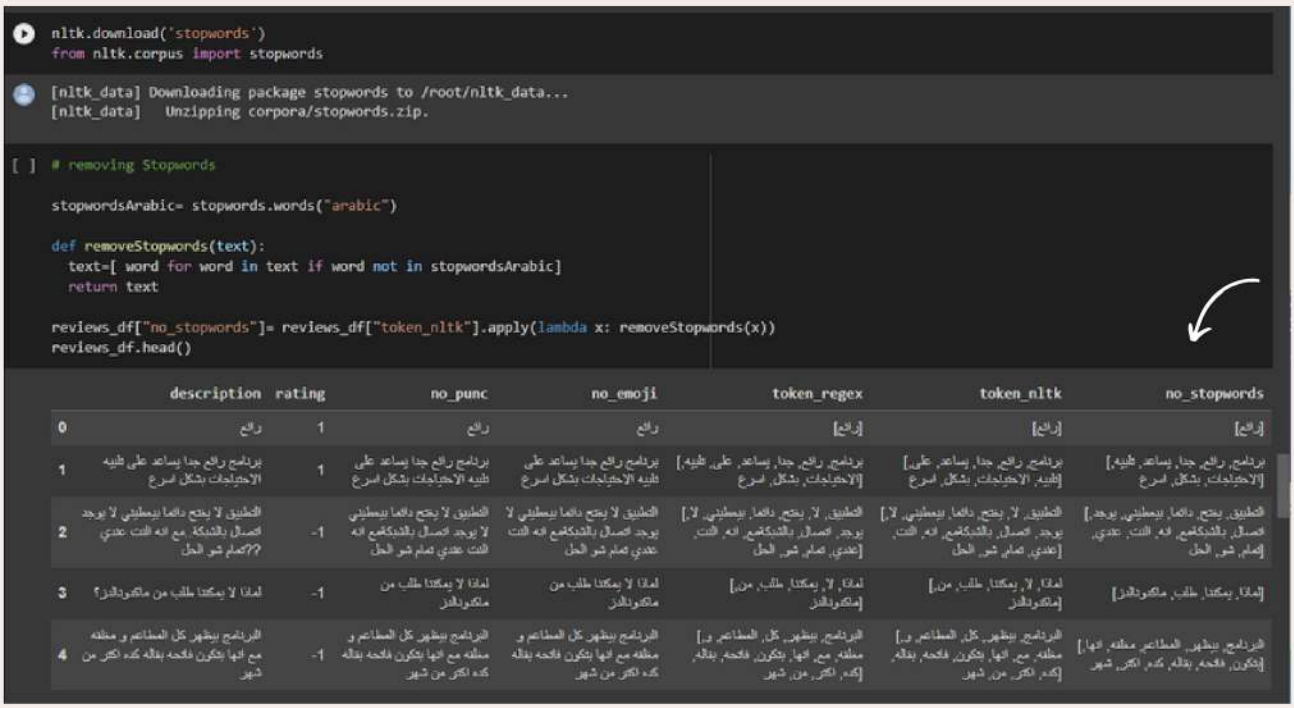
NLTK

```
# 2- NLTK

def tokenization_nltk(text):
  text = nltk.word_tokenize(text)
  return text

reviews_df ['token_nltk'] = reviews_df['no_emoji'].apply(lambda x: tokenization_nltk(x))
reviews_df.head()
```

	description	rating	no_punc	no_emoji	no_English	token_regex	token_nltk
0	رائع	1	رائع	رائع	رائع	[رائع]	[رائع]
1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	[برنامج, رائع, جدا, يساعد, على, تلبيه, الاحتياجات, بشكل, اسرع]	[برنامج, رائع, جدا, يساعد, على, تلبيه, الاحتياجات, بشكل, اسرع]
2	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكه_مع انه النت عندي تمام شو الحل??	-1	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكه انه النت عندي تمام شو الحل	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكه انه النت عندي تمام شو الحل	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكه انه النت عندي تمام شو الحل	[التطبيق, لا, يفتح, دائما, بيعطيني, لا, يوجد, اتصال, بالشبكه, انه, النت, عندي, تمام, شو, الحل]	[التطبيق, لا, يفتح, دائما, بيعطيني, لا, يوجد, اتصال, بالشبكه, انه, النت, عندي, تمام, شو, الحل]
3	لماذا لا يمكننا طلب من ماكونالدز؟	-1	لماذا لا يمكننا طلب من ماكونالدز	لماذا لا يمكننا طلب من ماكونالدز	لماذا لا يمكننا طلب من ماكونالدز	[لماذا, لا, يمكننا, طلب, من, ماكونالدز]	[لماذا, لا, يمكننا, طلب, من, ماكونالدز]
4	البرنامج يبظهر كل المطاعم و معظه مع انها بتكون فاتحه بقله كده اكثر من شهر	-1	البرنامج يبظهر كل المطاعم و معظه مع انها بتكون فاتحه بقله كده اكثر من شهر	البرنامج يبظهر كل المطاعم و معظه مع انها بتكون فاتحه بقله كده اكثر من شهر	البرنامج يبظهر كل المطاعم و معظه مع انها بتكون فاتحه بقله كده اكثر من شهر	[البرنامج, يبظهر, كل, المطاعم, و, معظه, مع, انها, بتكون, فاتحه, بقله, كده, اكثر, من, شهر]	[البرنامج, يبظهر, كل, المطاعم, و, معظه, مع, انها, بتكون, فاتحه, بقله, كده, اكثر, من, شهر]

Remove Stopwords

First, we download and import stopwords from nltk.corpus, then initialize it and specify the language (Arabic). After that, we created a function called "removeStopwords" that goes through the words one by one and removes the stopword. Finally, we created a new column and named it "no_stopwords" and we assigned to it the value of the "token_nltk" column after executing the Arabic remove stopwords function.



```
nltk.download('stopwords')
from nltk.corpus import stopwords
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
[ ] # removing Stopwords

    stopwordsArabic= stopwords.words("arabic")

    def removeStopwords(text):
      text=[ word for word in text if word not in stopwordsArabic]
      return text

    reviews_df["no_stopwords"]= reviews_df["token_nltk"].apply(lambda x: removeStopwords(x))
    reviews_df.head()
```

	description	rating	no_punc	no_emoji	token_regex	token_nltk	no_stopwords
0	رائع	1	رائع	رائع	[رائع]	[رائع]	[رائع]
1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	[برنامج, رائع, جدا, يساعد, على, تلبيه, الاحتياجات, بشكل, اسرع]	[برنامج, رائع, جدا, يساعد, على, تلبيه, الاحتياجات, بشكل, اسرع]	[برنامج, رائع, جدا, يساعد, تلبيه, الاحتياجات, بشكل, اسرع]
2	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكة مع انه انت عندي تمام شو الحل؟؟	-1	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكةمع انه انت عندي تمام شو الحل	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكةمع انه انت عندي تمام شو الحل	[التطبيق, لا, يفتح, دائما, بيعطيني, لا, يوجد, اتصال, بالشبكةمع, انه, انت, عندي, تمام, شو, الحل]	[التطبيق, لا, يفتح, دائما, بيعطيني, لا, يوجد, اتصال, بالشبكةمع, انه, انت, عندي, تمام, شو, الحل]	[التطبيق, يفتح, دائما, بيعطيني, يوجد, اتصال, بالشبكةمع, انه, انت, عندي, تمام, شو, الحل]
3	لماذا لا يمكننا طلب من ماكدونالدز؟	-1	لماذا لا يمكننا طلب من ماكدونالدز	لماذا لا يمكننا طلب من ماكدونالدز	[لماذا, لا, يمكننا, طلب, من, ماكدونالدز]	[لماذا, لا, يمكننا, طلب, من, ماكدونالدز]	[لماذا, يمكننا, طلب, ماكدونالدز]
4	البرنامج بيظهر كل المطاعم و مطلعه مع انها بتكون فاتحه بقالله كده اكثر من شهر	-1	البرنامج بيظهر كل المطاعم و مطلعه مع انها بتكون فاتحه بقالله كده اكثر من شهر	البرنامج بيظهر كل المطاعم و مطلعه مع انها بتكون فاتحه بقالله كده اكثر من شهر	[البرنامج, بيظهر, كل, المطاعم, و, مطلعه, مع, انها, بتكون, فاتحه, بقالله, كده, اكثر, من, شهر]	[البرنامج, بيظهر, كل, المطاعم, و, مطلعه, مع, انها, بتكون, فاتحه, بقالله, كده, اكثر, من, شهر]	[البرنامج, بيظهر, المطاعم, مطلعه, انها, بتكون, فاتحه, بقالله, كده, اكثر, شهر]

Stemming

We have used **Snowball** Steamer, which is also known as Porter2 because it is based on the Porter algorithm. It is designed to be language-agnostic, so we can use it with many languages, but the language must be specified in the code. We chose it because it is considered one of the most famous steamers due to its effectiveness, flexibility and ease of use.

First, we import the stemmer, then initialize it and specified the language. After that, we created a function "ArabicStemming" that go through the words one by one and executes the steaming. Finally, we create a new column and named it "StemmedText", we assigned to it the value of the "no_stopwords" column after executing the Arabic steaming function.

```python
#Stemming
from nltk.stem.snowball import SnowballStemmer

#initialize a snowball
stemmer = SnowballStemmer("arabic")

def ArabicStemming(text):
  stemmedWords = [stemmer.stem(word) for word in text]
  return stemmedWords

reviews_df["StemmedText"]= reviews_df["no_stopwords"].apply(lambda x: ArabicStemming(x))
reviews_df.head(20)
```

	description	rating	no_punc	no_emoji	token_regex	token_nltk	no_stopwords	StemmedText
0	رائع	1	رائع	رائع	[رائع]	[رائع]	[رائع]	[رايع]
1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	[برنامج, رائع, جدا, يساعد, على, تلبيه, الاحتياجات, بشكل, اسرع]	[برنامج, رائع, جدا, يساعد, على, تلبيه, الاحتياجات, بشكل, اسرع]	[برنامج, رائع, جدا, يساعد, تلبيه, الاحتياجات, بشكل, اسرع]	[رنامج, رايع, جدا, يساعد, تلب, احتياج, شكل, اسرع]
2	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكة_مع انه النت عندي تمام ??شو الحل	-1	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكةمع انه النت عندي تمام شو الحل	التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكةمع انه النت عندي تمام شو الحل	[التطبيق, لا, يفتح, دائما, بيعطيني, لا, يوجد, اتصال, بالشبكةمع, انه, النت, عندي, تمام, شو, الحل]	[التطبيق, لا, يفتح, دائما, بيعطيني, لا, يوجد, اتصال, بالشبكةمع, انه, النت, عندي, تمام, شو, الحل]	[التطبيق, يفتح, دائما, بيعطيني, يوجد, اتصال, بالشبكةمع, انه, النت, عندي, تمام, شو, الحل]	[تطبيق, يفتح, دايم, بيعط, يوجد, اتصال, شبكةمع, انه, النت, عند, تمام, شو, الحل]
3	لماذا لا يمكننا طلب من ماكدونالدز؟	-1	لماذا لا يمكننا طلب من ماكدونالدز	لماذا لا يمكننا طلب من ماكدونالدز	[لماذا, لا, يمكننا, طلب, من, ماكدونالدز]	[لماذا, لا, يمكننا, طلب, من, ماكدونالدز]	[لماذا, يمكننا, طلب, ماكدونالدز]	[لماذ, يمكن, طلب, ماكدونالدز]
4	البرنامج بيظهر كل المطاعم و مغلقه مع انها بتكون فاتحه بقالها كده اكتر من شهر	-1	البرنامج بيظهر كل المطاعم و مغلقه مع انها بتكون فاتحه بقالها كده اكتر من شهر	البرنامج بيظهر كل المطاعم و مغلقه مع انها بتكون فاتحه بقالها كده اكتر من شهر	[البرنامج, بيظهر, كل, المطاعم, و, مغلقه, مع, انها, بتكون, فاتحه, بقالها, كده, اكتر, من, شهر]	[البرنامج, بيظهر, كل, المطاعم, و, مغلقه, مع, انها, بتكون, فاتحه, بقالها, كده, اكتر, من, شهر]	[البرنامج, بيظهر, المطاعم, مغلقه, انها, بتكون, فاتحه, بقالها, كده, اكتر, شهر]	[برنامج, يظهر, مطاعم, مغلق, انه, تكو, فاتح, قال, كده, اكتر, شهر]
5	اصبح غالي جدا	-1	اصبح غالي جدا	اصبح غالي جدا	[, اصبح, غالي, جدا]	[اصبح, غالي, جدا]	[غالي, جدا]	[غال, جدا]
6	جميل جدا رائع . . .	1	جميل جدا رائع	جميل جدا رائع	[جميل, جدا, رائع]	[جميل, جدا, رائع]	[جميل, جدا, رائع]	[جميل, جدا, رايع]
7	للأسف الواحد بيصدم بعد زيادة الاسعار و للأسف بعض المطاعم اصبحت معاملتهم رديئة	-1	للأسف الواحد بيصدم بعد زيادة الاسعار و للأسف بعض المطاعم اصبحت معاملتهم رديئة	للأسف الواحد بيصدم بعد زيادة الاسعار و للأسف بعض المطاعم اصبحت معاملتهم رديئة	[للأسف, الواحد, بيصدم, بعد, زيادة, الاسعار, و, للأسف, بعض, المطاعم, اصبحت, معاملتهم, رديئة]	[للأسف, الواحد, بيصدم, بعد, زيادة, الاسعار, و, للأسف, بعض, المطاعم, اصبحت, معاملتهم, رديئة]	[للأسف, الواحد, بيصدم, زيادة, الاسعار, للأسف, المطاعم, اصبحت, معاملتهم, رديئة]	[اسف, واحد, يصدم, زياد, اسعار, اسف, مطاعم, اصبح, معامل, رديء]

Lemmatization

Qalsadi is an Arabic lemmatizer that returns the morpheme of the words. To use it, we load the module using the pip command and then import it. Next, we create a lemmatizer object that contains the "lemmatize(word)" method. This method takes a word as a parameter. So for each row, we loop each word in the list, apply this method and store them in "lemmas" Column. The lemmatization took too long, around 10 minutes, as shown below.

We did it in phase 2, but we didn't need it in Final phase.

```
[28]  1 # download arabic lemmatizer
      2 !pip install qalsadi
```

```
[31]  1 import qalsadi.lemmatizer
      2
      3 lemmer = qalsadi.lemmatizer.Lemmatizer()
      4
      5 def lem(list):
      6     lemmas = [lemmer.lemmatize(token) for token in list]
      7     return lemmas
```

```
[32]  1 import time
      2
      3 start = time.time()
      4
      5 reviews_df["lemmas"]= reviews_df["no_stopwords"].apply(lambda x: lem(x))
      6
      7 end = time.time()
      8 print(f"It took {(end - start)/60} min")
      9
     10 reviews_df.head()
```

```
It took 9.584727509816487 min
```

	description	rating	no_punc	no_emoji	no_English	token_regex	token_nltk	no_stopwords	StemmedText	lemmas
0	رائع	1	رائع	رائع	رائع	[رائع]	[رائع]	[رائع]	[رائع]	[رائع]
1	برنامج رائع هذا يساعد على تلبيه الاحتياجات بشكل اسرع	1	برنامج رائع هذا يساعد على تلبيه الاحتياجات بشكل اسرع	برنامج رائع هذا يساعد على تلبيه الاحتياجات بشكل اسرع	برنامج رائع هذا يساعد على تلبيه الاحتياجات بشكل اسرع	[برنامج, رائع, هذا, يساعد, على, تلبيه, الاحتياجات, بشكل, اسرع]	[برنامج, رائع, هذا, يساعد, على, تلبيه, الاحتياجات, بشكل, اسرع]	[برنامج, رائع, هذا, يساعد, تلبيه, الاحتياجات, بشكل, اسرع]	[برنامج, رائع, هذا, يساعد, تلبيه, احتياج, بشكل, اسرع]	[برنامج, رائع, هذا, يساعد, تلبيه, احتياج, بشكل, سرع]
2	التطبيق لا يفتح دائما يعطيني لا يوجد اتصال بالشبكة مع انه ??انت عندي تمام شو الحل	-1	التطبيق لا يفتح دائما يعطيني لا يوجد اتصال بالشبكة مع انه انت عندي تمام شو الحل	التطبيق لا يفتح دائما يعطيني لا يوجد اتصال بالشبكة مع انه انت عندي تمام شو الحل	التطبيق لا يفتح دائما يعطيني لا يوجد اتصال بالشبكة مع انه انت عندي تمام شو الحل	[التطبيق, لا, يفتح, دائما, يعطيني, لا, يوجد, اتصال, بالشبكة, مع, انه, انت, عندي, تمام, شو, الحل]	[التطبيق, لا, يفتح, دائما, يعطيني, لا, يوجد, اتصال, بالشبكة, مع, انه, انت, عندي, تمام, شو, الحل]	[التطبيق, يفتح, دائما, يعطيني, يوجد, اتصال, بالشبكة, مع, انه, انت, عندي, تمام, شو, الحل]	[تطبيق, يفتح, دايم, يعطي, يوجد, اتصل, شبكة, مع, انه, انت, عند, تمام, شو, الحل]	[تطبيق, يفتح, دائم, يعطيني, أوجد, اتصال, بالشبكة, مع, انه, انت, عند, تمام, شوى, حل]
3	ليدا لا يمكنا طلب من متكودالنز ؟	-1	ليدا لا يمكنا طلب من متكودالنز	ليدا لا يمكنا طلب من متكودالنز	ليدا لا يمكنا طلب من متكودالنز	[ليدا, لا, يمكنا, طلب, من, متكودالنز]	[ليدا, لا, يمكنا, طلب, من, متكودالنز]	[ليدا, يمكنا, طلب, متكودالنز]	[ليدا, يمكن, طلب, ماكودالنز]	[ليدا, امكن, طلب, ماكودالنز]
4	البرنامج يظهر كل المطاعم و محطة مع انها بتكون قاتحه بقالة كده اكثر من شهر	-1	البرنامج يظهر كل المطاعم و محطة مع انها بتكون قاتحه بقالة كده اكثر من شهر	البرنامج يظهر كل المطاعم و محطة مع انها بتكون قاتحه بقالة كده اكثر من شهر	البرنامج يظهر كل المطاعم و محطة مع انها بتكون قاتحه بقالة كده اكثر من شهر	[البرنامج, يظهر, كل, المطاعم, و, محطة, مع, انها, بتكون, قاتحه, بقالة, كده, اكثر, من, شهر]	[البرنامج, يظهر, كل, المطاعم, و, محطة, مع, انها, بتكون, قاتحه, بقالة, كده, اكثر, من, شهر]	[البرنامج, يظهر, المطاعم, محطة, انها, بتكون, قاتحه, بقالة, كده, اكثر, شهر]	[برنامج, يظهر, مطاعم, محطق, انها, تكون, قاتح, قال, كده, اكثر, شهر]	[برنامج, يظهر, مطاعم, محطق, انها, تكون, قاتح, بقال, ده, اكثر, شهر]

Work Distribution

Task	Revan	Ekram	Shahad	Atheer	Maarib
Introduction	*	*			
Dataset Details		*		*	
Load & View Dataset			*		
Remove Nan Values		*			
Remove Punctuation & Standardization					*
Remove Emoji					*
Tokenization using: regex and NLTK			*		
Remove Stopwords				*	
Stemming	*				
Lemmatization	*	*			

Chapter 3

Main Objective
Project Users
Project Methodology
The Code
Work Distribution

Introduction

In today's business environment, understanding customers and their needs is crucial for companies to improve. Especially for delivery companies such as Talabat due to their importance and frequency of use. One way to achieve this is by reading customer reviews. However, with the huge number of reviews, performing them manually can be a time-consuming and expensive process, leading to delays in implementing improvements and customer dissatisfaction.

Using Natural Language Processing (NLP), we can quickly and efficiently develop software for classifying customer reviews about a company. This software will identify issues by classifying negative and positive reviews to help retain satisfied customers. Then, the company can gain valuable insights into its customers and improve its services.

The result will be software, based on NLP, that classify customer reviews of Talabat Company and offer a solution to the problem of efficiently processing large volumes of customer reviews.

Main Objective

The main purpose of this project is to help business owners analyze customer reviews more easily and reduce their burden by making it easier for them to read the results.

Project Users

Business owners, especially owners of stores on Talabat application.

Project Methodology

Project methodology for the NLP project:

1. Problem Definition and Data Collection:
Defining the project problem and objectives, which is to analyze a dataset of customer reviews for a delivery company using NLP techniques.
Collect customer reviews dataset.

2. Data Preprocessing:
- Clean data by removing irrelevant information, such as punctuation marks, emojis.
- Mark up text into words and remove stop words.
-We used regex to clean the data
- Stem words to reduce vocabulary.
Encoding text data into numeric features using TF-IDF.

3. Model selection:
Select appropriate NLP algorithms for the project, such as SVM, Naive Bayes, and Random Forest.
Train and test each algorithm on pre-processed data and evaluate its performance using appropriate performance metrics, such as accuracy and time.

4 . Project documentation:
Document the entire project methodology, including data pre-processing, model selection, and optimization.
A detailed explanation of each step and the rationale for decisions made throughout the project is included.
Provide clear instructions on how to reproduce the project and operate models for future reference.

Data Prepreation

Load and Explore the Data
Remove Nan Values
Clean the Data
Remove Empty Rows

Load and Explore Data

The dataset is a collection of Arabic reviews of several companies stored in csv file. We extract the reviews of the first company to have 32,072 rows each rated in numerical order: (-1= Negative, 0= Neutral, 1= Positive).

Load and Explor Data

```
[4]
# read talbat reviews only the first  32073 rows.
# .iloc to delete the first and last column
reviews_df = pd.read_csv("https://drive.google.com/u/0/uc?id=1DrtPc0NLOv7hvIxlyCuFiQZIudSeeFy4&export=download", nrows=32073).iloc[: , 1:-1]

# rename the columns
reviews_df.columns = ['description','rating']
```

```
[5]
reviews_df.head()
```

	description	rating
0	رائع	1
1	برنامج رائع جدا يساعد على تلبيه الاحتياجات بشكل اسرع	1
2	??التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكة مع انه النت عندي تمام شو الحل	-1
3	لماذا لا يمكننا طلب من ماكدونالدز؟	-1
4	البرنامج بيظهر كل المطاعم و مغلقه مع انها بتكون فاتحه بقاله كذه اكثر من شهر	-1

```
[6]
reviews_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32073 entries, 0 to 32072
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   description  32072 non-null  object
 1   rating       32073 non-null  int64
dtypes: int64(1), object(1)
memory usage: 501.3+ KB
```

Remove NaN Values

The dataset has Nan value. So, we removed it to aviod the errors later.

Clean Data

In this step, we removed unnecessary characters that wouldn`t affect the results. The Arabic Language has a lot of them which are:

1. Arabic punctuation: ؛ ؟ ،
2. Diacritic:
3. Non-Arabic letters
4. Duplicated letters: إبداااااع
5. Emojis
6. digits and numbers

After that we applied the tokenization, stemming.

```python
def clean_text(text):

  # remove punct
  no_punc = re.sub(punctuation_re, ' ', text)

  # remove duplicated letters
  no_duplicate = re.sub(r'([\u0600-\u06FF])\1+', r'\1', no_punc)  #re.sub("ـ" ,"+ـ", no_punc)

  # remove non arabic letters and emojis
  no_english = re.sub(r'[^0-9\u0600-\u06ff\u0750-\u077f\ufb50-\ufbc1\ufbd3-\ufd3f\ufd50-\ufd8f\ufd50-\ufd8f\ufe70-\ufefc\uFDF0-\uFDFD.0-9]+', ' ', no_duplicate)

  # tokenize
  tokens = nltk.word_tokenize(no_english)

  #Stemming
  stemmedWords = [stemmer.stem(word) for word in tokens]

  return ' '.join(stemmedWords)
```

```python
reviews_df['clean_text'] = reviews_df['description'].apply(lambda x: clean_text(x))
reviews_df.head()
```

	description	rating	clean_text
0	رائع	1	رائع
1	برنامج رائع جدا يساعد علي اللغه الاحتياجات بشكل اسرع	1	برنامج رائع جدا يساعد علي لغه احتياج شكل اسرع
2	??التطبيق لا يفتح دائما بيعطيني لا يوجد اتصال بالشبكة مع انه النت عندي تمام شو الحل	-1	تطبيق لا يفتح دايم يعط لا يوجد اتصال شبكه مع انه النت عند تمام شو الحل
3	لماذا لا يمكنا طلب من ماكدونالدز؟	-1	لماذا لا يمكنا طلب من ماكدونالدز
4	البرنامج يظهر كل المطاعم و مطلق مع انها تكون فاتح قال كده اكثر من شهر	-1	برنامج يظهر كل مطاعم و مطلق مع انها تكون فاتح قال كده اكثر من شهر

Remove Empty Row

After cleaning, there were empty reviews so we removed them.

Remove Empty Reviews(Rows)

```python
reviews_df.rating.value_counts()
```

```
 1    19894
-1    10696
Name: rating, dtype: int64
```

```python
reviews_df = reviews_df[reviews_df.astype(str)['clean_text'] != '[]']
reviews_df.rating.value_counts()
```

```
 1    19894
-1    10696
Name: rating, dtype: int64
```

ML Model

Balance the Data
Extract Features
Split the Data
Apply ML Algorithm

Balance the Data

First, the number of neutral rating data is very small so we didn`t include it in the training. Another thing, the number of positives was the double of negative ones. Hence, we removed some of them orbitary.

Extract Features Using TD-IDF

To extract the import words in each review, and convert them to word vectors to train the Model, we use TD-IDF vectorization.

- using TFIDF

```
[18] from sklearn.feature_extraction.text import TfidfVectorizer
```

```
[19] tdidf_vect= TfidfVectorizer()
     tdidf = tdidf_vect.fit_transform(reviews_df['clean_text'])


     features_tdidf = pd.DataFrame(tdidf.toarray())
     features_tdidf.columns = tdidf_vect.get_feature_names_out()
     features_tdidf
```

	***	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	...	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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
1	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
2	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
3	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
4	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
21585	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
21586	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
21587	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
21588	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
21589	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

21590 rows × 13666 columns

Other Features

we claimed that there are other features that can affect the classification. Such as the number of words where usually angry people complain and explain their problems which increases the number of words in a negative review. Also, they use a lot of exclamation and questions mark.

- Number of Words

```
[29] reviews_df['number_of_words'] = reviews_df['description'].apply(lambda x: len(x.split(" ")))
```

- Number of Punctuation

```
[30] def count(text):
       count = sum([1 for char in text if char in ['!', '?', '؟']])
       return count

     reviews_df['number_of_?!'] = reviews_df['description'].apply(lambda x: count(x))

```

Also, we noticed that there are different emojis used with each rate. So, we extract them and convert them to vectors.

Type of emojies

```
[27]  1 def extractEmoji(text):
      2   return str(set([char for char in text if re.match(emoji_re, char)]))
      3
      4 reviews_df['type_of_emoji'] = reviews_df['description'].apply(lambda x: extractEmoji(x))
      5
      6 tdidf_e = TfidfVectorizer()
      7 tdidf = tdidf_e.fit_transform(reviews_df['type_of_emoji'])
      8 features_e = pd.DataFrame(tdidf.toarray())
      9 features_e.columns = tdidf_e.get_feature_names_out()
```

Unfortunately, using these features: the number of words, the number of exclamation and question marks, and emojis weren`t affected. Instead, they reduce the accuracy and we thought that because of their lack in the dataset. However, we still used TFIDF for features.

Feature and Label

Final Features

```
[48]  1 labels = reviews_df['rating']
      2 features = pd.concat( [features_tdidf], axis=1)
      3 features.shape
```

```
(21590, 13666)
```

Split Data to 3:1 for Traininng:Test

Split data

```
[49]  1 from sklearn.model_selection import train_test_split
      2
      3 x_train, x_test, y_train, y_test = train_test_split(features, labels, test_size=0.25)
      4
      5 print(x_train.shape, x_test.shape, y_train.shape, y_test.shape)
```

```
(16192, 13666) (5398, 13666) (16192,) (5398,)
```

Apply Different ML Algorithms

lastly, after applying different classification algorithms, we found that Random Forest, Multinomial Naive Bayes, and SVC achieved roughly same accuracy. However, SVC was very slow.

Random Forest

```
[50]  1 from sklearn.ensemble import RandomForestClassifier
      2 model = RandomForestClassifier(n_jobs=-1)
      3 model.fit(x_train, y_train)
      4 print(model.score(x_test, y_test))
```

```
0.8601333827343346
```

Naive Bayes

```
[51]  1 from sklearn.naive_bayes import GaussianNB
      2 NBmodel = GaussianNB()
      3 NBmodel.fit(x_train, y_train)
      4 print(NBmodel.score(x_test, y_test))
```

```
0.671359762875139
```

```
1 from sklearn.naive_bayes import MultinomialNB
2 NBmodel = MultinomialNB()
3 NBmodel.fit(x_train, y_train)
4 print(NBmodel.score(x_test, y_test))
```

```
0.8658762504631345
```

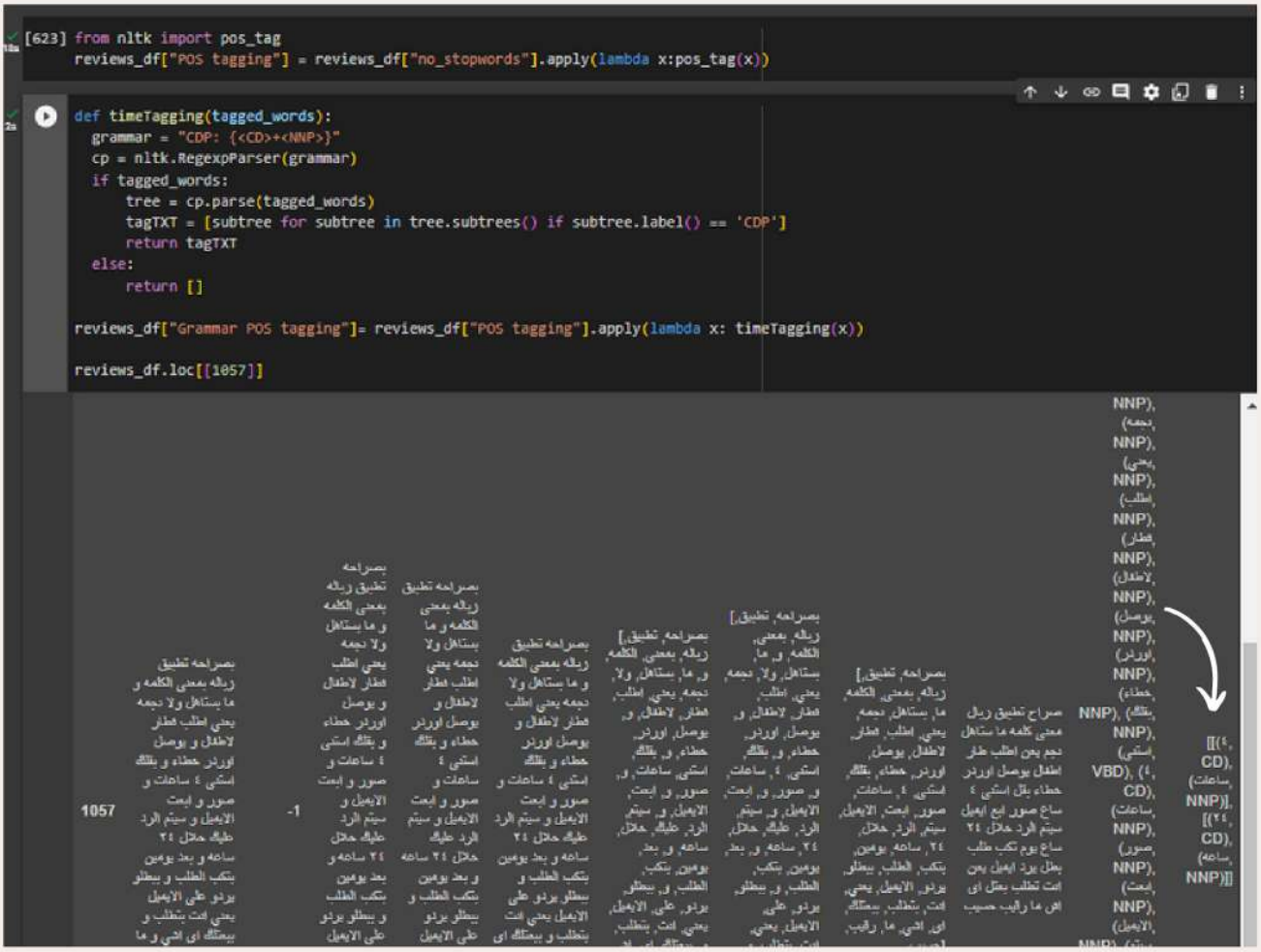
SVC

```
[38]  1 from sklearn import model_selection, naive_bayes, svm
      2
      3 SVM_model = svm.SVC(C=1.0, kernel='linear', degree=3, gamma='auto')
      4 SVM_model.fit(x_train, y_train)
      5 print(SVM_model.score(x_test, y_test))
```

```
0.8638384586884031
```

POS Tagging

In POS Tagging we wanted to find the hours in both positive and negative reviews to find the range that makes the users satisfied and the range that makes them not. unfortunately, we didn't get any output from the grammar "CDP: {<CD>+<ساعات>}" so we ended up writing it as "CDP: {<CD>+<NNP>}"

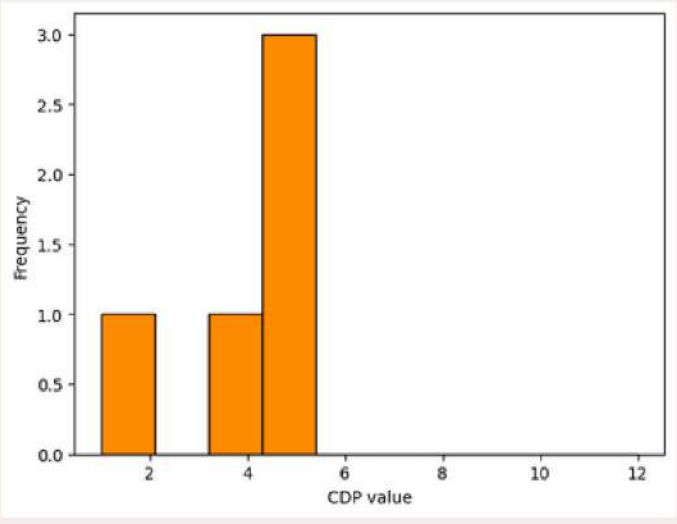


CD counting and chart

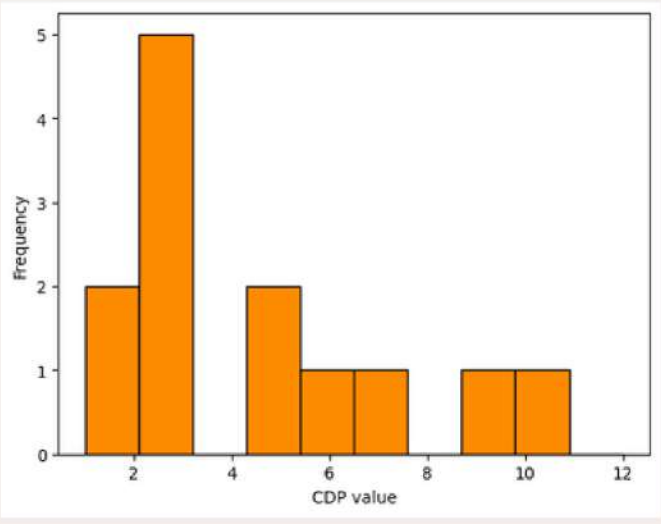
To help we removed the big numbers by writing a rule to count only CD values with 12 or less (more than 12 make them zero). Of course, it's not the best idea, and we can see it clearly with the non-logical difference between the positive and negative 'hours' numbers in the histogram chart.

The same code with positive reviews, but we switch -1 to 1.

```
def extractCD(text, rating):
    if rating == -1:
        tagged_words = [word for subtree in text for word in subtree]  # flatten the list of subtrees into a list of tuples
        cdp_values = [re.sub('[^\d\u0660-\u0669]', '', number) for (number, tag) in tagged_words if tag == "CD"] # filter out any non-numeric characters from the input text
        cdp_values = [int(num) if int(num) <= 12 else 0 for num in cdp_values if num]  # convert to integers, ignoring empty strings, and set any value greater than 12 to 0
        if len(cdp_values) == 0:
            return 0
        else:
            return sum(cdp_values)
    else:
        return 0

reviews_df["CD in CDP grammar"] = reviews_df.apply(lambda row: extractCD(row["Grammar POS tagging"], row["rating"]), axis=1)

# Filter the DataFrame to include only rows with non-zero cd_values
nonzero_cd_values_df = reviews_df[reviews_df["CD in CDP grammar"] != 0]

#  Histogram of the CD values with a maximum value of 12 and an orange color
plt.hist(nonzero_cd_values_df["CD in CDP grammar"], range=(1, 12), color='darkorange', ec="black")
plt.xlabel("CDP value")
plt.ylabel("Frequency")
plt.show()
```



positive reviews hours



negative reviews hours

Work Distribution

Task	Revan	Ekram	Shahad	Atheer	Maarib
Main Objective & Project Users	*				
Project Methodology					*
POS Tagging	*			*	
CD counting and chart	*			*	
Load Data and Explor it	*	*	*	*	*
Data Cleaning	*	*	*	*	*
Balance the Data and Remove Nan Values		*	*		
Extracting Features		*	*		
TFIDF Vectoraization			*		
Random Forest		*	*		
Naive Bayes	*				*
SVC	*				*
Documentaion	*	*	*	*	*

Colab Link

This is the python file in Colab

https://colab.research.google.com/drive/1dzGAN0F5QcpDqS1VB-29Bg5kif0AHrWU?usp=sharing

https://colab.research.google.com/drive/1dzGAN0F5QcpDqS1VB-29Bg5kif0AHrWU?usp=sharing