

# Predicting Future Products Rate using Machine Learning Algorithms

**Shaimaa Mahmoud, Mahmoud Hussein and Arabi Keshk**

Computer Science Department, Faculty of Computers and Information, Menoufia University, Shebin Elkom 32511, Egypt

E-mail: sh.mahmoud600@gmail.com, mahmoud.hussein@ci.menofia.edu.eg, and arabikeshk@yahoo.com

Received: 10 February 2020; Revised: 04 March 2020; Accepted: 16 March 2020; Published: 08 October 2020

**Abstract:** Opinion mining in social networks data is considered as one of most important research areas because a large number of users interact with different topics on it. This paper discusses the problem of predicting future products rate according to users' comments. Researchers interacted with this problem by using machine learning algorithms (e.g. Logistic Regression, Random Forest Regression, Support Vector Regression, Simple Linear Regression, Multiple Linear Regression, Polynomial Regression and Decision Tree). However, the accuracy of these techniques still needs to be improved. In this study, we introduce an approach for predicting future products rate using LR, RFR, and SVR. Our data set consists of tweets and its rate from 1:5. The main goal of our approach is improving the prediction accuracy about existing techniques. SVR can predict future product rate with a Mean Squared Error (MSE) of 0.4122, Linear Regression model predict with a Mean Squared Error of 0.4986 and Random Forest Regression can predict with a Mean Squared Error of 0.4770. This is better than the existing approaches accuracy.

**Index Terms:** Twitter, Sentiment Analysis, Machine Learning, prediction.

## 1. Introduction

The importance of opinion mining is increasing every day because of the large number of information published in social media and people interact with it. Opinion mining focuses on studying the interaction and communications of people on different topics in order to benefit from their feedback [1].

Many applications is benefiting from using opinion mining. For example, in a business where opinions of customers about the published products are extracted and analyzed for improving the company products performance. The performance is improved through predicting the rate of each product and then the future profit of each product is predicated. Another example is in the education system, where student's feedback about courses is extracted to predict courses that students are interested in. The main problem in such applications is how to classify the extracted opinions with a high level of accuracy and then predict future information such as future products rate and future interested courses with fewer errors [2,3]. In this work we introduce an approach for improving the prediction accuracy through modifying data preprocessing phase of the predication process. We perform this work using the machine learning prediction algorithms for example (LR, RFR, and SVR) and both (unigrams and bigrams) to extract the features from the data in order to training the algorithms. There are many researchers interested in this point in order to benefit from user's feedback using machine learning prediction algorithms. For example, Random Forest Regression Algorithm is used for Predicting the tweets popularity based on Re-tweets by users [1]. researchers used dataset has 12,470,144 (English) tweets that go from 1st July 2016 until 15th of July gathered from the twitter streaming API in order to present a study on the importance of various text features in predicting the tweets popularity (e.g. number of re-tweets), as the importance of the user's history of re-tweets. The main goal in this work is how to build a Re-tweet Predictive Model and show how the distinct number of text features. Another approach is introduced for predicting the future interests of users on Twitter based on known a minimum user interaction with the topic to perform prediction [4]. Researchers use the twitter dataset that consists of approximately 3M tweets posted by 135.731 unique users in order to predict the future interests of these users with regard to potentially trending topics. The work aims to extend the state of the art through the prediction of users' interests with regard to future unobserved topics.

Twitter considers the richest sources to learn about people's opinions and interactions [5]. But, the existing techniques' accuracy still needs to be improved. This paper explains an approach to improve the accuracy of predicted products' rate depending on the users' comments. We use (API) to extract a large number of tweets which include users' opinions about products and its rate from 1:5. The tweets are then split to two parts in order to train and test different machine learning prediction algorithms on it. Before the training phase, the data is cleaned by (a) converting all words

to lower case, (b) removing usernames, mentions, links, repeated characters, numbers, empty spaces, and tweets, punctuations and stop words, and (c) converting all words like “isn’t” to “is not”. Tokenization, lemmatization, and stemming are also used to join all words to the data frame.

We implement the proposed approach using Python language with different prediction algorithms such as LR [6], RFR [7] and SVR [8,9] to achieve the best accuracy of the prediction process compared to existing techniques. The features are extracted using unigrams and bigrams. SVR can predict future product rate with a Mean Squared Error (MSE) of 0.4122, Linear Regression model predict with a Mean Squared Error of 0.4986 and Random Forest Regression can predict with a Mean Squared Error of 0.4770. These results mean that our approach accuracy for the prediction process is better than the existing approaches.

This paper is organized as follows. In Section 2, we present related work of data mining and sentiment analysis, which uses different machine learning for the prediction. Section 3 is about the machine learning methods that are used in our proposed approach. In section 4, our approach for improving prediction accuracy is introduced. The results when using different prediction algorithms are presented in Section 5. Conclusion and future work are put forward in Section 6.

## 2. Related Work

Twitter consider the most important and used social network in which people can communicate with each other and interact with different published topics. Twitter allows peoples to write a tweet in order to express their opinions about different topics. Extracting these opinions is not easy because of there is a huge number of tweets written everyday by people, and then it is an important process. We can benefit from it as feedback about users’ opinions on the different topics.

An approach has been introduced to predict the tweets popularity based on users’ re-tweets [10]. In this study, researchers have used a dataset that has 12,470,144 (English) tweets that go from first of July 2016 until the 15<sup>th</sup> of July 2016 gathered from the twitter streaming API. Twitter search API is used in order to present a study on the importance of different text features in predicting the tweets popularity (e.g. the number of re-tweets, as the importance of the user’s history of re-tweets). The features used in this study are divided into two main groups: user features, and tweets features. The user features include number of followers, number of states, number of favorites, number of times the user was listed, number of days of the accounts, and the user is verified (or not). The tweet features includes number of hashtags, URLs, and mentions, length of the tweets, and number of words.

They have applied the Random Forest Regression Algorithm with data splitting ratio of 80:20 where 80% is used for training and 20% for testing. The goal of this work is how to build a Re-Tweet Predictive Model and show how the distinct number of text features. The resulting Re-tweet predictive Model takes into consideration different kinds of tweets (e.g. tweets with hashtags and URLs, among the used quality categories). Results show that there is a strong relationship between specific features, e.g. user’s popularity. Fig.1 shows the steps for predicting the tweets popularity. It contains twitter data extracted from twitter, which is then analyzed in order to create the data set. Data set is split into two data sets first one for training learning algorithms, and a second for testing. Finally, prediction process implemented in order to know its accuracy.

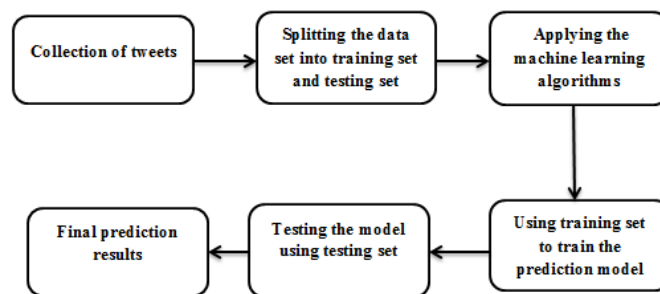


Fig.1. Re-tweet Predictive Model [10]

Social networks become big data production engines that can be analyzed in order to know trending topics which used in different applications [11]. This work addresses the problem of common topics and trends prediction out of social network information streams, using a time series classification model, which is implemented over a flexible and adaptive distributed framework. Trend prediction is employed in a real-time fashion, utilizing big data principles and technologies, under a framework designed in the Lambda architecture outline. Lambda design has been chosen because of its flexibility in the process each evolving and static information threads supported a tunable latent source model.

Data are collected from social networks via Twitter streaming API, over several time windows. Due to lack of space, a sample of 1-month period (Nov-Dec 2013) is discussed. Its total size exceeding 300 GB, along with the Twitter trending topics announced for the same time window. Data arrival originates from the twitter streaming API. Thus,

Hadoop3 and its implementation of the MapReduce model have been chosen [11]. The prediction of trending topics in this work follows a series comparison process which requires a sample of the dataset in a specific time. This data set is built based on twitter trending topics and by performing retrospective analysis on the data set. The planned model sets the supposed latent source “signals” such as a model event of a certain kind, and a clustering process is combining with the classification tasks of labeling the data threads over specific classes (either detected as trends or not). Trending topic prediction has reached an accuracy of 78.4 %.

Another work is performed to predict User’s Future Interests on Twitter based on known a minimum user interaction with the topic to perform prediction [12]. In this work, researchers use the twitter dataset, which consists of approximately 3M tweets posted by 135,731 unique users to predict user’s future interests with regards to potentially trending topics of the future. They used a method to extract the historical topic profile of users then predict the future interests of them depending on the prediction model. The work aims to extend the state of the art by predicting users’ interests concerning future unobserved topics. For example, they are interested in determining whether a given user would be interested in following the news about the release of a new mobile operating system that would compete with iOS.

This approach uses the Wikipedia category structure in order to model high-level user interests. It also considers the temporal evolution of the user’s interests to predict the user’s future interests. This paper has three key contributions. First, they proposed a model that transfers the user’s interests from different time intervals onto Wikipedia’s category structure. Second, they showed how semantic information derived from the Wikipedia knowledge base as well as temporal information can be integrated into this model to predict user’s interests with regards to unobserved topics of the future on Twitter. Third, they perform experimentation to show the impact of considering Wikipedia categories on the accuracy of predicting the future interests of users on Twitter. Finally, they used root mean absolute error (RMAE) and mean absolute error (MAE) in order to evaluate the accuracy of prediction.

All of the works shown above are good and have achieved reasonable accuracy. But, the ratio of splitting the data needs to be modify and the data preprocessing needs to be also modified in order to improve their experiment’s results.

### 3. Machine Learning Methods

Machine learning algorithms are fast to train, easy to understand and implemented with a different programming language such as python. They also give the best results of both text classification and prediction. Therefore, we use such algorithms in our proposed approach in next section.

In this section, we explain in brief machine learning algorithms that are used in our proposed approach. We use different prediction algorithms and feature extractors to achieve less error in prediction accuracy. The prediction algorithms used in our approach are LR, RFR, and SVR while the use of unigrams and bigrams to extract the features from the dataset.

#### 3.1. “Support Vector Machine (SVM)”

One of the most important supervised machine learning algorithms that we can use it in both classification and prediction process to benefit from their results in our life [8,9]. It gives us the best result for our experiment. However, it is mostly used in classification problems. In this algorithm, each data item is plotted as a point in n-dimensional space (where n is the number of features) with the value of each feature being the value of a particular coordinate. Then, prediction is performed according to the information, which extracted from the dataset.

#### 3.2. “Random Forest Regression (RFR)”

Is a supervised machine learning algorithm which used in both the classification and regression process [7]. It may be a material technique and not a boosting technique. The trees within the random forests area unit run in parallel. There is no interaction between these trees whereas building the trees. It operates by constructing a large number of decision trees at training time. The, it outputs the category that is the mode of the categories (classification) or means prediction (regression). It combines the results of multiple predictions that aggregate several decision trees, with some useful modifications. The amount of features, which will be split on at every node, is proscribed to some share of the entire (which is thought because of the hyperparameter). This ensures that the ensemble model does not believe too heavily on somebody’s feature, and makes use of all most likely prophetic choices. Every tree draws a random sample from the initial information set when generating its splits, adding a further part of randomness that prevents overfitting. The random forest makes predictions by combining decisions from a sequence of base models.

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \dots \quad (1)$$

$g$  is the sum of simple base models  $f_i$

### 3.3. “Logistic regression (LR)”

Logistic regression may be an applied math analysis methodology used to predict an information value supported by previous observations of a data set [6,13]. Logistic regression has become a very important tool within the discipline of machine learning. The approach allows an algorithm being employed during a very machine learning application to classify incoming data supported by historical data. As additional relevant data comes in, the algorithmic program ought to improve at predicting classifications inside data sets. Logistic regression also can play a task in data preparation activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, and load (ETL) the process to stage the data for analysis. LR model can predicts a dependent data variable by analyzing the link between one or additional existing independent variables.

$$p(x) = e^{b_0+b_1x} / (1 + e^{b_0+b_1x}) \tag{2}$$

It can be transformed into:  $-\ln\left(\frac{p(x)}{1-p(x)}\right) = b_0 + b_1x$

The goal is to use the training data to find the values of coefficients  $b_0$  and  $b_1$  in order to minimize the error between the predicted data and the actual data.

## 4. Proposed Approach

Our proposed approach focus on improving the prediction accuracy by modifying the data pre-processing phase (see Fig.2). The modification includes: removing stop words, tokenization, lemmatization, and stemming. Our approach consists of three phases. In phase 1, tweets are extracted from twitter using an API. Then, in phase 2, features are selected which are used in the training step. Finally, in phase 3, each product rate is predicted with a range from 1 to 5.

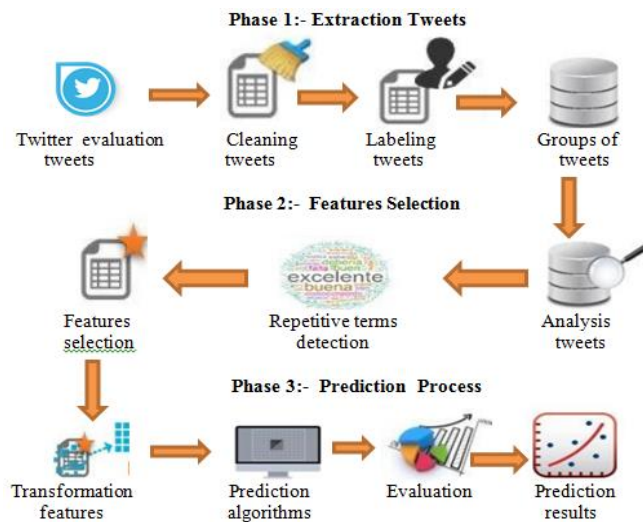


Fig.2. The architecture of our proposed approach

#### 4.1. Phase 1: Extracting Tweets.

In this phase, tweets are extracted from Twitter using an (API). Then the tweets converted into a data frame and names are given to the data frame in order to clean it. The data frame consists of tweets and its rate from 1 to 5.

1) Cleaning tweets: For each tweet, it is important to clean it. Therefore, we (a) convert all words to lower case, remove (usernames, mentions, links, repeated characters, numbers, empty spaces and tweets, punctuations, and stop words), and (b) convert all words like “isn’t” to “is not” using tokenization, lemmatization, and stemming. Then, all words are joined to the data frame. In the following, we describe how user names, links, and repeated letters are treated.

2) Usernames: Users regularly incorporate usernames in tweets to coordinate their messages. An accepted standard is to incorporate the “@” image before the username (e.g. @alecmgo). Thus, comparability class token (USERNAME) replaces all words that begin with the @symbol.

3) Use of links: Users all the time incorporate connections in their tweets. A comparability class is utilized for all URLs where a URL like "http://tinyurl.com/cvvg9a" is converted to the token "URL".

4) Repeated letters: Tweets contain normal language. For instance, if you search about "love" with a repetitive number of o's in the center (for example loove, loooooove, looooooove) there likely a non-empty result set. In our approach, we use preprocessing so any letter repetitive more than two times in any word is replaced with two letters. In

the examples above, these words would be changed into the token loove.

5) Labeling tweets: Through labeling, each tweet is given a label, which is from 1:5 according to its rate. Then, these tweets are split into two data sets (training set, testing set)

4.2. Phase 2: Features Selection

After creating the tweets groups, they are analyzed in order to detect the most repetitive terms according to the polarity of each term. The most repetitive terms are considered as features, these features extracted by the use of unigrams and bigrams. They can give useful information on the relative importance data or relevance of features for a given problem. This information can help to filter the data set and increase the accuracy of our models. Then, we used these features in order to train our algorithms [logistic regression, Random Forest Regression and Support Vector Regression] that help us to achieve high accuracy of sentiment prediction models.

4.3. Phase 3: Prediction Process.

In this phase, the dataset is split with the ratio 70:30, where 70% used to train the algorithms while 30% is used for testing. Then, a new instance vector of Tf-idf is made and the approach is fed with (a) a parameter as analyzer ="Word", (b) a parameter called stop\_words="English", and (c) ngrams range for feature selection. The n-grams range is (1, 2) where unigrams and bigrams are used [14]. Finally, the machine learning algorithms [15]: LR, RFR and SVR are trained by the training data, and the testing data are used to calculate the accuracy of the prediction process by comparing the predicted labels and the labels of the original data.

5. Results

In this section, we present the results of prediction future Mobiles rate by using different loss error measures (i.e. Mean Absolute Error, Mean Squared Error, Median Squared Error), and different machine Learning prediction algorithms (i.e. Logistic Regression, Random Forest Regression and Support Vector Regression). The dataset used in this experiment consists of four mobile phone categories (Blackberry, iPhone, Lenovo, and Samsung). Each category contains 34,000 comments about the user’s opinions of each product, and each comment labeled with the rate from 1 to 5 depending on the user feedback.

Mean Squared Error formula:-

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \tag{3}$$

Mean Absolute Error formula:-

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{4}$$

Where in equation (3) and equation (4), (n) is the number of data points,  $y_i$  present actual value, and  $\hat{y}_i$  present predicted value which returned by the model

5.1. Blackberry

Predictions of future blackberry rate when using Logistic Regression, Random Forest Regression and Support Vector Regression and extracted features by using both unigrams and bigrams (1, 2) are shown in Fig.3, Fig.4, and Fig.5 respectively.



Fig.3. Predicted values (LR) model



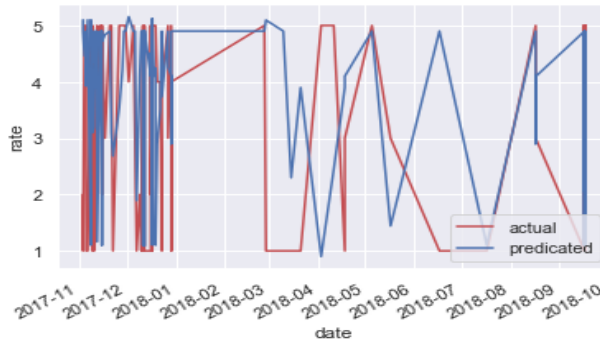


Fig.4. Predicted values (SVR) model

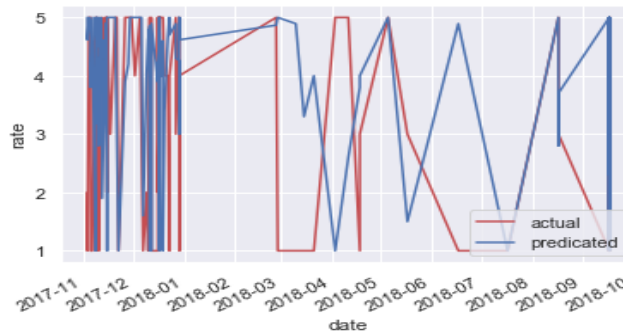


Fig.5. Predicted values (RFR) model

Table 1. presents summary about the accuracy of Prediction future blackberry rate when using Logistic Regression, Random Forest Regression and Support Vector Regression and extracted features by using both unigrams and bigrams (1, 2)

Table 1. Accuracy of prediction future blackberry rate with different prediction algorithms

Model	Train score	Test score	Mean Absolute Error	Mean Squared Error	Median Squared Error
Linear Regression	0.9750	0.8483	0.3006	0.3811	0.0000
Support Vector Regression	0.9371	0.8509	0.3531	0.3747	0.1002
Random Forest Regression	0.9496	0.8157	0.3332	0.4632	0.1000

### 5.2. iPhone

Predictions of future iPhone rate when using Logistic Regression, Random Forest Regression and Support Vector Regression and extracted features by using both unigrams and bigrams (1, 2) are shown in Fig.6, Fig.7, and Fig.8 respectively.

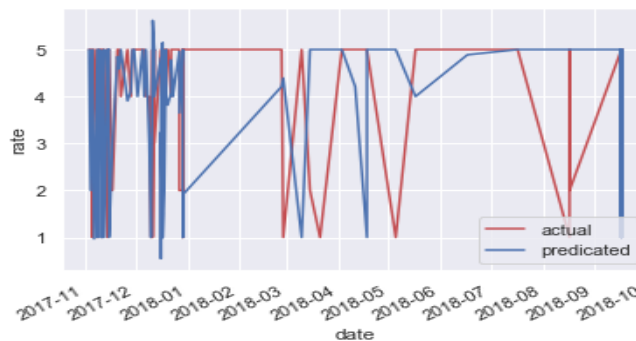


Fig.6. Predicted values (LR) model

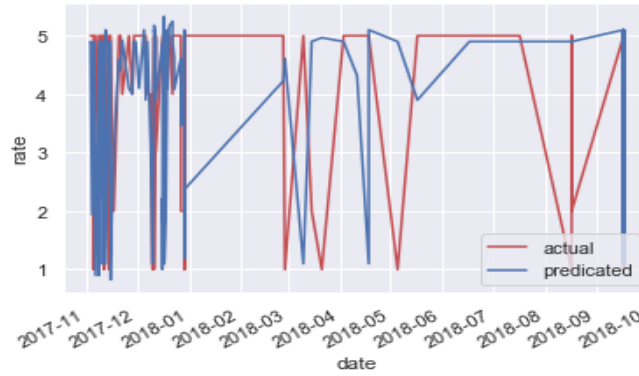


Fig.7. Predicted values (SVR) model

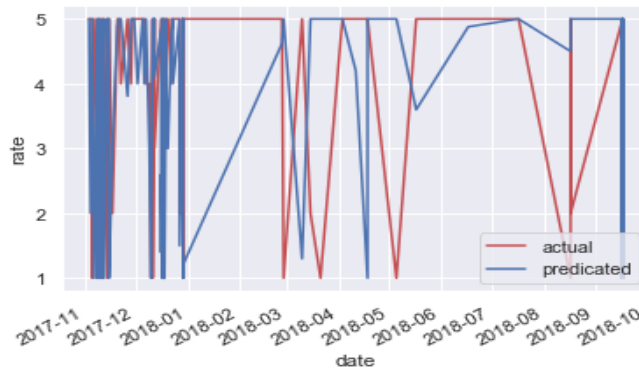


Fig.8. Predicted values (RFR) model

Table 2. presents summary about the accuracy of Prediction future iPhone rate when using Logistic Regression, Random Forest Regression and Support Vector Regression and extracted features by using both unigrams and bigrams (1, 2)

Table 2. Accuracy of prediction future iPhone rate with different prediction algorithms

Model	Train score	Test score	Mean Absolute Error	Mean Squared Error	Median Squared Error
Linear Regression	0.9891	0.8866	0.2304	0.3010	0.0000
Support Vector Regression	0.9520	0.8837	0.3170	0.3085	0.1002
Random Forest Regression	0.9674	0.8666	0.2572	0.3539	0.0000

### 5.3. Lenovo

Predictions of future Lenovo rate when using Logistic Regression, Random Forest Regression and Support Vector Regression and extracted features by using both unigrams and bigrams (1, 2) are shown in Fig.9 , Fig.10, and Fig.11 respectively.



Fig.9. Predicted values (LR) model



Fig.10. Predicted values (SVR) model



Fig.11. Predicted values (RFR) model

Table 3. presents summary about the accuracy of Prediction future Lenovo rate when using Logistic Regression, Random Forest Regression and Support Vector Regression and extracted features by using both unigrams and bigrams (1, 2)

Table 3. Accuracy of prediction future Lenovo rate with different prediction algorithms

Model	Train score	Test score	Mean Absolute Error	Mean Squared Error	Median Squared Error
Linear Regression	0.9877	0.7750	0.4047	0.5216	0.0757
Support Vector Regression	0.9237	0.7856	0.4544	0.4970	0.1547
Random Forest Regression	0.9403	0.7165	0.4708	0.6574	0.2000

#### 5.4. Samsung

Predictions of future Samsung rate when using Logistic Regression, Random Forest Regression and Support Vector Regression and extracted features by using both unigrams and bigrams (1, 2) are shown in Fig.12, Fig.13, and Fig.14 respectively.





Fig.12. Predicted values (LR) model

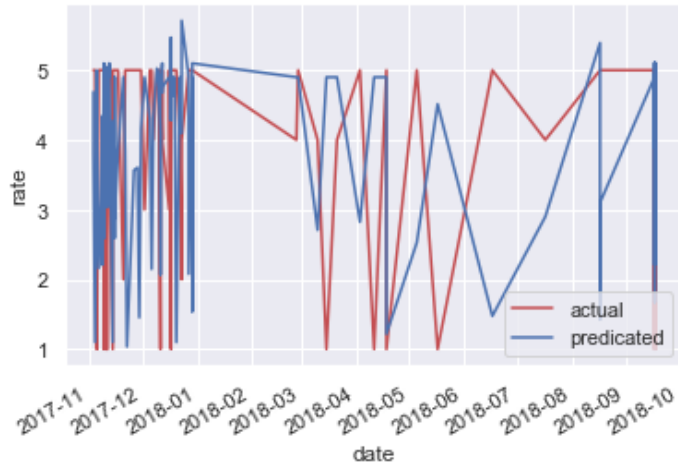


Fig.13. Predicted values (SVR) model



Fig.14. Predicted values (RFR) model

Table 4. Presents summary about the accuracy of Prediction future Samsung rate when using Logistic Regression, Random Forest Regression and Support Vector Regression and extracted features by using both unigrams and bigrams (1, 2)

Table 4. Accuracy of prediction future Samsung rate with different prediction algorithms

Model	Train score	Test score	Mean Absolute Error	Mean Squared Error	Median Squared Error
Linear Regression	0.9806	0.7376	0.4211	0.5765	0.0787
Support Vector Regression	0.9099	0.7682	0.4567	0.5093	0.1573
Random Forest Regression	0.9322	0.7087	0.4623	0.6399	0.2000

Finally, after modifying the preprocessing phase by, (a) removing punctuation, (b) numbers, (c) stop words from the data, (d) empty tweets and spices, (e) repeated characters and tweets. (f) Using tokenization, (g) lemmatization, (h) stemming, (i) N-grams and Tf-idf vectorizer. In prediction process (a) prediction future blackberry rate, when using logistic regression we achieve 0.3006 when using mean absolute error, 0.3811 with mean squared error, and 0.0000 with median squared error.

When using support vector regression we achieve 0.3531 when using mean absolute error, 0.3747 with mean squared error, and 0.1002 with median squared error. When using random forest regression we achieve 0.3332 when using mean absolute error, 0.4632 with mean squared error, and 0.1000 with median squared error. (b) Prediction future iPhone rate, when using logistic regression we achieve 0.2304 when using mean absolute error, 0.3010 with mean squared error, and 0.0000 with median squared error. When using support vector regression we achieve 0.3170 when using mean absolute error, 0.3085 with mean squared error, and 0.1002 with median squared error. When using random forest regression we achieve 0.2572 when using mean absolute error, 0.3539 with mean squared error, and 0.0000 with median squared error. (c) Prediction future Lenovo rate, when using logistic regression we achieve 0.4047 when using mean absolute error, 0.5216 with mean squared error, and 0.0757 with median squared error. When using support vector regression we achieve 0.4544 when using mean absolute error, 0.4970 with mean squared error, and 0.1547 with median squared error. When using random forest regression we achieve 0.4708 when using mean absolute error, 0.6575 with mean squared error, and 0.2000 with median squared error. (d) Prediction future Samsung rate, when using logistic regression we achieve 0.4211 when using mean absolute error, 0.5765 with mean squared error, and 0.0787 with median squared error. When using support vector regression we achieve 0.4567 when using mean absolute error, 0.5093 with mean squared error, and 0.1573 with median squared error. When using random forest regression we achieve 0.4623 when using mean absolute error, 0.6399 with mean squared error, and 0.2000 with median squared error.

## 6. Conclusion and Future Work

In this work, we implemented an approach for opinion mining in social networks data with the aim of improving the accuracy of predicting future products rate according to users' feedback about each product. The feedback is extracted from Twitter using an application programming interface (API). Then, machine learning algorithms such as LR, RFR, and SVR are used to predict future products rate. we applies our improving by modifying data preprocessing phase in order to clean it by using (tokenization, stemming and lemmatization), convert all words to lower case, remove (usernames, mentions, links, repeated characters, numbers, empty spaces and tweets, punctuations, and stop words), and convert all words like isn't to is not to cleaning the data, using both (unigrams and bigrams) to extract the features from the data. We present the results of prediction by using different loss error measures (i.e. Mean Absolute Error, Mean Squared Error, and Median Squared Error). The dataset is split with the ratio of 70:30, where 70% used to train the algorithms while the 30% is used for testing. SVR model can predict future products rate with a Mean Squared Error (MSE) of 0.4122, Linear Regression can predict with a Mean Squared Error of 0.4986 and Random Forest can predict with a Mean Squared Error of 0.4770, which is better than existing approaches accuracy.

In the future, we will plan to estimate the future profit of each product, and the time complexity for the proposed approach.

## References

- [1] Gutiérrez, G., et al. "Analyzing Students Reviews of Teacher Performance Using Support Vector Machines by a Proposed Model." International Symposium on Intelligent Computing Systems. Springer, Cham, 2018.
- [2] Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. "Sentiment analysis in Facebook and its application to e-learning." *Computers in human behavior* 31 (2014): 527-541.
- [3] Esparza, Guadalupe Gutiérrez, et al. "A sentiment analysis model to analyze students reviews of teacher performance using support vector machines." International Symposium on Distributed Computing and Artificial Intelligence. Springer, Cham, 2017.
- [4] Kaewyong, Phuripoj, et al. "The possibility of students' comments automatic interpret using lexicon based sentiment analysis to teacher evaluation." 3rd International Conference on Artificial Intelligence and Computer Science (AICS2015). 2015.
- [5] Sarlan, Aliza, Chayanit Nadam, and Shuib Basri. "Twitter sentiment analysis." *Proceedings of the 6th International conference on Information Technology and Multimedia*. IEEE, 2014.
- [6] Wi, D. A. V. I. D. "Applied logistic regression." (2000).
- [7] Palmer, David S., et al. "Random forest models to predict aqueous solubility." *Journal of chemical information and modeling*

- 47.1 (2007): 150-158.
- [8] Suthaharan, Shan. "Support vector machine." *Machine learning models and algorithms for big data classification*. Springer, Boston, MA, 2016. 207-235.
  - [9] "Support Vector Machines." *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, by Nello Cristianini and John Shawe-Taylor, Cambridge University Press, Cambridge, 2000, pp. 93–124.
  - [10] Oliveira, Nelson, et al. "Retweet Predictive Model for Predicting the Popularity of Tweets." *International Conference on Soft Computing and Pattern Recognition*. Springer, Cham, 2018.
  - [11] Vakali, Athena, Nikolaos Kitmeridis, and Maria Panourgia. "A distributed framework for early trending topics detection on big social networks data threads." *INNS Conference on Big Data*. Springer, Cham, 2016.
  - [12] Zarrinkalam, Fattane, et al. "Predicting users' future interests on Twitter." *European Conference on Information Retrieval*. Springer, Cham, 2017.
  - [13] Westreich, Daniel, Justin Lessler, and Michele Jonsson Funk. "Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression." *Journal of clinical epidemiology* 63.8 (2010): 826-833.
  - [14] Marouli, Georgios. "Comparison between Maximum Entropy and Naïve Bayes classifiers: Case study; Appliance of Machine Learning Algorithms to an Odesk's Corporation Dataset." (2014).
  - [15] Maetschke, Stefan R., et al. "Supervised, semi-supervised and unsupervised inference of gene regulatory networks." *Briefings in bioinformatics* 15.2 (2014): 195-211.

### Authors' Profiles



**Shaimaa Mahmoud** is a master student at Menoufia University, Egypt. She received her BSc in Computer Science from 6 October University, Faculty of Computers and Information in 2017 respectively. Her main research interest includes Data Mining, and Machine Learning.



**Mahmoud Hussein** received his BSc. and MSc. in Computer Science from Menoufia University, Faculty of Computers and Information in 2006 and 2009 respectively and received his PhD in Software Engineering from Swinburne University of Technology, Faculty of Information and Communications Technology in 2013. His research interest includes Software Engineering, Data Mining, Machine Learning, Data Privacy, and Security.



**Arabi E. keshk** received the B.Sc. in Electronic Engineering and M.Sc. in Computer Science and Engineering from Menoufia University, Faculty of Electronic Engineering in 1987 and 1995, respectively and received his PhD in Electronic Engineering from Osaka University, Japan in 2001. His research interest includes software testing, software engineering, distributed system, database, data mining, and bioinformatics.

**How to cite this paper:** Shaimaa Mahmoud, Mahmoud Hussein, Arabi Keshk, "Predicting Future Products Rate using Machine Learning Algorithms", *International Journal of Intelligent Systems and Applications(IJISA)*, Vol.12, No.5, pp.41-51, 2020. DOI: 10.5815/ijisa.2020.05.04