Using Social Media Data To Forecast Telecom Companies Revenues with Machine Learning

Eng.Yazan Aswad[1]
Faculty of Information Technology and Communications,
Master in Web Sciences, Syrian Virtual University
Damascus – Syria

Dr. Aghiad Kh. Alkatan[2]
[2]Assistant Professor,
Computer and Automation Engineering
Alhamak College - Damascus University
Syria

***Abstract*— Traditional models for predicting future sales of a product or service are based on previous, not updated data, resulting in unsatisfactory and inaccurate forecasting results, meaning that the data used as inputs to the forecasting process is stable and not dynamic during the forecasting process.**

The research aims to leverage social media data by extracting features from Facebook platform (features are reactions to posts) and using them as input to the automated forecasting system to try to predict corporate revenues.

Machine learning algorithms have been trained to predict returns according to pre-stored data and can be updated on demand, which means that the proposed forecasting system will work in a dynamic environment.

The following algorithms were used to predict the profitability of new services and the one with the highest accuracy was selected: (Random Forest, DT, Gradient Boosting, K nearest neighbors, NB).

The results showed that Random Forest algorithm is the one with the best accuracy, with an accuracy of 67%, and a slight correlation was observed between the interactions on the target post and the profitability of the service within the post.

Keywords*— Machine learning, social media marketing, classification (Random Forest, DT, K nearest neighbors, NB), sentiment analysis, Facebook graph API.

I. INTRODUCTION

Predicting the profits of service companies has been studied for years with different methods and tools, studying the past to understand the present and anticipate the future. On the other hand, companies have started trying to reach the target audience via social media, where platforms such as Facebook, Twitter and Instagram allow users to express their thoughts and opinions through the buttons on the platform.

Therefore, many companies and business strategists consider social media an important area and are constantly trying to figure out different ways to increase their profitability using the data extracted from it [1], as the importance of social media in business grows very rapidly as the number of people who join and use social media sites regularly increases.

As button is a part of interactions available to users, it can be used directly to identify users' feelings and emotions, and it can help determine satisfaction with the post in social media [2-11].

Social media marketing is not a new trend, and is a way to help companies easily reach targeted customers. Social media marketing can be defined simply as using social media channels to promote a company and its products [3-4].

Some researchers consider social media marketing to be one of the basic components of marketing, and it is a tool for analyzing and knowing customers' behavior and an opportunity to learn about their desires to analyze and try to achieve them, and this is done by establishing relationships with customers and providing them with the opportunity to express their opinions in the development of the company's products[5-12].

Social media can provide unlimited information about customers without human intervention, this is an advantage over other forms of communication because the amount of information that can be provided is much larger than any other form of communication, in addition, a company that adopts social media marketing can allow customers to design products and services that meet their specific requirements, for example"[13-14].

II. RELATED WORK

Research [6] "Using social media and machine learning to predict financial performance of a company", stated that traditional models to predict future sales of a product or service depend on past and not up-to-date data, which provide unsatisfactory and inaccurate forecast results.

Another research predicted the profitability of a film using social media data via film-related input parameters such as actor, director, writer, etc. [7].

This study includes three key components (data collection, advance data processing, Data Mining), and this project follows a mathematical model of the k-NN algorithm used in the classification and results obtained from the project were observed close to IMDB classifications.

On the other hand, research [8] took advantage of internet data such as the number of online product reviews, the number of comments, the number of answers to questions asked and the evaluation of customer review, and then these data were entered into the forecasting system in order to predict product sales in an online environment.

Chavan et. in [9], handled profit forecasting in a different way through the training of machine learning algorithms (Linear Regressor, K-Neighbors Regressor, Xgboost Regressor, Random Forest), with data collected from previous grocery store sales. The data consisted of variables such as item weight, item fat content, item shape, item type, and random Forest's algorithm has the highest accuracy with 93%.

In research "A machine learning-based approach to enhancing social media marketing" [10], researchers examined social media data analytics using machine-learning tools, this approach used the WEKA environment to develop a social media marketing strategy.

The proposed model (ML-SMM) includes the concepts of social media marketing and machine learning, and integrates the WEKA machine-learning tool to predict online consumer behaviour to ensure effective marketing.

Based on these previous studies, it was noted that the data used are static non-dynamic data and the current research will try to solve this problem by relying on constant and variable data.

The research is based on collecting data from the company's posts via Facebook platform, training machine-learning algorithms on this data, comparing them and selecting an algorithm with the highest accuracy, to try to predict the profitability of a new service through interactions on the post of this service via Facebook.

III. PROPOSED MODEL

(Figure .1) depicts our research methodology, the first stage is data collection (numerical data and text data), the numerical data are the interactions which users interact on company's posts, in addition to the profit feature which expresses the profitability of services advertised through the company's posts, and text data are number of comments (about 4500 comments).

Comments classification system will be trained (using NLP techniques) on text data, this system will determine the percentage of positive comments within the posts to complete the system database.

System inputs are interactions (likes, haha...) + comments positivity percentage + profitability of services, and system output is a numerical value (one to six), where value 6 expresses higher profitability.

The next stage is to train prediction algorithms on the system database and compare them to determine the best algorithm to be able to correlate interactions with the profitability of the service in the post.

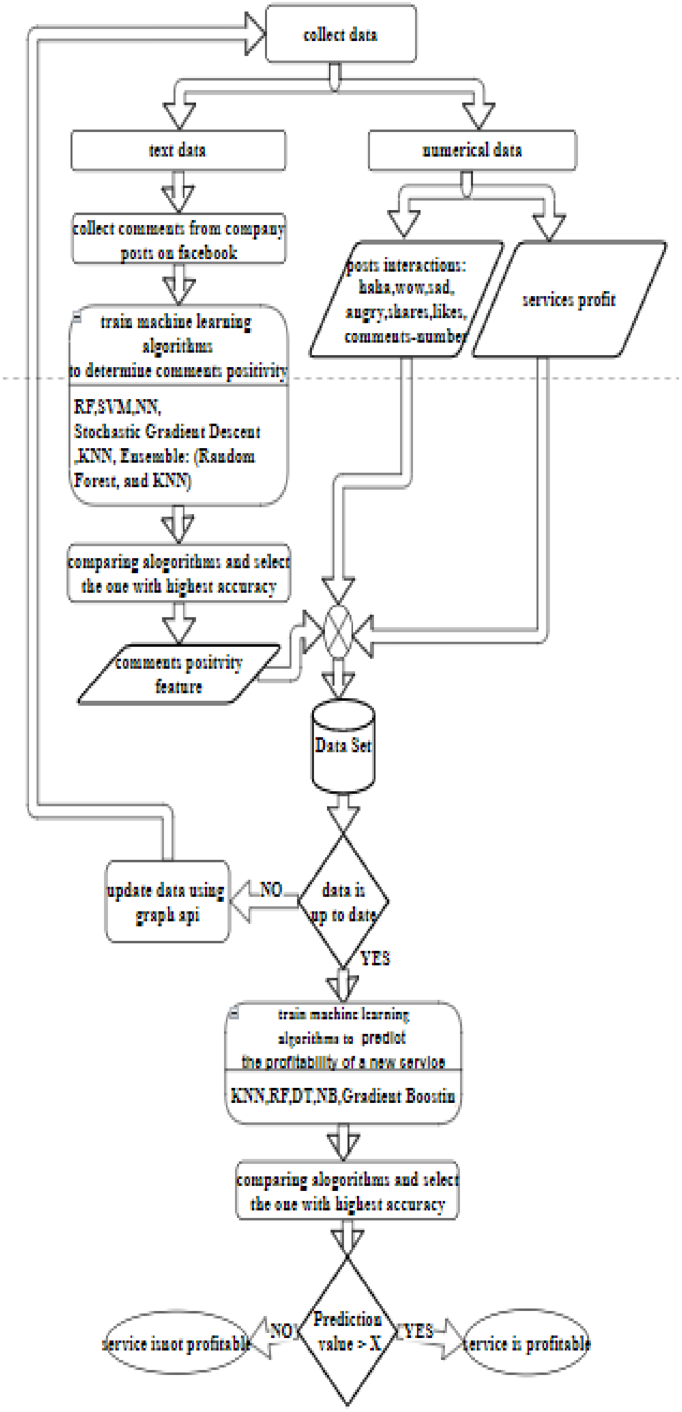

collect data
text data
numerical data
collect comments from company posts on facebook
posts interactions: haha,wow,sad, angry,shares,likes, comments-number
services profit
train machine learning algorithms to determine comments positivity
RF,SVM,NN, Stochastic Gradient Descent ,KNN, Ensemble: (Random Forest, and KNN)
comparing alogorithms and select the one with highest accuracy
comments positivity feature
Data Set
update data using graph api
NO
data is up to date
YES
train machine learning algorithms to predict the profitability of a new service
KNN,RF,DT,NB,Gradient Boostin
comparing alogorithms and select the one with highest accuracy
service isnot profitable
NO
Prediction value > X
YES
service is profitable

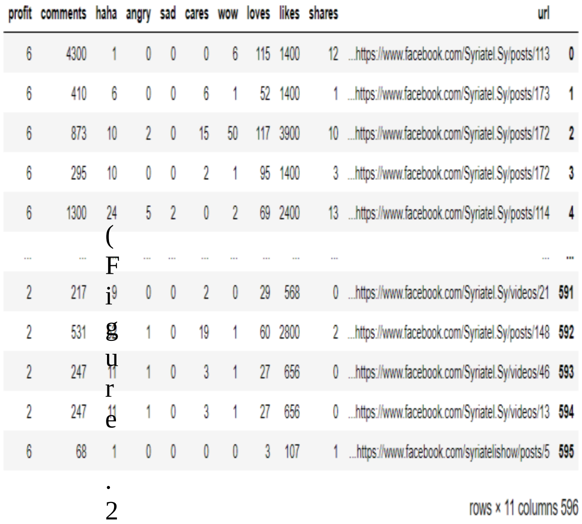
Figure 1. The proposed framework for predicting new service profit

IV. IMPLEMENTATION

A. COLLECT DATA

a. NUMERICAL DATA COLLECTION

The data will be extracted from each post published by the company and contains a service, product or offer offered by the company for a period from 2021 to 2013, and data collected via Facebook graph api.



profit	comments	haha	angry	sad	cares	wow	loves	likes	shares	url	
6	4300	1	0	0	0	6	115	1400	12	...https://www.facebook.com/Syriatel.Sy/posts/113	0
6	410	6	0	0	6	1	52	1400	1	...https://www.facebook.com/Syriatel.Sy/posts/173	1
6	873	10	2	0	15	50	117	3900	10	...https://www.facebook.com/Syriatel.Sy/posts/172	2
6	295	10	0	0	2	1	95	1400	3	...https://www.facebook.com/Syriatel.Sy/posts/172	3
6	1300	24	5	2	0	2	69	2400	13	...https://www.facebook.com/Syriatel.Sy/posts/114	4
...	...		...	...	...	...	...	...	...	...	...
2	217	9	0	0	2	0	29	568	0	...https://www.facebook.com/Syriatel.Sy/videos/21	591
2	531		1	0	19	1	60	2800	2	...https://www.facebook.com/Syriatel.Sy/posts/148	592
2	247	11	1	0	3	1	27	656	0	...https://www.facebook.com/Syriatel.Sy/videos/46	593
2	247	11	1	0	3	1	27	656	0	...https://www.facebook.com/Syriatel.Sy/videos/13	594
6	68	1	0	0	0	0	3	107	1	...https://www.facebook.com/syriatelishow/posts/5	595

rows × 11 columns 596

(Figure.2)

Figure 2. Numerical data

b. TEXT DATA COLLECTION

Text data are the texts (comments) written by users on Syriatel's page posts.

The textual data are 4,550 comments, compiled in one file, and then these comments were classified into three sections [positive (1), negative (0) and neutral (2)], and this was done manually on the 4550 comments where each comment was classified into one of the sections, and then comments processed through several techniques (Text pre-processing).

Classification algorithms were trained on these comments to select the best model and build a model that can determine the positive percentage of comments written on posts.

i. COMMENTS CLASSIFICATION

The goal of classification is to predict accurately the target class for each case in the data, to achieve this goal we perceptually split comments dataset into two disjoint sets 80% training set and 20% test set, training set used to build the model and test set used to validate it.

There are many types of classification algorithms, and we will go to use (Random Forest, Support Vector Machine, Stochastic Gradient Descent,

K nearest neighbors, and Ensemble: Random Forest, and KNN).

ii. COMMENTS CLASSIFICATION EXPIREMENTS

Comments classification algorithms will be compared across different metrics to choose the most accurate algorithm:

Accuracy: The basic metric used to evaluate the model is often accuracy, describing the number of correct predictions on all predictions.

Recall: is a measure of the number of positive cases correctly predicted by the classifier, for all positive cases in the data.

F1-Score: A measure that combines precision and recall, generally described as the consensual average of both.

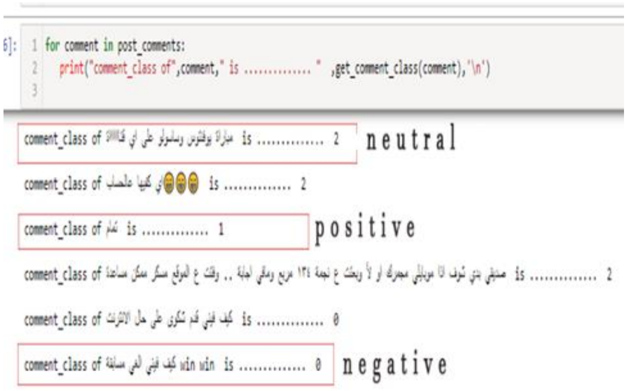
Training results as shown in (table .1).

Table 1. Text classification algorithms training results.

Model	Accuracy	F1-Score	Recall
Random Forest	85 %	81.7%	82.3%
Support Vector Machine	83.1%	83.2%	81.2%
Stochastic Gradient Descent	65.7 %	65.5%	62.7%
K nearest neighbors	75.4 %	73.4%	71.4%
Ensemble: Random Forest, and KNN)	83.8 %	83.7%	81.7%

As shown in (table .1), Random Forest is the best algorithm to classify comments.

(Figure .3) shows testing the model on three different sentences.



```
for comment in post_comments:
   print("comment_class of",comment," is ..............  " ,get_comment_class(comment),'\n')
```

```
comment_class of مباراة يوفنتوس وميلانو على اي قناة  is .............. 2

comment_class of ي كيا ملحساب😁😁😁  is .............. 2

comment_class of تمام  is .............. 1

comment_class of صديقي بدي شوف انا موبايلي محروك او لا وبحث ع نجمة ١٣٤ مربع ومافي اجابة .. وقت ع الموقع مسكر ممكن مساعدة  is .............. 2

comment_class of كيف فيني قدم شكوى على حل الانترنت  is .............. 0

comment_class of كيف فيني الغي مسابقة win win  is .............. 0
```

Figure 3. Examples of classifying different comments

Let's assume that the 3 sentences in (figure .4) are comments on a post, the system will return a value of 0.33 as the proportion of positive comments from the number of total comments.

B. NEW SERVICE PROFIT

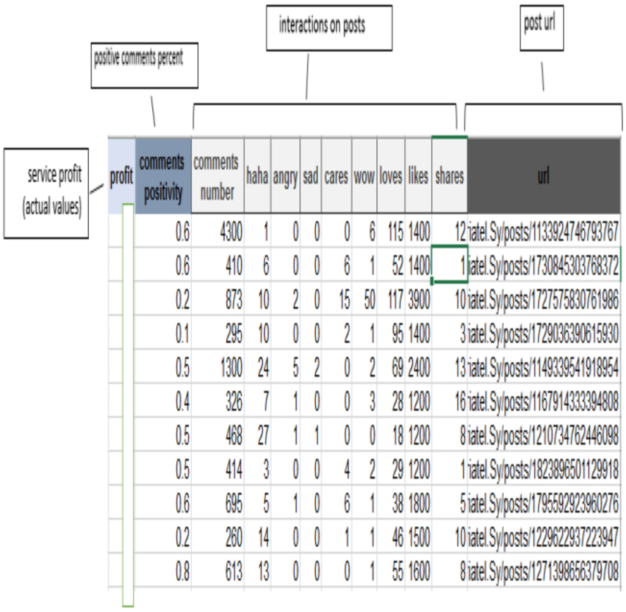
Initially, the numerical data has been combined with the percentage of positive comments within each post.

All features of the data set are presented in (Figure .4).

By using machine-learning algorithms, we try to implement a predictor model for the Telecom Company.

Now we have a data set, and by pre-processing and feature selection, we divide the data set for training and testing. For these algorithms, we have made some feature engineering to have more efficient and accurate results. (Here we divided data into 70% - training, 30%-testing). We used five algorithms to know which will provide us with results that are more accurate: (Random Forest, DT, Gradient Boosting, K nearest neighbors, NB).

We used the same dataset to train all models and tested it


interactions on posts
post url
positive comments percent
service profit (actual values)

profit	comments positivity	comments number	haha	angry	sad	cares	wow	loves	likes	shares	url
	0.6	4300	1	0	0	0	6	115	1400	12	iatel.Sy/posts/113392474679376
	0.6	410	6	0	0	6	1	52	1400	1	iatel.Sy/posts/173084530376837
	0.2	873	10	2	0	15	50	117	3900	10	iatel.Sy/posts/172757583076198
	0.1	295	10	0	0	2	1	95	1400	3	iatel.Sy/posts/172903639061593
	0.5	1300	24	5	2	0	2	69	2400	13	iatel.Sy/posts/114933954191895
	0.4	326	7	1	0	0	3	28	1200	16	iatel.Sy/posts/116791433339480
	0.5	468	27	1	1	0	0	18	1200	8	iatel.Sy/posts/121073476244609
	0.5	414	3	0	0	4	2	29	1200	1	iatel.Sy/posts/182389650112991
	0.6	695	5	1	0	6	1	38	1800	5	iatel.Sy/posts/179559292396027
	0.2	260	14	0	0	1	1	46	1500	10	iatel.Sy/posts/122962293722394
	0.8	613	13	0	0	0	1	55	1600	8	iatel.Sy/posts/127139865637970

Figure 4. Full Data Set.

V. EXPERIMENTAL RESULTS

We performed several experiments on the proposed profit prediction model using machine-learning algorithms on the dataset

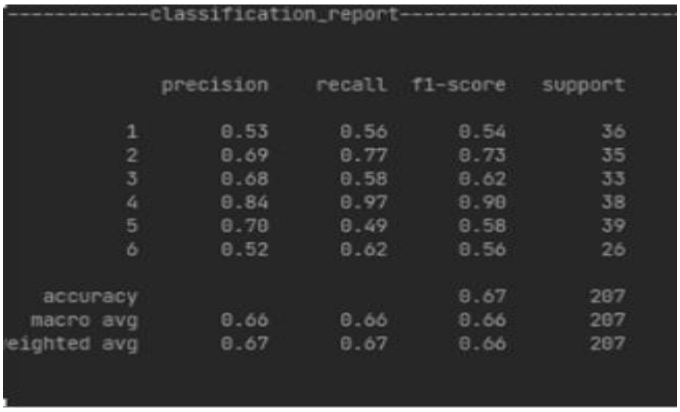
The result obtained from the algorithms are shown in (table .2), where we can observe the accuracy obtained by using the algorithms.

Table 2. Predicting profit s algorithms training results

Model	Accuracy	F1-socre	Recall	roc_ auc_ score
Random Forest	67 %	66%	66%	0.75
DT	48 %	48%	47%	0.50
Gradient Boosting	63 %	62%	63%	0.60
K nearest neighbors	59 %	58%	59%	0.55
NB	35 %	34%	37%	0.35

Random Forest (RF) is a useful algorithm that suite for classification and can handle nonlinear data very efficiently. RF produced better results and better accuracy and performance compared to the other techniques.

We can observe the results obtained when using the Random Forest technique (Figure.5) and NB (Figure.6).



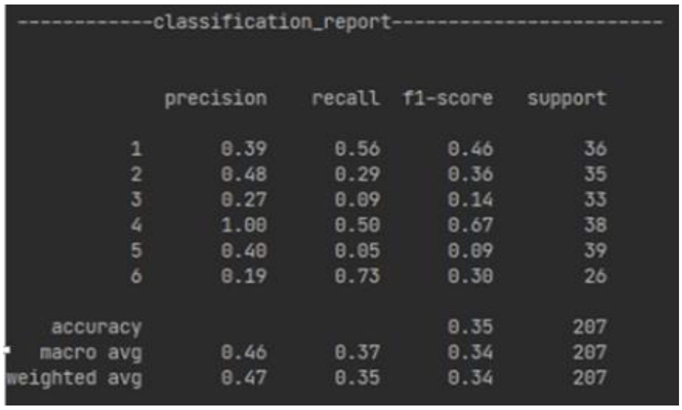
```
------------classification_report----------------------

              precision    recall  f1-score   support

           1       0.53      0.56      0.54        36
           2       0.69      0.77      0.73        35
           3       0.68      0.58      0.62        33
           4       0.84      0.97      0.90        38
           5       0.70      0.49      0.58        39
           6       0.52      0.62      0.56        26

    accuracy                           0.67       207
   macro avg       0.66      0.66      0.66       207
eighted avg       0.67      0.67      0.66       207
```

Figure 5. Classification report of Random Forest.



```
------------classification_report----------------------

              precision    recall  f1-score   support

           1       0.39      0.56      0.46        36
           2       0.48      0.29      0.36        35
           3       0.27      0.09      0.14        33
           4       1.00      0.50      0.67        38
           5       0.40      0.05      0.09        39
           6       0.19      0.73      0.30        26

    accuracy                           0.35       207
   macro avg       0.46      0.37      0.34       207
weighted avg       0.47      0.35      0.34       207
```

Figure 6. Classification report of NB.

Real values can be compared with the predicted values to evaluate the performance of the algorithm's, (figure .7) and (figure .8) show the predicted value and real value of random forest algorithm and Decision Tree algorithm for 22 samples.

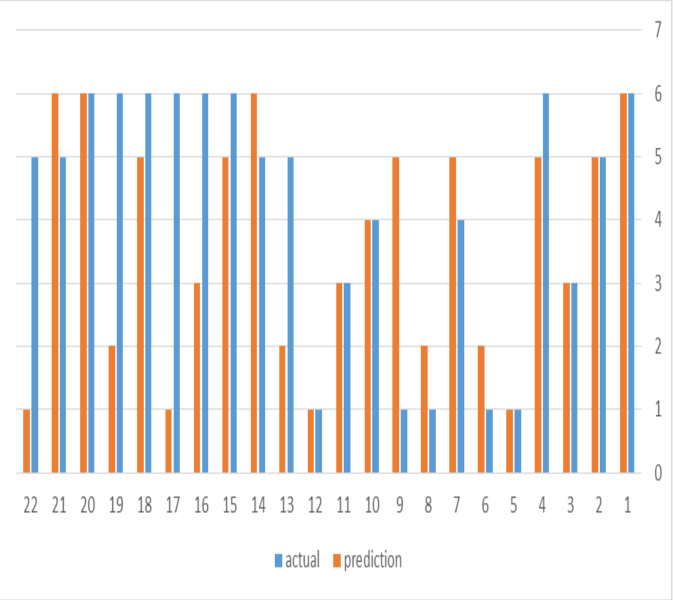


Figure 7. Comparing predicted value and real value of random forest.

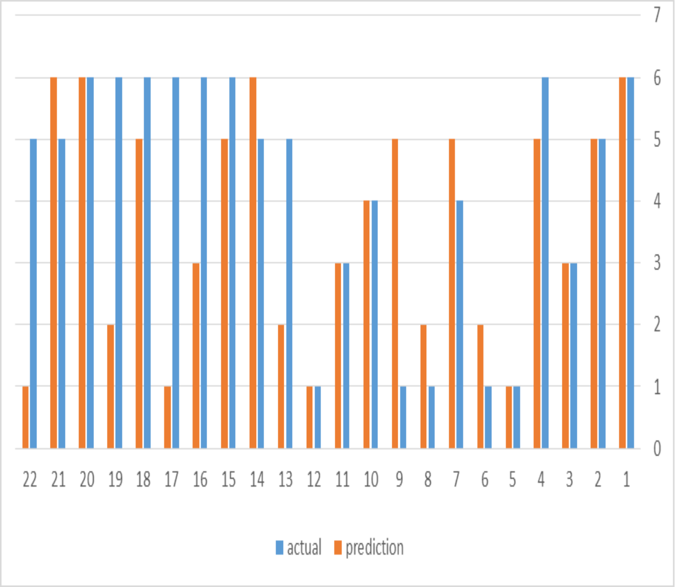
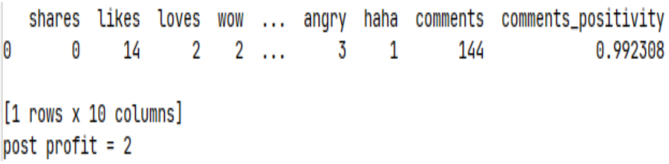


Figure 8. Comparing predicted value and real value of Decision tree.

Then to get the final decision by the system about the profitability of the service within a criterion of 1-6, a file containing the value of interactions on the selected post is sent to the Random Forest model and the output is as shown in (Figure .9).



```
   shares  likes  loves  wow  ...  angry  haha  comments  comments_positivity
0       0     14      2    2  ...      3     1       144             0.992308

[1 rows x 10 columns]
post profit = 2
```

Figure 9. Predicting profit of new service

Previous results showed that Random Forest algorithm is the algorithm with the best performance with an accuracy of up to 67%, and the ROC_AUC value is 0.75, which means a good ability of the model to distinguish between the classes when predicting the value of a class, in addition to its superiority in the values of (F1-score) and (Recall).

From the results, we conclude that:

- The number of data qa is relatively small with 690 records.
- The reason for the low accuracy is due to the nature of the target community and the nature of the studied company, since there is no clear difference between the interaction obtained on publications with high-profit and low-profit services.
- There is no real correlation between the service's profitability and the interactions on the post containing the service.
  For example, there are services whose profitability rating is 2 and some is 6, but are very similar in the number of interactions on posts when they are promoted in the company's Facebook posts.
- We conclude from the previous point that there is some randomness due to the appearance of the post for multiple categories of users and not only the target group of the service provided.
  All categories consider the company's Facebook post as a way of expressing their problems and asking page admins about various complaints and not necessarily expressing their opinion about the service provided in the post.

VI. CONCLUSION AND FUTURE WORK

The importance of profit forecasting will help many companies, especially in the telecommunications industries, to increase its income and achieve good revenues.

Social media is an environment that includes a huge amount of data, which, if taken advantage of, provides very important information about the market and users.

The research focused on the possibility of predicting the profitability of telecom company's new services through machine learning technologies, data processing and taking advantage of the data provided company's previous posts via Facebook.

Text and numerical data were collected, to take advantage of every detail of the post, and machine-learning algorithms were trained on the text data to help determine the overall orientation of the comments on the post.

In the last step, machine-learning algorithms were trained to try to find a link between the interactions of the post containing the service and its profitability obtained from Syriatel.

Several algorithms were chosen for comparison between them, due to their applicability and versatility in this type of application (Random Forest, DT, Gradient Boosting, K nearest neighbors, NB).

By using Random Forest, we will get more accuracy comparing other algorithms.

A possible improvement in the future to improve the accuracy of the system and obtain better results, is increasing the number of data used, and to look for a possible relationship between the type of post (photo, text, video), with the interactions obtained by the post with the profitability of the service within the post.

REFRENCES

[1] Ray, A., Bala, P.K. and Jain, R. , Utilizing emotion scores for improving classifier performance for predicting customer's intended ratings from social media posts, Benchmarking: An International Journal, (2021), Vol. 28 No. 2, pp. 438-464.

[2] Zimmerman, J., Ng, D., & Tusing, M. (2020). Social media marketing all-in-one for dummies: 4th edition. Unabridged. [United States]: Tantor Audio.

[3] Matthew A. Russell, Mikhail Klassen. Mining the Social Web, 3rd Edition, *O'Reilly Media, Inc.* January 2019

[4] Kaur, Wandeep & Balakrishnan, Vimala & Rana, Omer & Sinniah, Ajantha. (2018). Liking, Sharing, Commenting and Reacting on Facebook: User behaviors' impact on Sentiment Intensity. *Telematics and Informatics. 39.* 10.1016/j.tele.2018.12.005.

[5] مشارة نورالدين. دور التسويق عبر شبكات التواصل الإجتماعي في إدارة العلاقة مع الزبون -دراسة حالة متعاملي قطاع الهاتف النقال بالجزائر. *جامعة قاصدي مرباح- ورقلة* (2014).

[6] Sepehr Forouzani .Using social media and machine learning to predict financial performance of a company *.uppsala unversitet / teknisk- naturvetenskaplig fakultet* .(2016).

[7] Matthias Bogaert, Michel Ballings, Dirk Van den Poel, Asil Oztekin, Box office sales and social media: A cross-platform comparison of predictive ability and mechanisms, *Decision Support Systems*, (2021).

[8] Dipak Gaikar, Riddhi Solanki, Harshada Shinde, Pooja Phapale , Ishan Pandey .Movie Success Prediction Using Popularity Factor from Social Media *.International Research Journal of Engineering and Technology (IRJET)*, page 6, 04Apr, 2019.

[9] Chavan, Sandeep & Panchal, Simsri & Sawant, Tanvi & Shinde, Janhavi. (2020). Predicting Online Product Sales using Machine Learning. *International Journal of Engineering Research.*

[10] Purvika Bajaj ،Renesa Ray ،Shivani Shedge ،Shravani Vidhate, Prof. Dr. Nikhilkumar Shardoor .June, 2020 .Sales Prediction Using Machine Learning Algorithms *.International Research Journal Of Engineering And Technology (IRJET)*.

[11] B. Senthil Arasu, B. Jonath Backia Seelan, N. Thamaraiselvan, A machine learning-based approach to enhancing social media marketing, *Computers & Electrical Engineering*,Volume 86,2020.

[12] Gabor Szabo, Gungor Polatkan, P. Oscar Boykin, Antonios Chalkiopoulos. Social Media Data Mining and Analytics, *Wiley*, October 2018,

[13] Duarte, J. J., Montenegro González, S., & Cruz, J. C. (2020). Predicting stock price falls using news data: Evidence from the Brazilian market. *Computational Economics*.

[14] Cui, Ruomeng & Gallino, Santiago & Moreno, Antonio & Zhang, Dennis. The Operational Value of Social Media Information. *Production and Operations Management.* (2017).