Effectiveness of Telehealth in Saudi Arabia During COVID-19 Pandemic.

Probed by

Ghadeer, Rawabi, Fatimah, Raghad, Afrah.

Background

As the novel coronavirus disease (Covid-19) spreads across Saudi Arabia, the need for innovative measures to provide high-quality patient care and manage the disease's spread becomes more pressing. The use of telehealth has steadily increased, and it has become a viable modality of patient care. As a result, early adopters try to use telehealth to provide high-quality care, and patient satisfaction is an important indicator of how well the telehealth modality met patient expectations.

COVID19 spreads swiftly, and each infected individual can infect multiple people, resulting in an exponential and extremely high rate of spread. During the outbreak, the Saudi Ministry of Health has urged individuals to use smartphone apps instead of going to primary care facilities. Telehealth visits grew from 102.4 to 801.6 per day between March 2nd and April 14th, 2020. Over 80% of Medicare beneficiaries reported that their usual providers offered telehealth during the COVID-19 pandemic. The goal is to discuss the current status of the use of remote health services applications during the emerging Corona pandemic in the Kingdom, in addition to the effectiveness of these applications in supporting public health measures, and to know the opinions of users of applications such as the Tawakkalna and Sehaty applications. In this

project, we focus on the applications most used based on the survey (Tawakkalna and Sehaty).
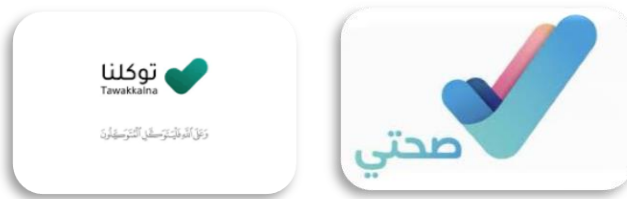

Figure1: Applications for telehealth.

Questions/Needs

- What are the most used applications?

- What are the beneficiaries' comments about the applications?

- Which model will be better to cluster the comments?

- How can we test the predicted model?

- What are the users' reactions to the applications?

- We need to extract the most topics using the wordcloud tool.

Methodology

Dataset acquisition protocol

The data for this project will be obtained from [QASEEM UNIVERSITY-Scientific Research Deanship]. Data will contain about 1040 rows. Tool of data collection: it includes three main parts which the first part included 4 items regarding Socio Demographic data; The second part included 9 items about the knowledge of current status of telehealth; The third part contains 13 items for the effectiveness of health care of telehealth. A user reviews column has also been added from Google Play for the most used applications.
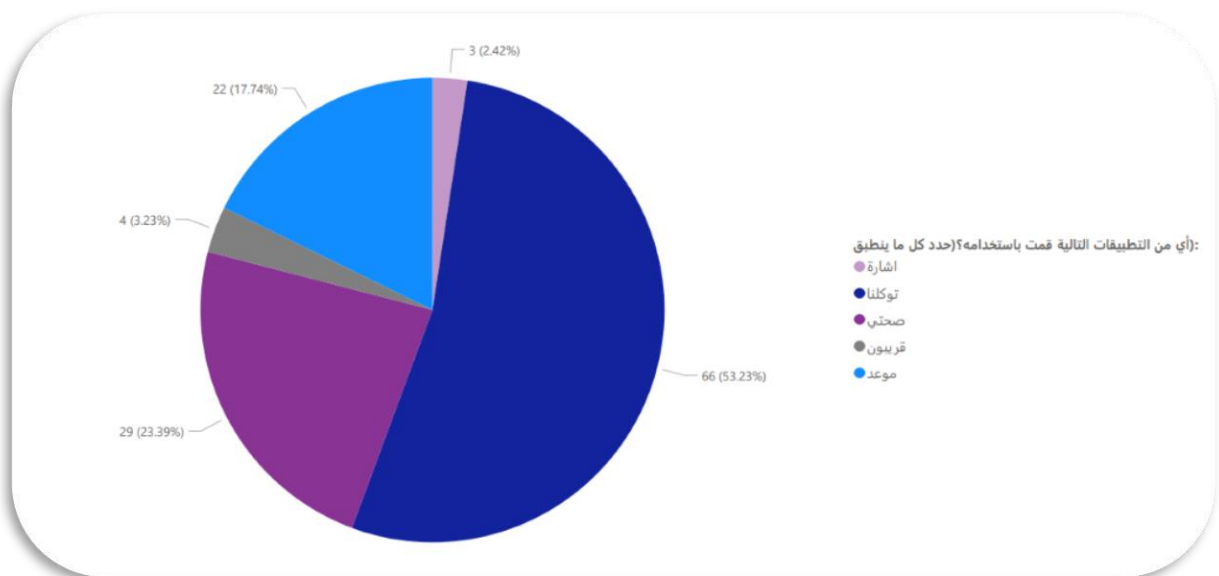
Figure2:Range of applications used for th services.

In the comparison above in the pie chart, it is clear that the two most important applications are Tawakkalna and Sehaty.

Pre-processing

Data preprocessing involves transforming raw data to well-formed data sets so that data mining analytics can be applied. Raw data is often incomplete and has inconsistent formatting. In this step, we removed the null values, dropped any duplicates, and dropped unnecessary columns. We also needed to do more processing to format the input text. We removed(numbers, any white spaces, punctuations, newline characters, emojis....etc)

- To summarize the overall opinions

Let's take a look at the full dataset to see what the opinions are in Tawakkalna and Sehaty .
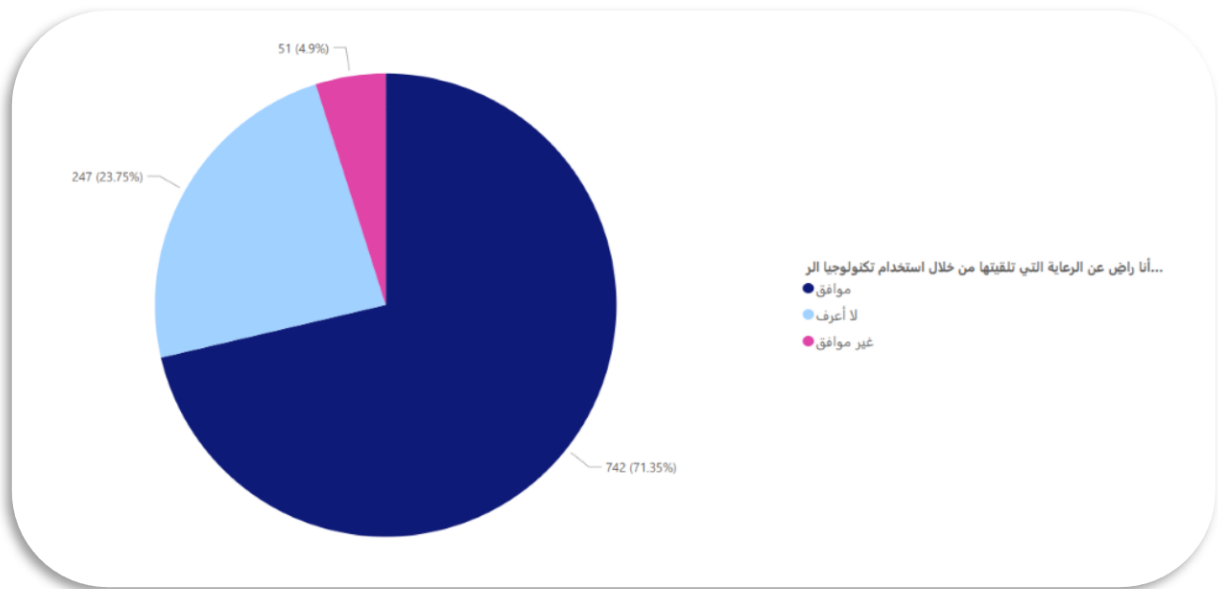
51 (4.9%)

247 (23.75%)

...أنا راض عن الرعاية التي تلقيتها من خلال استخدام تكنولوجيا الر

● موافق
● لا أعرف
● غير موافق

742 (71.35%)

Figure 3:Tawakkalna opinions.

92 (11.22%)

Column2
● موافق
● غير موافق
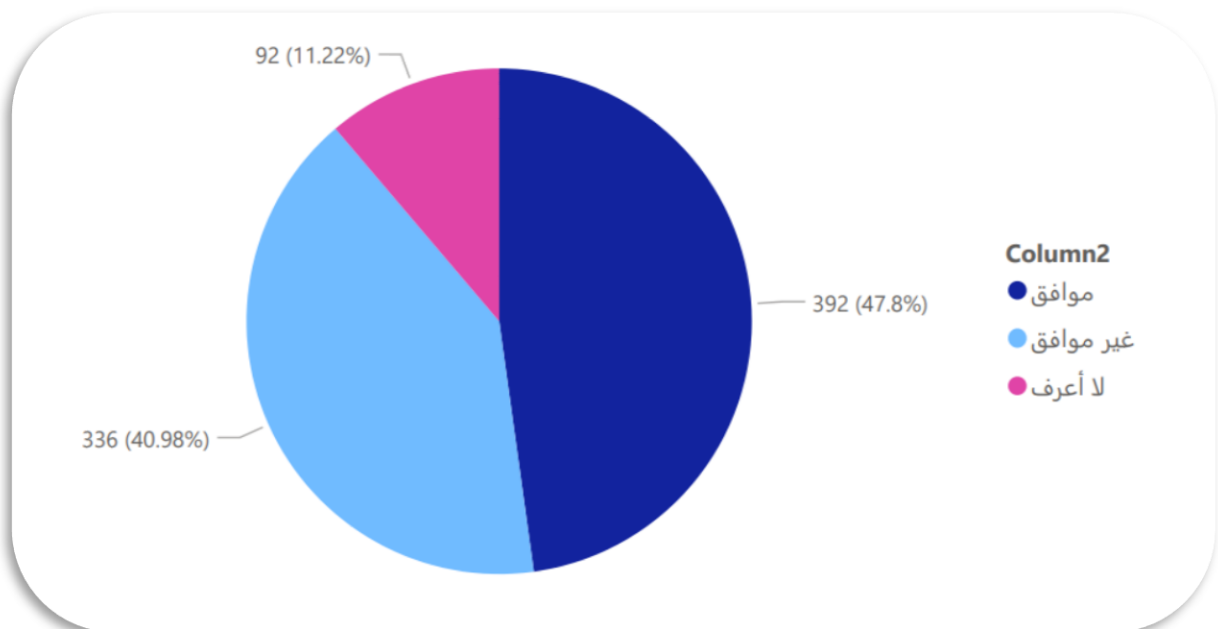● لا أعرف

392 (47.8%)

336 (40.98%)

Figure 4:Sehaty opinions.

The two applications have positive reactions, but the Sehaty application has the most negative reactions, and after looking at the comments of the beneficiaries, we noticed one of the updates caused a problem in locating

the patient. In addition, In the comments of the beneficiaries of all applications, most of the negative responses came from many updates. To see more clearly, we used the machine learning application NLP.

Algorithm

1. Count Vectorizers

Countvectorizer is a way to convert a given set of strings into a frequency representation. This is a very simple case of NLP where you get a tagged text data set and then using it you have to predict the tag of another text data. The texts can be converted into count frequency using the CountVectorizer function of the sklearn library.
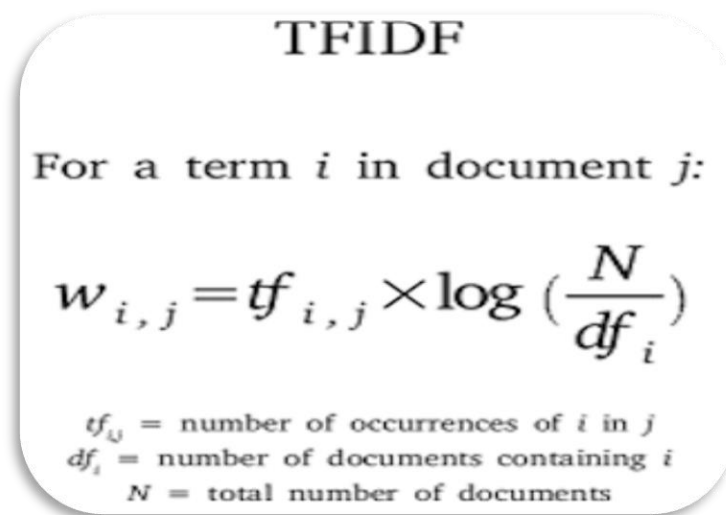
Their frequencies are calculated as 0 while other words are present Once hence their frequencies are equal to 1.This is, in a nutshell, how we use the Countvectorizer. Counting vectors can be helpful in understanding the type of text by the frequency of words in it. But its major disadvantages are:

- Its inability in identifying more important and less important words for analysis.
- It will just consider words that are abundant in a corpus as the most statistically significant word.
- It also doesn't identify the relationships between words such as linguistic similarity between words.

2. TF-IDF

TF-IDF means Term Frequency - Inverse Document Frequency. This is a statistic that is based on the frequency of a word in the corpus but it also provides a numerical representation of how important a word is for statistical analysis.

TF-IDF is better than Countvectorizer because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words. We can then remove the words that are less important for analysis, hence making the model building less complex by reducing the input dimensions. This is how tf-idf is calculated:

TFIDF

For a term i in document j:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Figure 5:TFIDF calculating.

The term "tf" is basically the count of a word in a sentence. The term "df" is called document frequency which means in how many documents the word "subfield" is present within the corpus. Here is the final TF-IDF matrix for a corpus.

Topic Modeling

Topic modeling is the process of identifying topics in a set of documents. This can be useful for search engines, customer service automation, and any other instance where knowing the topics of documents is important. There are multiple methods of going about doing this.

- **LDA**

(Latent Dirichlet Allocation) In LDA, latent indicates the hidden topics present in the data, then Dirichlet is a form of distribution. Dirichlet distribution is different from the normal distribution. When ML algorithms are to be

applied, the data has to be normally distributed or follow Gaussian distribution. The normal distribution represents the data in real numbers format whereas the Dirichlet distribution represents the data such that the plotted data sums up to 1. It can also be said as Dirichlet distribution is a probability distri.

Figure 6:Topic Modeling LDA.

- **NMF**

Non-Negative Matrix Factorization (NMF)
Non-Negative Matrix Factorization is a statistical method that helps us to reduce the dimension of the input corpora or corpora. Internally, it uses the factor analysis method to give comparatively less weightage to the words that are having less coherence.
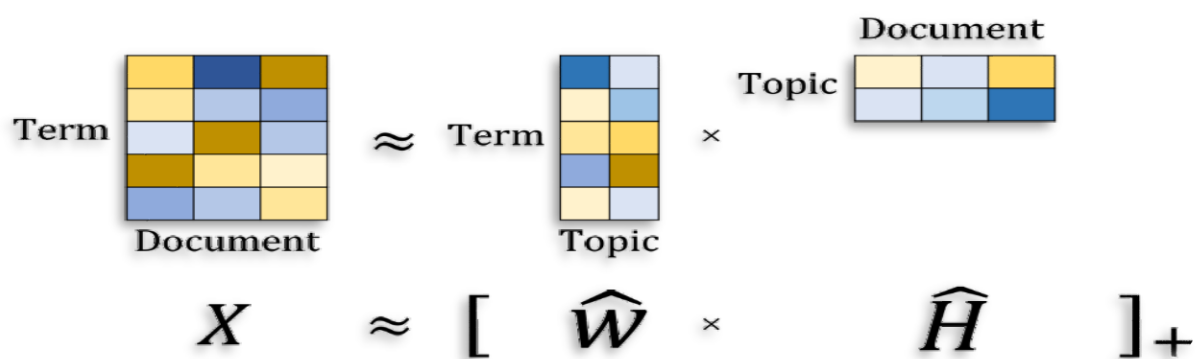

Figure 7:Topic Modeling NMF.

dominant_topic	خدمات	تعليق	برنامج	الوطن	سهل الاستخدام	تحديث	
5	0.191	0.000	0.000	0.000	0.000	0.028	Doc0
0	0.000	0.001	0.014	0.012	0.007	0.060	Doc1
3	0.000	0.000	0.293	0.000	0.000	0.000	Doc2
0	0.002	0.000	0.019	0.000	0.000	0.087	Doc3
2	0.000	0.000	0.006	0.224	0.000	0.004	Doc4
...	...	...	...	...	...	...	...
5	0.045	0.018	0.000	0.010	0.000	0.000	Doc723
5	0.079	0.001	0.006	0.001	0.000	0.000	Doc724
3	0.003	0.000	0.164	0.002	0.000	0.034	Doc725
0	0.000	0.000	0.000	0.040	0.001	0.073	Doc726
5	0.143	0.059	0.000	0.000	0.057	0.000	Doc727

Table 1: NMF.Tawakkalna_document topics.

dominant_topic	انترنت	مساعدة	تحديد الموقع	صورة	ممتاز	إزعاج	
0	0.005	0.000	0.001	0.000	0.001	0.008	Doc0
2	0.000	0.000	0.099	0.186	0.159	0.000	Doc1
1	0.045	0.003	0.046	0.000	0.059	0.005	Doc2
2	0.061	0.000	0.000	0.092	0.000	0.079	Doc3
2	0.000	0.024	0.000	0.192	0.000	0.007	Doc4
...	...	...	...	...	...	...	...
4	0.000	0.547	0.000	0.000	0.000	0.000	Doc569
2	0.000	0.008	0.001	0.103	0.000	0.000	Doc570
5	0.193	0.000	0.000	0.000	0.149	0.000	Doc571
2	0.000	0.022	0.019	0.023	0.014	0.000	Doc572
0	0.000	0.000	0.000	0.063	0.000	0.218	Doc573

Table 2: NMF.Sehaty_document topics.

- **SVD**

We would clearly expect that the words that appear most frequently in one topic would appear less frequently in the other- otherwise that word wouldn't make a good choice to separate out the two topics. Therefore, we expect the topics to be orthogonal. The SVD algorithm factorizes a matrix into one matrix with orthogonal columns and one with orthogonal

rows (along with a diagonal matrix, which contains the relative importance of each factor).
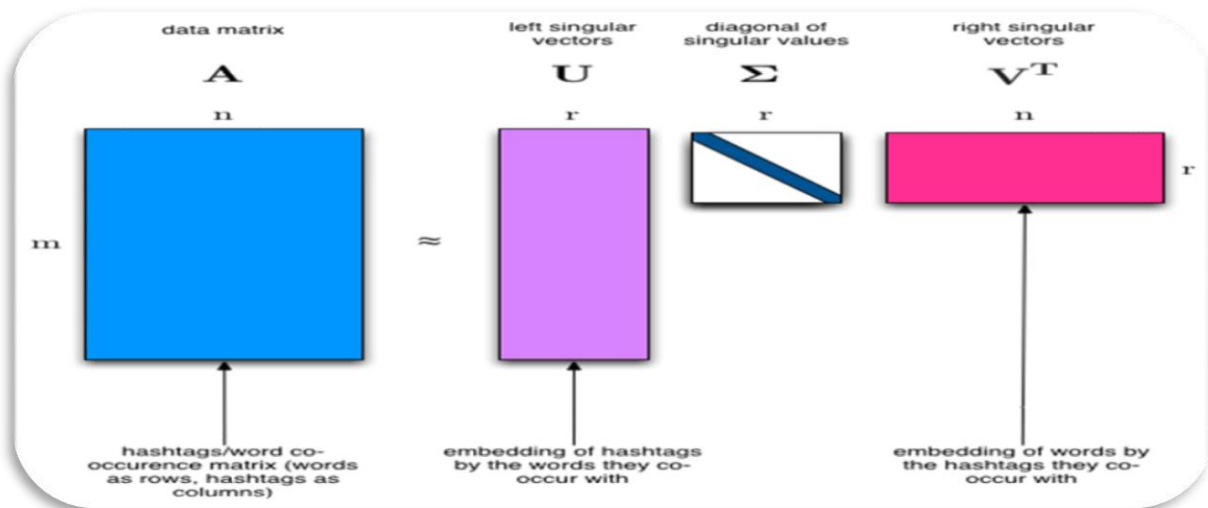

Figure 8:Topic Modeling SVD.

SVD is an exact decomposition, since the matrices it creates are big enough to fully cover the original matrix. SVD is extremely widely used in linear algebra, and specifically in data science.

- A plot of the number of most common opinions in Tawakkalna application using the plot method:

- A plot of the number of most common opinions in the Sehaty application using the plot method:

VISUALIZATION

- Visualize all Words count in Tawakkalna

- Visualize all Words count in Sehaty

Tools

Technologies :Jupyter Notebook , Python.

Libraries:nltk, spacy, re, genism, fasttext,wordcloud,farasa,sklearn,etc...