كلية علوم الحاسب والمعلومات
قسم هندسة البرمجيات

Student's name	id
1- Taraf Alsuhaim	438201957
2- Majd Bin musibeh	438200829
3- Muntaha jaber	437200150
4- Shammaa Alsomaikhi	438200218
5- Maha Alshammari	438201023

SWE486 PROJECT

Group#11
section:54979
Phase 4

content

Overview

The remaining of the document includes description of our chosen company and the problem that we want to solve at the end of the project, extracted data, three problems that we solve it for preparing the data. General information about our data. Report of the frameworks/libraries that we used in our project, and provide a list of all our development files with descriptions.

Why did we start this project?

We start this project for "Big Data and Cloud Computing" (SWE486) course. We believe that this project will improve our technical and practical data analytical skills.

What is "Al-Rajhi Bank"?

Al-Rajhi Bank is a Saudi Arabian bank and the world's biggest Islamic bank that was established in 1957. The bank is a significant speculator in Saudi Arabia and is one of the biggest business entities in the realm, with more than 600 branches. Its administrative center is situated in Riyadh, with six provincial workplaces. Al-Rajhi Bank likewise has branches in Kuwait and Jordan, and an auxiliary in Malaysia and Syria. The bank offers an assortment of banking administrations, for example, stores, advances, speculation guidance, protections exchanging, settlements, Visas, and purchaser financing. All administrations are offered by Islamic necessities. The bank has won various honors for its Middle East activities.

Description of the problem:

There is no doubt that is every bank exist wants as much as possible of customers, and to do that you need to get customers first, then make sure that they will stay using your bank and then target new people (potential customers), so what we are trying to say here if you keep getting customers and they left your bank this is wasted effort for sure! the "Al-Rajhi bank" is very huge here in Saudi Arabia, but it's one of these banks that have this serious problem because a lot of their customers actually leave "Al-Rajhi bank" and go to another bank.

Our main goal of this project:

1-Try to find out the main reason of customers leaving Al-Rajhi bank and go to other banks. Also, collect every single thing about this problem.
2-find out the counter solutions that reduce our problem.
3-find out the weakness points in Al-Rajhi bank and the services that is provided in the other banks that tempt the customers to it.
4-Try to find out strengths that lead to develop and prosperity to Al-Rajhi bank.

Description of data:

As we said before our problem is “why Al-Rajhi bank customers leave and go to other banks?” so we looked for negative hashtags that people were talking about Al-Rajhi bank badly, to search about the reasons which made them leave.

General information about our data

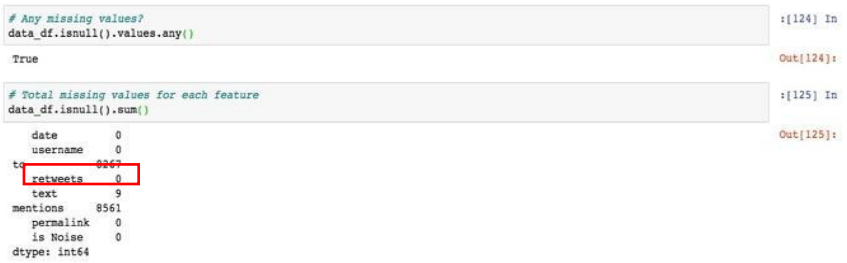
We've explored our data which was 9452 number of variety tweets, as we were exploring our data, we found out multiple tweets were against Al-Rajhi bank and some of the tweets were supporting them. We've also noticed that our data contains multiple unrelated tweets to our needs such as advertisement tweets, images and videos, we also found unrelated hashtags to Al-Rajhi bank and many null value data.
The number of tweets after cleaning is 4112.

What we noticed when we investigated our data:

Since we utilized "get old tweets" library, that we discussed about it in phase1, to extract our data we noticed the advantages and disadvantages of this library. This library is not only about extracting an unlimited number of tweets, yet it also did a little bit of cleaning. Since we didn't end up with a duplicate tweet or an emoji problem thanks to this powerful library. However one thing that trouble us that once we extract a tweet that includes "hashtag" we asked for it, it also does extract all tweets that replies to that tweet regardless of whether they have the hashtag or not, but it was a little price for an unlimited amount of tweets extraction for us.

Four problems in our data

- Problem 1 : remove photo or video (null tweets) .

Simple explain	Some tweets contain video or photo and we just need to analyze tweets which have text only. Because, it is difficult to analyze video/photo as they may not have any word in the excel file.
How we find it ?	When we explore our data and apply code of missing data, we found 9 tweets that have null value then we traceback to actual tweets we found out it was a video or photo. 
How we solve it ?	We used same code that we obtained from the workshop and apply it on our data. The code removes the tweets that have null value .
How the solution helped us?	Since in this course we're only learning how to analyze text data, photos or videos won't be helpful for us to analyze them so we decided to delete them so we can analyze our text data easily without any other data won't be beneficial for us.

- Problem 2 : remove tweets containing advertisements .

Simple explain	Some tweets contain advertisements and we don't need these tweets in our data, we just need to analyze tweets which contain helpful text only.
How we find it ?	When we explored our data we saw many worthless tweets so we considered these tweets as noisy ones since it's not helpful to us when analyzing our data in the next phase.
How we solve it ?	We write a piece of code to remove any noisy tweets since advertisements is considered noisiness to us. 

How the solution helped us?	After this code our data became a lot helpful to us without the noisy tweets containing the advertisements

- Problem 3 : remove unrelated tweets .

Simple explain	it's not like "delete tweets containing advertisements" problem because as we talked before in “What we noticed when we investigated our data" using GetOldTweets lead us to this problem since it also gives us replies tweet like: "hi", "good morning"...etc. these are not advertisements tweets but yet does not make any sense and it’s not related to our problem.
How we find it	When we explored our data, we saw many tweets that doesn’t make sense to us, so we considered these tweets as "not related" since they are not related to Al-Rajhi Bank in any way.
How we solve it	We used the same code that we obtained from the workshop with little edit and apply it to our data. The code removes the tweets that label as “not related”.
How the solution helped us?	We minimize the number of tweets and that helped us to classify them to positive and negative data.

- Problem 4 : clean the text from (punctuation marks , links ,repeated letters , hashtags)

Simple explain	We found in our data some tweets contain symbols, hashtags, punctuation marks, links and repeated letters like(# , !, _ , @) these are unnecessary for us to keep them.
How we find it ?	When we explored our data, we saw many unwanted symbols like punctuation marks ,repeated letters and hashtags.
How we solve it ?	We used the same code that we obtained from the workshop and apply it on our data. The code removes symbols, hashtags and repeated letters, therefore that will make data analysis easier in next phase.
How the solution helped us?	After this code our data became a lot helpful to us without the noisy tweets containing the strange signals, symbols and hashtags.

Descriptive Analysis:

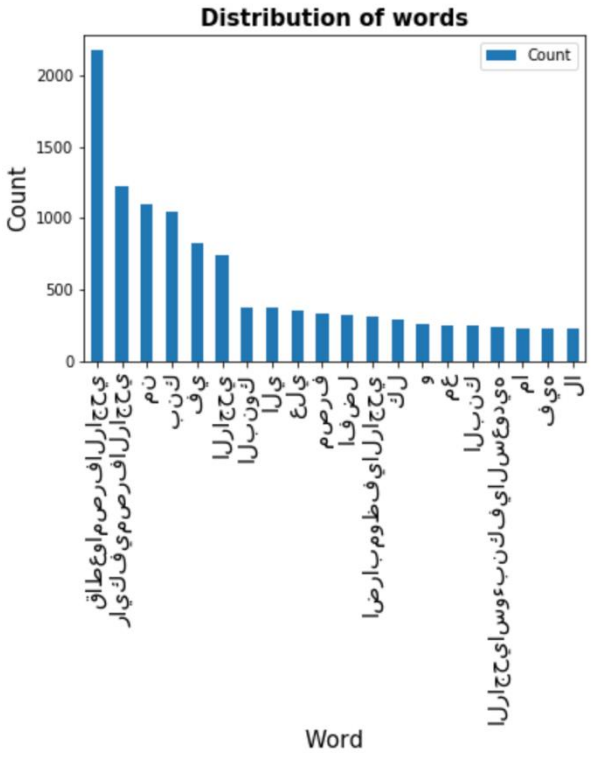
"A descriptive analysis is an important first step for conducting statistical analyses. It gives you an idea of the distribution of your data, helps you detect outliers and typos, and enable you identify associations among variables, thus making you ready to conduct further statistical analyses." [6]

```
# show the most common words
word_counter.most_common(10)
```

```
,(2176 ,'قاطعوامصرفالراجحي')]
,(1225 ,'رايكفيمصرفالراجحي')
,(1094 ,'من')
,(1042 ,'بنك')
,(827 ,'في')
,(743 ,'الراجحي')
,(374 ,'البنوك')
,(372 ,'الي')
,(351 ,'علي')
[(330 ,'مصرف')
```

```
# Display 10 least common lines
word_counter.most_common()[-10:]
```

```
,(1 ,'الس')]
,(1 ,'الفصل')
,(1 ,'التعسفي')
,(1 ,'ومضايقه')
,(1 ,'الوئام')
,(1 ,'اكتبوا')
,(1 ,'الديوان')
,(1 ,'وسجلوا')
,(1 ,'وارفعوا')
[(1 ,'برقيات')
```



services_count	system_count	sentiment	neg_count	pos_count	text length	retweets	
4111.000000	4111.000000	4111.000000	4111.000000	4111.000000	4111.000000	4111.000000	count
0.465337	0.237898	0.399173	0.229141	0.118463	77.331063	0.791778	mean
0.987120	0.755730	0.489788	0.501615	0.367575	49.436805	10.290853	std
0.000000	0.000000	0.000000	0.000000	0.000000	8.000000	0.000000	min
0.000000	0.000000	0.000000	0.000000	0.000000	41.000000	0.000000	25%
0.000000	0.000000	0.000000	0.000000	0.000000	65.000000	0.000000	50%
0.000000	0.000000	1.000000	0.000000	0.000000	105.000000	0.000000	75%
7.000000	5.000000	1.000000	3.000000	3.000000	279.000000	496.000000	max

Analysis methods that we used:

- **logistic regression:**

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. The outcome is measured with a dichotomous variable (in which there are only two possible outcome) ,the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.). or in case of our data set whether the AL-Rajhi bank customers are satisfied or not. The purpose of logistic regression is to estimate the probabilities of events, including determining a relationship between features and the probabilities of particular.[10]

Why we used this model: since our data set can take two possible values such as satisfied ,not satisfied. we decided that the Logistic regression is most suitable for our data set. Also, Logistic regression has less prone to over-fitting ,and it's easier to implement, interpret and very efficient to train using Logistic regression.
We have used Logistic regression to classify our data into two parts, the first part the data has classified based on whether the customer is "satisfied , not satisfied" with AL-Rajhi bank . The second part classified our data based on the reasons the customers are not satisfied with the bank and its classified into , not satisfied because of services or not satisfied because of bank system was not working properly as " service, system" respectively.[8]

First Part

Confusing matrix table:[15]

Confusion Matrix is a performance measurement for machine learning classification and its used to evaluate the performance of our classification model .

	Predicted satisfied (1)	**Predicted not satisfied (0)**
Positive(1)	354 TP	516 FN
Negative(0)	30 FP	1156 TN

We summarized from the confusion matrix the following information:[14][9]

- There are 2 predicted classes: " not satisfied " and " Satisfied" customers.
- By adding the four values from the table we found out the classifier made 2056 prediction.
- The classifier predicted " not satisfied " 1672 times (1156+516=1672)
- The classifier predicted " Satisfied" 384 times (30+354=384)
- **Accuracy:** (TP+TN)/(TP+FP+FN+TN)(156+354)/(1156+354+30+516)=0.73
- **Misclassification rate:** (accuracy-1),1-0.73=0.27
- **True positive rate (TPR)** :TPR=TP/(TP+FN)354/(354+516)=0.40
- **True negative rate(TNR):** TNR=TN/(TN+FP) 1156/(1156+30)=0.97
- **False positive rate(FPR)** : FPR= FP/FP+TN=30/(30+1156)=0.02
- **False negative rate(FNR)** : FNR= FN/FN+TP 516/(516+354)=0.59

classification report of logistic classifier table:

class	precision	recall	F1-score	support
(0) **not Satisfied**	0.69	0.97	0.81	**1186**

	(1) **Satisfied**	0.92	0.41	0.56	**870**
accuracy	-----	-----	-----	0.73	**2056**
Macro avg	-----	0.81	0.69	0.69	**2056**
Weighted avg	-----	0.79	0.73	0.71	**2056**

We summarized from classification report of logistic classifier the following information :[11][12]

- **Class(0):** not satisfied with AL-Rajhi bank.
- **Class(1):** satisfied with AL-Rajhi bank.

Precision: is the ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as the ratio of true positives to the sum of true and false positives in other words Precision – "for all instances classified positive, what percent was correct?".

Precision :

- **when its predicted "**not satisfied **" how often its correct** 0.69 TP/TP+FP (1156)/(1156+516) {from Confusing matrix}
- **when its predicted "** satisfied**" how often its correct** 0.92 (354)/(354+30) {from Confusing matrix}

Recall : is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives. In other word, "for all instances that were actually positive, what percent was classified correctly?"

- **Recall** =TP/TP+FN 354/354+516=0.41
- From the table above we observed the Recall for 'Satisfied' customers is 0.41 which represent a bad parentage.

F1 score: What percent of positive predictions were correct
The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0 , 'not satisfied' has higher F1 however both classes conceder to has a good F1 score.

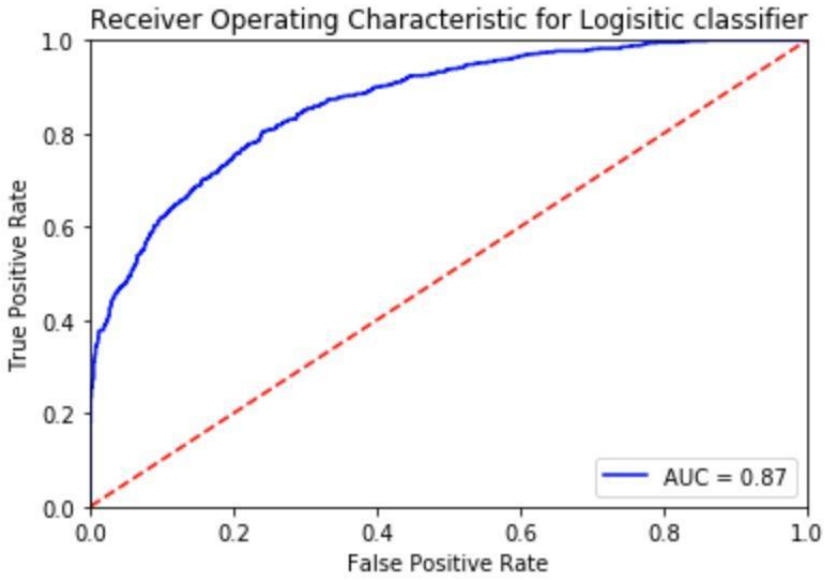
Support: Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

- Since in our data the support of "Satisfied" is not close to " Satisfied " this indicate the balance is not quite good.

ROC curve:[13]

It's a graph used to summarize the performance of a classifier over all possible threshold. Its generated by plotting the true positive rate(y-axis) against the false positive rate(x-axis).

- False positive rate: FP/FP+TN□30/(30+1156)=0.02
- True positive rate: TP/(TP+FN)□354/(354+516)=0.40



- The relationship between FPR AND TPR is immediate when FPR increase the TPR also increases as it asperses in the ROC curve the area under the curve is large therefor the result classified as good. And the accuracy of the curve is 0.73[13]

What we summarize from analyzing our data using logistic regression:

Since in our logistic regression model the accuracy=0.73 is high ,so we can consider the classification acceptable, although there is a small percentage of error 0.27 this error due to wrong classification data , there could be many reasons for wrong classification such as systemic error in the test for the outcome, or maybe the model is too simple to capture the patterns in data, data has too much noise etc.[16]

Second Part

Confusing matrix table:[15]

	Predicted system (1)	**Predicted Services (0)**
Positive(1)	70 TP	107 FN
Negative(0)	0 FP	479 TN

We summarized from the confusion matrix the following information:[9][14]

- There are 2 predicted classes: " Services" and " system".
- By adding the four values from the table we found out the classifier made 656 prediction.
- The classifier predicted " Services" 586 times (479+107=586).
- The classifier predicted " system" 70 times (0+70=70).

- **Accuracy** : (TP+TN)/(TP+FP+FN+TN)=(479+70)/(479+70+0+70)=0.84
- **Misclassification rate:** (accuracy-1),1-0.84=0.16
- **True positive rate (TPR)** :TPR=TP/(TP+FN)70/(0+70)=0.39
- **True negative rate(TNR):** TNR=TN/(TN+FP) 479/(479+0)=1
- **False positive rate(FPR**) : FPR= FP/FP+TN=0/(0+479)=0
- **False negative rate(FNR**) : FNR= FN/FN+TP 107/(107+70)=0.60

classification report of logistic classifier table:

	class	precision	recall	F1-score	support
	(0) **services**	0.82	1.00	0.90	**479**
	(1) **system**	1.00	0.40	0.57	**177**
accuracy	-----	-----	-----	0.84	**656**
Macro avg	-----	0.91	0.70	0.73	**656**
Weighted avg	-----	0.87	0.84	0.81	**656**

We summarized from classification report of logistic classifier the following information :

- **Class(0):** "services" is the reason customers are not satisfied with AL-Rajhi bank.
- **Class(1):** " system" is the reason customers are not satisfied with AL-Rajhi bank.

Precision :

- **when its predicted** " services " **how often its correct** 0.82
- **when its predicted** " system " **how often its correct** 0.1

Recall : for all instances that were actually positive, what percent was classified correctly?

- " services " class has 1 recall witch is very good.
- " system " class has 0.4 witch is not very good percentage.

F1 score: " services " has a very high F1 score but " system " considered good.

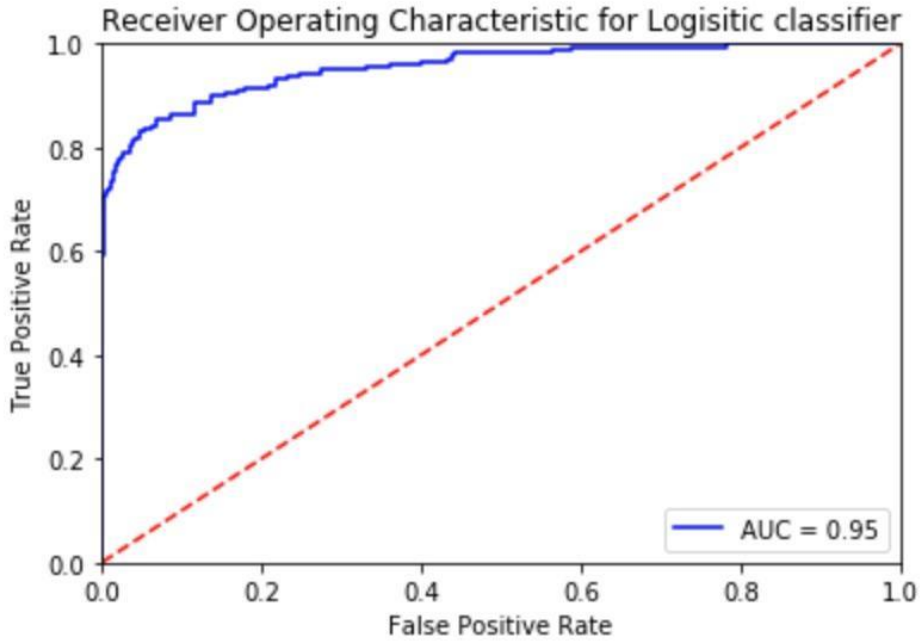
Support:

- Since in our data the support of " services " is not close to " system " this indicate the balance is not quite good.

ROC curve:[13]

It's a graph used to summarize the performance of a classifier over all possible threshold. Its generated by plotting the true positive rate(y-axis) against the false positive rate(x-axis).

- False positive rate(FPR) : FPR= FP/FP+TN=0/(0+479)=0
- True positive rate (TPR) :TPR=TP/(TP+FN)70/(0+70)=0.39



The relationship between FPR AND TPR is immediate when FP increase the TP also increases as it asperses in the ROC curve the area under the curve is large therefor the result classified as good. And the accuracy of the curve is 0.84[13]

What we summarize from analyzing our data using logistic regression:

Since in our logistic regression model the accuracy=0.84 is high ,so we can consider the classification acceptable, although there is a small percentage of error 0.16 this error due to wrong classification data , there could be many reasons for wrong classification such as systemic error in the test for the outcome, or maybe the model is too simple to capture the patterns in data, data has too much noise etc.[16]

what we did in phase 3:

we focus to achieve our goals from phase 1 the first goal was:
"Try to find out the main reason of customers leaving Al-Rajhi bank and go to other banks."

So how we achieve it?

First, we classify our data to satisfied customers / not satisfied customers (high probability they will leave the bank, or they did)

For classification we first try to collect words that satisfied/ not satisfied customers used and classify them by using the words, but the result was not good since we have a lot of tweet like: "Al-Rajhi bank is the worst, *** bank is the best!" and then we try to use analysis tool called "**Mazajak**" which was suggested by teacher Bayan Al-Arifi, the tool helps us a lot and give excellent result compare to how we did it before.

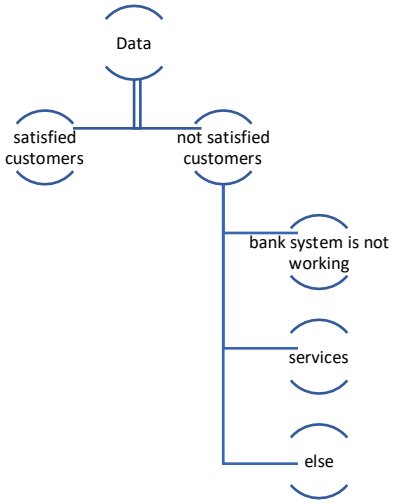
After that we drop satisfied customers (we did have 4111 tweets after drop 2470) more than half was classify as not satisfied customers and there no surprise since we was looking for them and it time to "find out the main reason of customers leaving Al-Rajhi bank and go to other banks" here we notice that customers used straight forward word so it was easy to us this time to do classification by collect words and then run the code to get the result.

We did make a new column name "reason" and classify the tweet to "not satisfied because services" , "not satisfied because bank system is not working" or "else" we chose these preictal problem because during the cleaning phase we look in our data deeply and we notice they did actually repeat a lot so we make a small test to see if we was right about chose these problems , the test:

2470 (number of not satisfied customers) □ drop the ones that classify as "else" □
1312 (number of not satisfied customers after drop)

That mean more than half their problems was related to services/ bank system is not working and that proof that these problems consider as main reason of customers leaving Al-Rajhi bank and go to other banks.

Classification summary:



After applying the classification on our data , it classifies our data to three types Positive , negative and natural tweets . But our question is **which one have more tweets and what are the reasons**?

We found that the negative tweets have the largest number of tweets, **2470 of 4111** in our data. Since achieving this result, we look for the main reasons that was behind it by applying our classification on the negative tweets only to classifying them into the main reasons **bank system is not working** and **customer service** and other reasons . and we found that our main reasons have more than the half of negative tweets it has **1312 of 2470**. For this reason we will explain this problems in detailed:

We find many tweets that represent that customers upsets of **bank system is not working** like ATMs not working and sometimes the ATMs hold the card without any reason. Also, sometimes the employees tell the customers that the system is not working just for they do not want to serve this customer .

For the second reason **customer service** we find many tweets that talk about the bad manners of the customer service employees. The customers also talk about the bad administration of the department of management after looked for the reasons of bad customer service we found in our data strike movement of the employees they asked for premium of there efforts.

The last phase of the project phase 4:

Achieving Our 4th goal from phase one was: "try to find out strengths that lead to develop and prosperity to Al-Rajhi bank".

From our point of view since there is a lot of problems on the service system its good for Al-Rajhi bank to improve their system by an effective application that can serve many people in an efficient way. So, when we search, we found that they really got an amazing application that raised customers satisfaction in the last years, this application enrolls Al-Rajhi bank with a strength point.

we suggest solutions for these two fundamental problems to reduce customer exit from Al-Rajhi bank that is:

For customer services:

- Training: provide a training for the customer services' employee.
- Recognition award: for the incredible delivering of the services.
- Monthly employee star: best employee performance in a month.
- Increasing number of employees: this will help to degrade the load of work.
- Financial reward: to encourage the employees.

For system is not working:

- Use secondary system when the basic system is not working.
- Regular checking of the system periodically.

After suggesting and recommending some solutions it is time to visualization:

What is visualization and why it is important?

Data visualization is the representation of data or information in a graph, chart, or other visual format. It communicates relationships of the data with images.[18]

Data Visualization makes the complex easier to understand. Also, it is a quick and easy way to convey concepts in a youniversal manner. Data visualization can help identifying areas that need attention or important.[19]

What visualization technique we used?

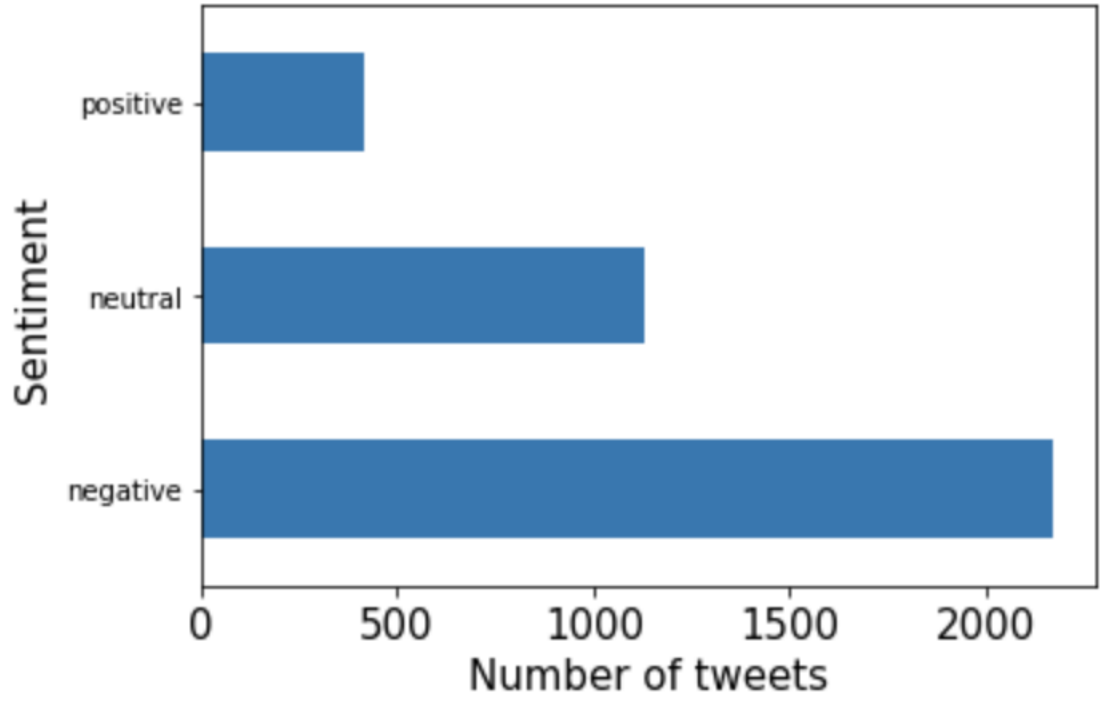
We used two different visualization techniques: Time series analysis chart to showing our result clearly of our classification depending on the years , Bar

diagram to showing our results and the difference between the results when we add or remove a specific values clearly .

Who the consumer of our visualization?

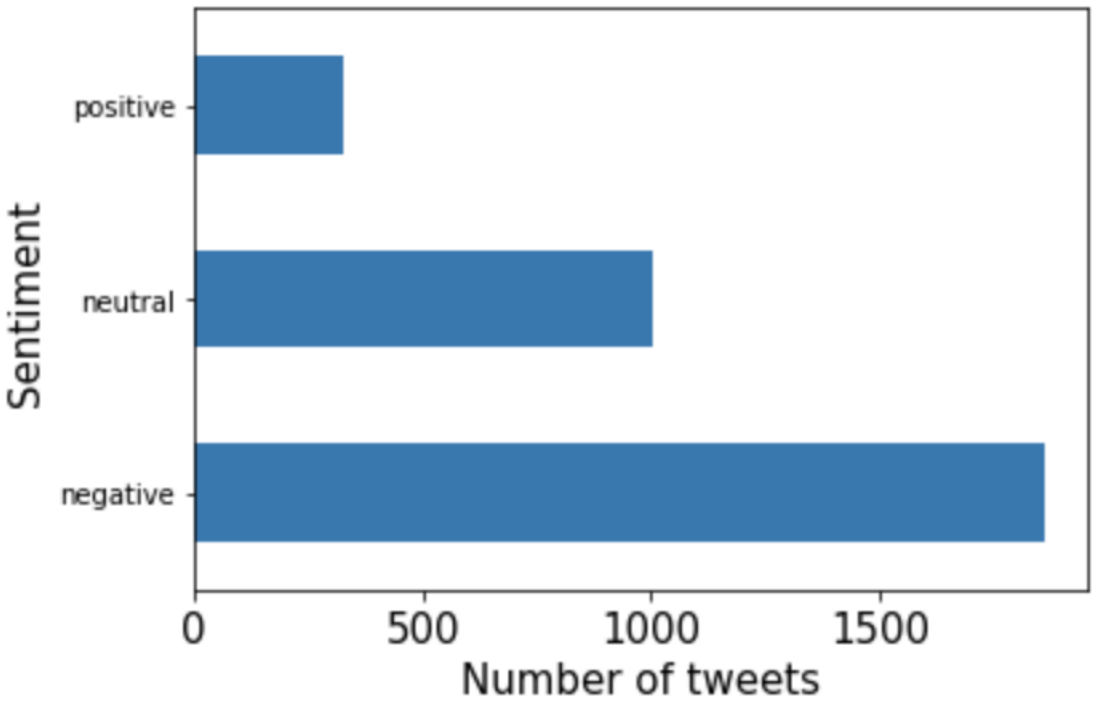
The main consumer of our visualization are customers, since it's all about considering their sentiments and trying to solve the problems facing them.

visualization and Explanation:



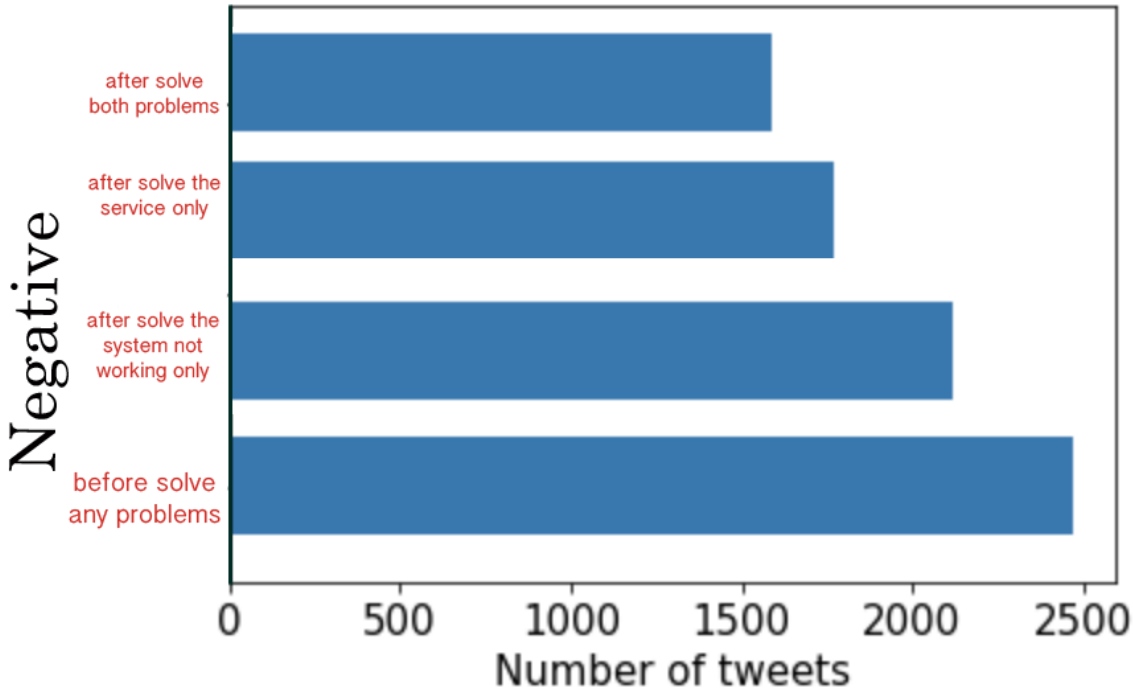
(1): This graph shows sentiment of customers after solve system is not working problem

This graph shows what the differences will be when Al-Rajhi bank solve one of their problems which is kind of important one "system is not working" problem, as we can see that the number of tweets differ when customers sentiments are negative, positive, and neutral. And it is clear that the number of tweets became less after solving the problem, this guides to customers satisfaction.



(2): This graph shows sentiment of customers after solve customer services problem

This diagram shows what the difference will make when Al Rajhi Bank actually tries to solve one of its main problems which is customer service
Apparently after solving the service issue, the negative tweets fell from about 2,500 to 1,800 with 700 difference, which accounted for 29% of all negative tweets, this mean negative tweet dropped 29% after Al Rajhi Bank solved this problem
In conclusion solving customer service issue by using previously defined suggestion will help the bank to satisfy 29% of overall 2470 clients, Approximately 716 clients that were not satisfied by customer service



(3): This graph shows negative comments from customers after the different events

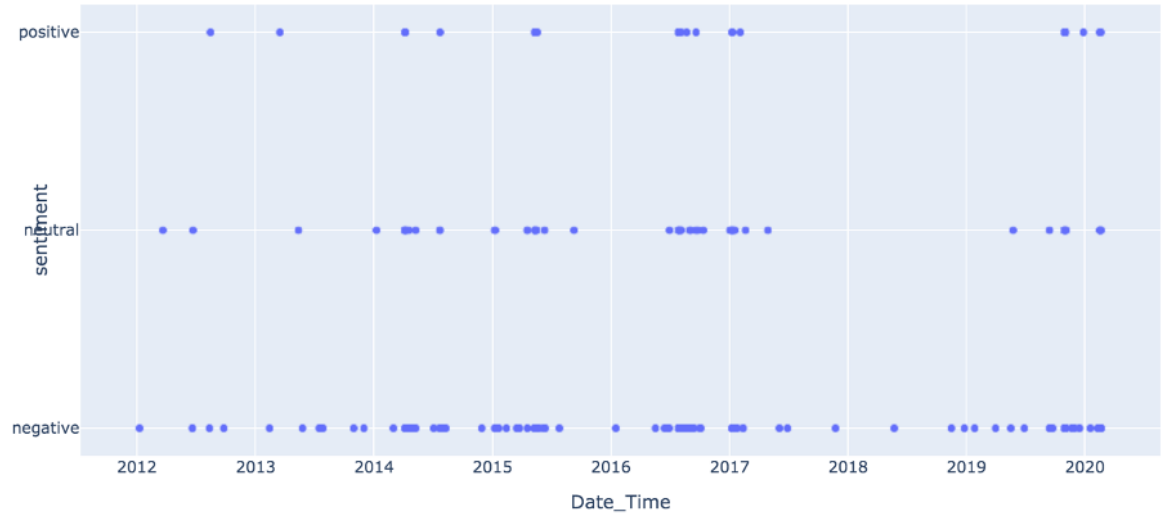
After talking about what will happen if “Al-Rajhi Bank “ solve "customer services" problem or solve "system is not working" problem but what will happen if they did solve none of the problems or they solve both problems?

of course, solving a big number of problems as you can give you the best result that why the shorter number of negative tweets was when they solve both the problems but solving both problems is hard to do at the same time, but what if they are able to solve one problem right now that give them the best result?

we thought the same that why we did this graph in the first place from the first look you can easily see the difference in a number of negative tweets when they only solve "customer services" problem and when they solvent "system is not working" problem it trends out that the biggest problem here was "customer services" problem! so, our recommendation for “Al-Rajhi Bank” to deal with the "customer services" problem first and then the "system is not working" problem.

For the "system is not working" problem yes, the number of negative tweets become less but it did not give a good result compared to the "customer services" problem and then we thought maybe because Al-Rajhi Bank acutely deals with

this problem and it is not considered as a problem anymore. we will talk about how they solve it in the graph (4)
But since our rustle shows a difference in the number of negative tweets we think they need to keep the focus on the "system is not working" problem since there some customers still complain about it.



(4): This graph shows the sentiment of customers in different period times. source code for the graph17

This diagram shows the time series of our classification in different period times so we will focus on the negative tweets or the customers against the bank. They are concentrated between 2014 and half of 2015 after looking for the gains and losses in the bank of these years we found that there is a loss in the profits on 2014 about -8.09% compared with the previous year 2013[20]. For the period between the half of 2017 and 2019 we see that there is a decreasing of the number of negative tweets or the customers against the bank. After looked for the reasons of this outcome we found that this is the period that was releasing the new version of al-Rajhi application that solve the problems on the previous version, So we can see that the customers are more comfortable with the application and they don't need to go to the bank any more[21] .after that we see there are increasing of the number of customers against the bank but we think that because of the new standards and rules nowadays and taxes .

list of all our development files

File name	files descriptions
extract tweet code.ipynb	The code that we used to extract tweets.
Data.csv	The file that contains all tweets that we extracted before and after the cleaning.
with the bank .csv	This file contains positive words that we discover from our quick look at tweets.
against the bank.csv	This file contains negative words that we discover from our quick look at tweets.
code.ipynb	This file contains the code we used to clean the data + apply our model
system.csv	This file contains words about system problem that we discover from our quick look at tweets.
services.csv	This file contains words about services problem that we discover from our quick look at tweets.

Library that we used:

GetOldTweets3:

GetOldTweets3 is an improvement fork of the original Jefferson Henrique's GetOldTweets-python. It fixes known issues and adds features such as counting retweets, searching over multiple user's accounts, etc. GetOldTweets3 supports only Python 3.

- Since Twitter Official API has the bother limitation of time constraints, we can't get older tweets than a week. So, we have used GetOldTweets3 library to extract older tweets and with no limitation.

Nltk(Natural Language Toolkit):

NLTK is a powerful Python package that provides a set of diverse natural languages algorithms. It is free, opensource, easy to use, large community, and well documented. NLTK consists of the most common algorithms such as

tokenizing, part-of-speech tagging, stemming, sentiment analysis, topic segmentation.

- We have used NLTK to help the computer to analysis, preprocess, and understand the written text the we extracted from twitter.

Pandas:

Pandas is a library created to help developers work with "labeled" and "relational" data intuitively. It's based on two main data structures: "Series" (one-dimensional, like a list of items) and "Data Frames" (two-dimensional, like a table with multiple columns). Pandas allows converting data structures to DataFrame objects, handling missing data, and adding/deleting columns from DataFrame, imputing missing files, and plotting data with histogram or plot box. It's a must-have for data wrangling, manipulation, and visualization.

- We have used panda for cleaning our data.

Numpy:

The library offers many handy features performing operations on n-arrays and matrices in Python. It helps to process arrays that store values of the same data type and makes performing math operations on arrays (and their vectorization) easier. In fact, the vectorization of mathematical operations on the NumPy array type increases performance and accelerates the execution time.

- We have used NumPy for cleaning our data by specifying some texts, such as tweet that involves Advertisements so we can delete it also we have used it to specify null value tweets for deleting it etc.

The tool that we used:

Tweepy : Tweepy is an open-sourced Python library to communicate with Twitter and access the Twitter API. It is great for simple automation and creating twitter bots. And we have used the tool to Get tweets from our timeline ,for analyzing them.

- We have used it to extract tweets from Twitter.

Mazajak: An Online Arabic Sentiment Analyzer

Sentiment analysis is one of the most useful natural language processing applications. There are many papers and systems addressing this task, but most of the work is focused on English.[7]

- We have used it to classification our data .

Challenges

in (phase one):

We faced a struggle during extracting tweets from twitter before our lab time. Why struggle? We find different sources, the old and the new ones, so we got confused and lost. then we decided to wait for the lab to clear our confusion.

in (phase two):

since it's our first time using python, we had a lot of difficulties more than phase 1. Yes, we do have the workshop code but sometimes we spend one hour or more just to understand what this variable for, what does this error means, and what case this type of error? and more.

We put for ourselves a schedule to meet and work together in the project but since suddenly we start taking our lectures online because coronavirus a lot of lectures change their original time so we reschedule our plan for the project to be able to do both attend the lectures and keep the work in the project at the same time, and we hope it did not affect the quality of work that we provide.

REFERENCES

[1] https://pypi.org/project/GetOldTweets3/
[2] https://gist.github.com/vickyqian/f70e9ab3910c7c290d9d715491cde44c
[3]https://drive.google.com/drive/folders/1ANnepiUMumMoUZYZ5kFNEOrQOsr3GBuo This reference was given to us in the workshop .
[4] https://bigdata-madesimple.com/top-20-python-libraries-for-data-science/
[5] https://www.dataquest.io/blog/15-python-libraries-for-data-science/
[6] http://www.statulator.com/blog/descriptive-analysis-take-it-easy/
[7] http://mazajak.inf.ed.ac.uk:8000/
[8] http://logisticregressionanalysis.com/33-when-to-use-logistic-regression/
[9] https://www.youtube.com/watch?v=TtIjAiSojFE
[10] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/
[11]https://datascience.stackexchange.com/questions/64441/how-to-interpret-classification-report-of-scikit-learn
[12] https://muthu.co/understanding-the-classification-report-in-sklearn/
[13] https://www.youtube.com/watch?v=A7G30xN_2n4
[14]https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/One_ROC_Curve_and_Cutoff_Analysis.pdf
[15]https://towardsdatascience.com/demystifying-confusion-matrix-confusion-9e82201592fd--https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
[16]https://stats.stackexchange.com/questions/281288/what-makes-a-classifier-misclassify-data
[17] https://plotly.com/python/line-and-scatter/
[18] https://www.import.io/post/what-is-data-visualization/
[19] https://www.sas.com/en_sa/insights/big-data/data-visualization.html
[20]http://argaamplus.s3.amazonaws.com/85c13f2f-e4dd-428c-8ebd-7542901c1ebb.pdf
[21]https://twitter.com/alrajhibank/status/931507314217111552?lang=ar