Classification and Prediction of Opinion Mining in Social Networks Data

Shaimaa Mahmoud, Mahmoud Hussein, and Arabi Keshk
Computer Science Department, Faculty of Computers and Information,
Menoufia University, Shebin Elkom 32511, Egypt,
sh.mahmoud600@gmail.com, mahmoud.hussein@ci.menofia.edu.eg, arabikeshk@yahoo.com

Abstract- Opinion mining in social networks data considers one of the most significant and challenging tasks in our days due to the huge number of information distributed each day. We can profit from these opinions by utilizing two significant procedures (classification and prediction). Although there is many researchers' interest in and work at this point, it still needs improvement. Therefore, in this paper, we present a method to improve the accuracy of the classification and prediction processes. The improvement is done through cleaning the data set by converting all words to lower case, removing usernames, mentions, links, repeated characters, numbers, delete more than two spaces between words, empty tweets, punctuations and stop words, and converting all words like "isn't" to "is not". In the feature selection phase, we use both unigrams and bigrams in order to extract the features from the data to training it. Our data set contains the user's feelings about distributed products, tweets labeled positive or negative, and each product rate from one to five. We implemented this work using different supervised machine learning algorithms like Naïve Bayes, Support Vector Machine and Max Entropy for the classification process, and Random Forest Regression, Logistic Regression, and Support Vector Regression for the prediction process. At last, we have accuracy in the classification and prediction process better than existing works. In classification, we achieved accuracy of 90% and in the prediction process, Support Vector Regression model is able to predict future product rate with a mean squared error (MSE) of 0.4122, Logistic Regression model is able to predict with a mean squared error of 0.4986 and Random Forest Regression model is able to predict with a mean squared error of 0.4770.

Keywords— *Twitter, Sentiment Analysis, Machine Learning, Classification, and Prediction.*

I. INTRODUCTION

The significance of opinion mining is expanding each day due to the enormous number of data distributed in social media, and people interact with this. Opinion mining focuses on studying the interaction and communications of people on different topics in order to benefit from their feedbacks [1].

Numerous applications profit from opinion mining. For instance, in the education system, students' feedback on the performance of teachers can be extracted to evaluate their performance. Thus, the learning process is improved in ways that are more useful to students by identifying teachers' disadvantages and provide them with appropriate training to improve their performance [4].

Another model where the assessments of clients about the distributed items extracted and analyzed for improving the performance of the company products and predicting the profit of each product. The main issue in such applications is how to classify and predict the extracted opinions with a high level of accuracy.

Currently, there is much research in the area of opinion mining in social network data, but this work still needs to be improved. Therefore, in this paper, we introduce a method to improve the accuracy of the classification and prediction processes. The dataset consists of four mobile phone categories extracted from twitter using application programming interface (API). Each category contains 34,000 comments about the user's opinions of each product, each comment labeled into positive or negative, and the rate from one to five [1]. Then, we split it into two parts in order to train and test different machine learning classification and prediction algorithms on it. The data is cleaned by several steps (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 implemented this work

using different supervised machine learning algorithms like NB, SVM, and Max Entropy for the classification process and RFR, LR, and SVR for the prediction process. Finally, in classification, we achieved an accuracy of 90%, and in the prediction process, SVR model can predict future product rate with a mean squared error (MSE) of 0.4122, LR model can predict with a mean squared error of 0.4986 and RFR model can predict with a mean squared error of 0.4770.

This paper is organized as follows. In Section 2, we present some related work of data mining and sentiment analysis, which using different machine learning classifiers and predictors. In section 3, our approach for improving the classification and prediction accuracy is introduced. The results of our approach are presented in Section 4. Conclusion and future work are put forward in Section 5.

II. RELATED WORK

In this section, we present related work that uses AI with different features and presents their classifiers' and predictor's accuracy. The related work is divided into two groups: classification and predication.

A. Classification Process

A method has been proposed to detect the feeling of students on some topics and support the teacher to improve their teaching process [4]. Social networks become the most important and easy way for everybody to express their opinions about any topic in our life. In this work, they want to know student's opinions about teacher performance in order to benefit from these opinions as feedback about teacher performance then provide them with appropriate courses to improve their performance. A number of learning algorithms have been used to classify reviews data sets and comments that extracted from twitter using application programming interface (API) in order to extract students' opinions. Then, the results are used as feedback about teacher performance into positive or negative to improve the learning process by identifying teachers' disadvantages and provide them with appropriate training. Data set consists of comments by 75 students collected from Moodle (mixed graph of the terms building module as shown in Fig 1). The model draws a graph of terms from a set of documents belong to knowledge, and it is labeled according to the feeling expressed in them. Fig 1 showing the architecture (related to the beautiful design and construction of buildings, etc.) of the feeling analysis model that contains a training set, the mixed graph of terms building module, document, and feeling mining module in order to know the result about the feeling of a document. Finally, they had an accuracy of 82% when SVM is used to train the data, while it is 81% with Naive Bayes and Maximum Entropy.

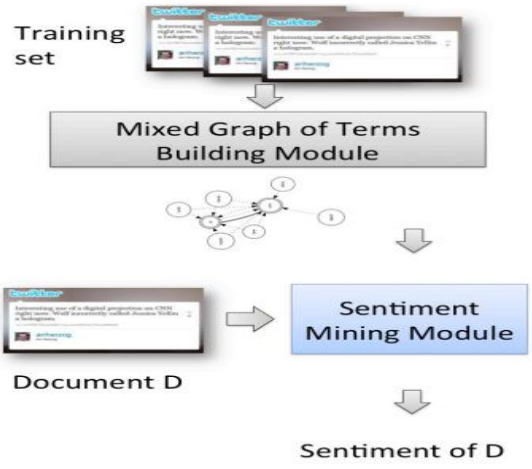

Fig 1. The architecture of sentiment analysis model [4]

An approach for automatically classifying the sentiment of twitter messages extracted from users' reviews into positive or negative have been proposed [2,12]. In this work, they are implemented the classification process by creating a dataset consist of a group of 752 negatives and a group of 1301 positive reviews in order to trading and testing the classification model. Then, NB, Max Entropy, and SVM classifiers applied in order to classify these reviews. Finally, they had accuracy around 80% when these algorithms are trained, considering emotion data as noise.

A sentiment analysis model to analyze students' reviews of teacher performance using Support Vector Machines has been introduced [1]. In this study, researchers used groups of comments in the Spanish language which extracted from student's reviews. They have applied support vector machine algorithms with linear, radial, and polynomial kernels in order to classify these comments into positive, negative or neutral. The accuracy achieved is 80%, 78%, and 67% when SVM Linear, SVM Radial, and SVM Polynomial are used respectively.

Students comments about teacher performance assessment using machine learning algorithms has been analyzed using algorithms such as Support Vector Machines and Random Forest [15]. The goal is to classify students' feedbacks on the performance of teachers into positive or negative. Then evaluate their performance to improve the learning process in ways that are more useful to students. They collected data from twitter tweets and self-encrypting drives (SED) model, and then applied (Support Vector Machines with three kernels: Linear, Radial Basis and Polynomial) and Random Forest. The accuracy achieved is 85%.

B. Prediction Process

An approach has been introduced to predict the popularity of Tweets based on re-tweets by users [9]. In this study, researchers used a dataset that has 12,470,144 tweets with the English language that collected from 1st of July 2016 until the 15th of July. Features used in this study divided into two main groups. They have applied the Random Forest Regression Algorithm. Data splitting with ratio 80: 20, where 80% is used for training and 20% for testing. The main goal is to predict the popularity of Tweets based on re-tweets by users. Fig 2 shows the steps of the re-tweet Predictive model. It contains twitter data extracted from twitter then analyzes it to create the data set. Data set split into two data sets first one for training machine-learning algorithms, second for testing. Finally, the prediction model is applied in order to know this accuracy.

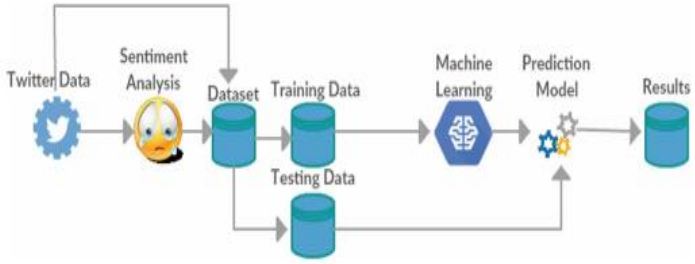

Fig 2. Re-tweet predictive model [9]

Social networks become big data production engines that can be carefully studied in order to know the common topics used in different applications [13]. This work deals with the problem of common topics and predicts big things outside social network information flows, using the time series classification model. Data is collected from social networks using API. Its total size goes beyond 300 GB, along with popular twitter topics at the same time. This work follows a chain comparison that needs data set at a time series. This data set is built based on twitter trending topics and by performing retrospective analysis on the data. The planned model sets the supposed potential source "signals," such as a model event of a specific 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 used to predict the user's future interests on social networks data for example twitter [14]. Researchers implemented their experiment using data set consists of 3M tweets written by 134,731 users

to predict their future interests. They used a method in order to extract this data from each user's profile then used prediction model to implement their experiment. This work aims to extend the state of art by predicting the user's future interests concerning future topics. Finally, the researchers used root mean absolute error and mean absolute error to evaluate the prediction accuracy.

III. PROPOSED APPROACH

In this work, we introduce an approach for improving the classification and prediction accuracy of opinion mining in social networks data by modifying the data-preprocessing phase (see Fig 3). Our approach consists of four phases. In Phase 1, tweets are extracted from twitter using application programming interface (API) then converted into a data frame in order to clean it. We clean the data set by several steps as: (a) converting all words to lower case, (b) removing usernames, mentions, links, repeated characters, numbers, (c) delete more than two spaces between words, empty tweets, punctuations and stop words, and (d) converting all words like "isn't" to "is not". In Phase 1, we also use tokenization, stemming and lemmatization. In Phase 2, (feature selection), we using both unigrams and bigrams in order to extract the features from the data to training it. In Phase 3, each product is classified into positive or negative. Finally, in Phase 4, the product rate is predicted in a range from 1 to 5 depending on users' feedback about each product.

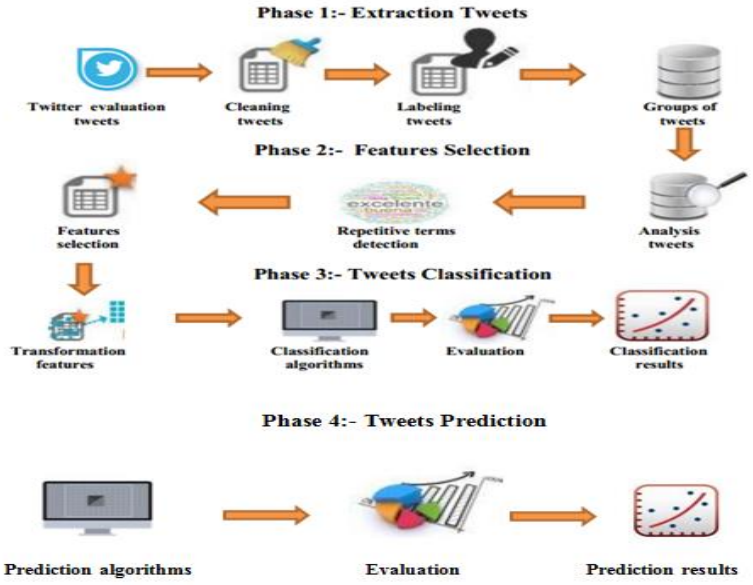

Fig 3. The architecture of our proposed approach

A. Phase 1: Extracting Tweets.

Tweets collected from twitter using application programming interface (API) then converted to data frame to clean it. The data set, which used in this experiment, consists of four mobile phone categories. Each category contains 34,000 comments about the user's opinions of each product, each comment labeled positive or negative and the rate from one to five. The cleaning pass by several steps as shown in Figure 4, we (a) convert all words to lower case, remove (usernames, mentions, links, repeated characters, numbers, empty tweets, more than two spaces between words, punctuations, and stop words), and (b) converting all words like "isn't" to "is not", tokenization, lemmatization, and stemming are used. Then, all words are joined to the data frame.

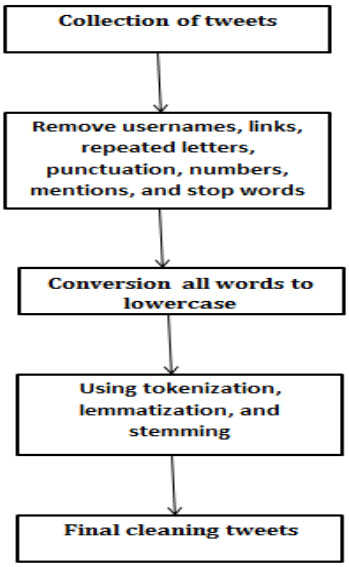

Fig 4. Cleaning tweets steps

B. Phase 2: Features Selection.

After the tweets groups are created, they are analyzed in order to discover the most frequent terms according to the polarity of each term. The most frequent terms are features; these features are extracted using unigrams and bigrams. They can provide useful information about the relative importance data or how features relate to a specific problem. This information can help in filtering the data set and increase the accuracy of our models. Then, we used these features in order to train our algorithms (Naïve Bayes [3], Support Vector Machines [7, 8, 12], Logistic Regression [5,10], Random Forest Regression [6], MaxEntropy [11]) that help us to achieve high accuracy of sentiment classification and prediction models.

C. Phase 3: Tweets Classification.

We split the data with ratio 70:30: 70% for the training step and 30% for the testing. Then, create a vector of Tf-idf and the approach is fed with (a) a parameter as analyzer ="Word", (b) a parameter called stop words="English", and (c) n-grams range for feature selection. We are using both unigrams and bigrams in order to extract the features from the data set. Finally, the machine learning classifiers: NB, MaxEntropy, and SVMs are trained, and the testing data is used to calculate the accuracy of the classifiers.

D. Phase 4: Prediction Process.

After classification each tweet into positive or negative, we use machine learning prediction algorithms (Logistic Regression, Support Vector Regression, Random Forest Regression) in order to predict each product rate from 1:5 depending on user's comments about each product and its classification positive or negative. Finally, the testing data are used to calculate the accuracy of the prediction process by comparing the prediction results rate with the original tweets.

IV. RESULTS

In this section, we present the results when using different features (i.e. unigrams, bigrams and both), and different machine learning techniques: classification and prediction (i.e. Naïve Bayes, MaxEntropy, Support Vector Machines, Logistic Regression, Random Forest Regression).

A. Classification Process

As shown in Figure 5, based on the modification of the preprocessing phase, we have accuracy of 88.18% when Maximum Entropy is used to train the data while it is 85.75% with Naive Bayes and SVM when unigrams is used as a feature selection.

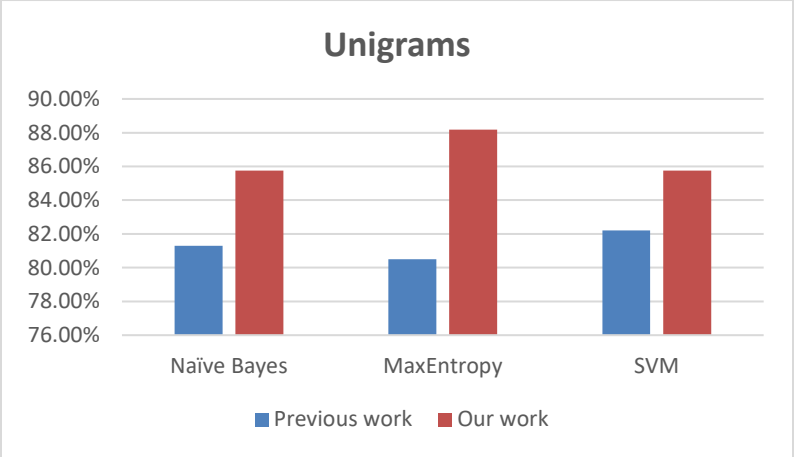

Fig 5. Classifiers results with unigrams

Figure 6 shows a comparison between existing approach and our approach accuracy when bigrams is used as features. The accuracy is 87.15 % when using Naive Bayes and SVM and 87.08% with MaxEntropy.

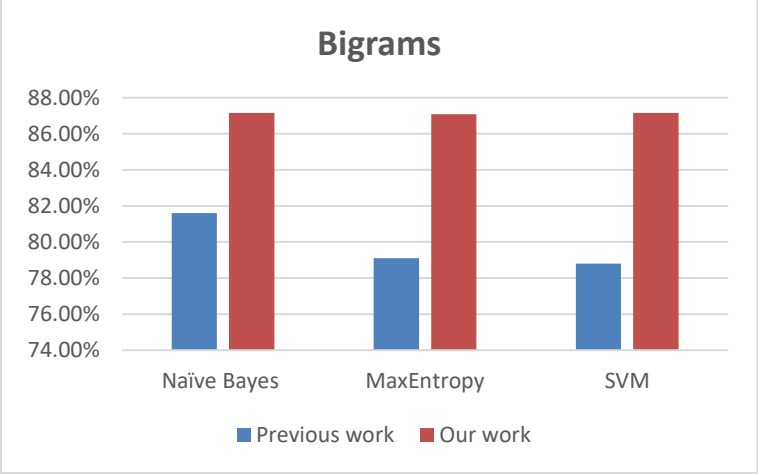

Fig 6. Classifiers results with bigrams

Figure 7 shows the results when both unigrams and bigrams are used as features. The accuracy is improved from (87.15 % to 88.63%) in Naive Bayes and SVM, and from (87.08% to 90.02) in MaxEntropy.

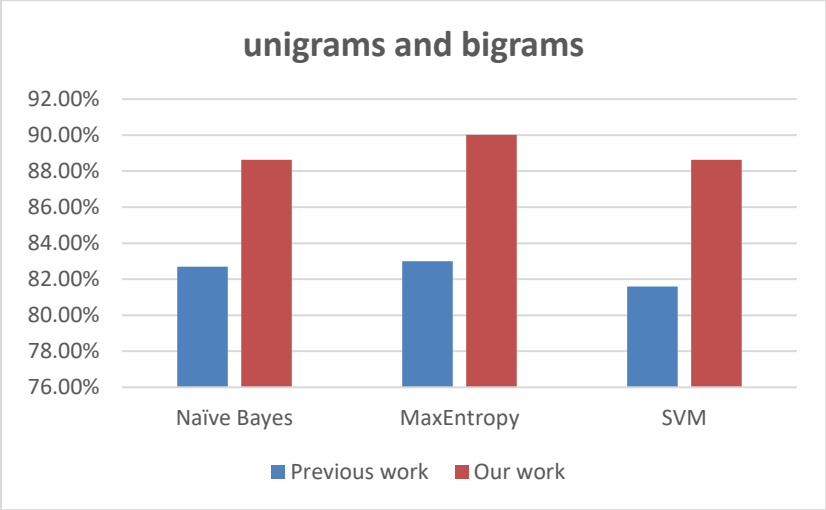

Fig 7. Classifiers results with unigrams and bigrams

A comparison between the existing approach and our approach accuracy is shown in Table 1. The combination includes the classifiers: Naïve Bayes, Support Vector Machine and MaxEntropy, and feature selection using: unigrams, bigrams and both (i.e. unigrams and bigrams).

Table 1. Summary of classifiers accuracy

Features	Naïve Bayes	MaxEntropy	SVM	
Unigrams	81.3%	80.5%	82.2%	**Previous work[4]**
	85.75%	88.18%	85.75%	**Our Work**
Bigrams	81.6%	79.1%	78.8%	**Previous work[4]**
	87.15%	87.08%	87.15%	**Our Work**
Unigrams and Bigrams	82.7%	83.0%	81.6%	**Previous work[4]**
	88.63%	90.02%	88.63%	**Our Work**

B. Prediction Process

In the prediction process, we used mean squared error, and mean absolute error in order to calculate the loss error.

Mean Squared Error formula:

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

Mean Absolute Error formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (2)$$

Where in equation (1) and equation (2), n is the number of data points, y_i present actual value, and $\hat{y}_i$ the present predicted value which returned by the model

1. Logistic Regression

Results of predicting future Blackberry, IPhone, Lenovo, and Samsung rates when using *LR* and extracted features by using both unigrams and bigrams (1, 2) are shown in figure 8.

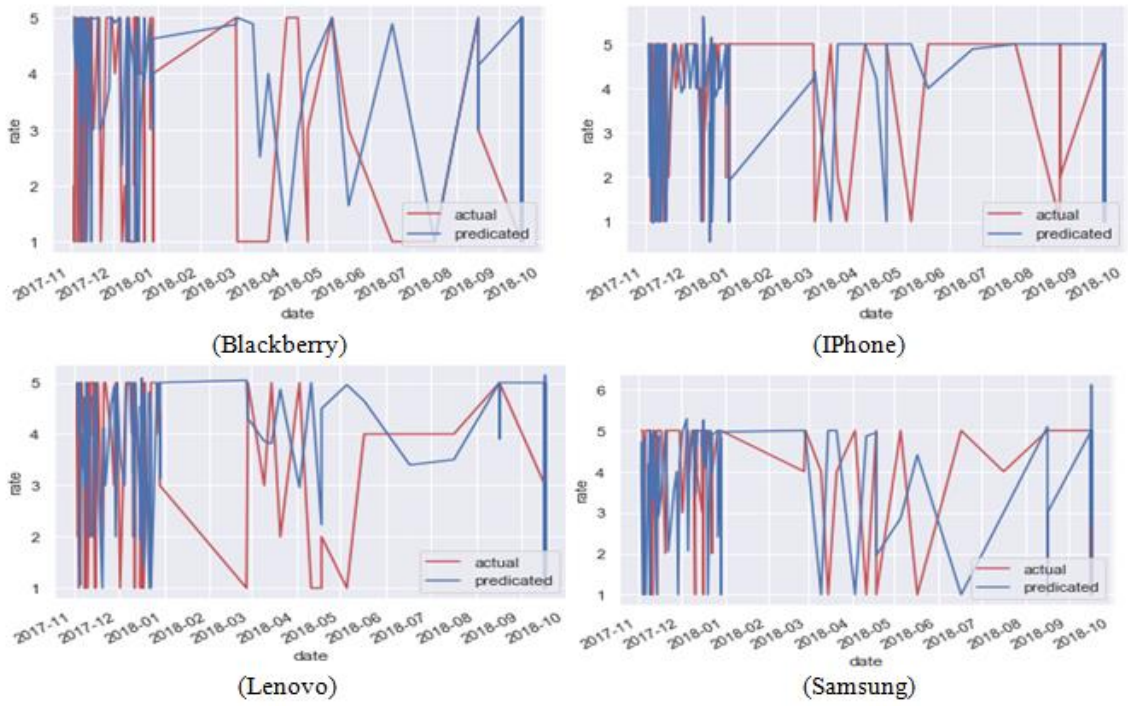

Fig 8. Comparison between actual and predicted rate using Logistic Regression

2. Random Forest Regression

Results of predicting future Blackberry, IPhone, Lenovo, and Samsung rates when using *RFR* and extracted features by using both unigrams and bigrams (1, 2) are shown in figure 9.

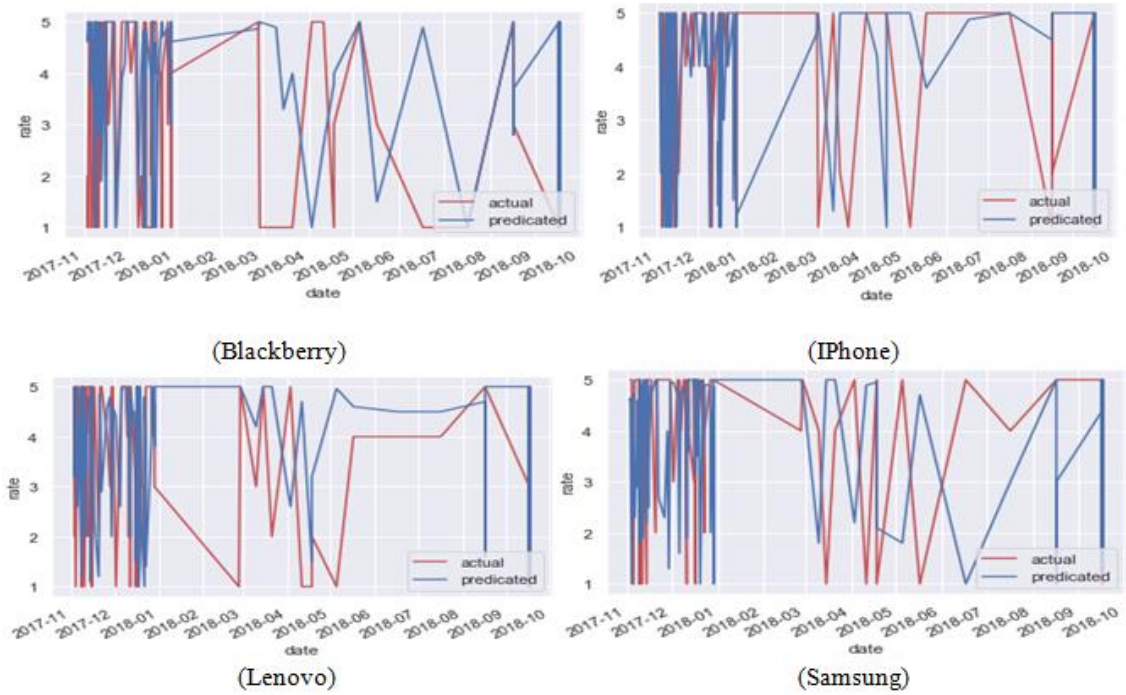

Fig 9. Comparison between actual and predicted rate using Random Forest Regression

3. Support Vector Regression

Results of predicting future Blackberry, IPhone, Lenovo, and Samsung rates when using *SVR* and extracted features by using both unigrams and bigrams (1, 2) are shown in Figure 10.

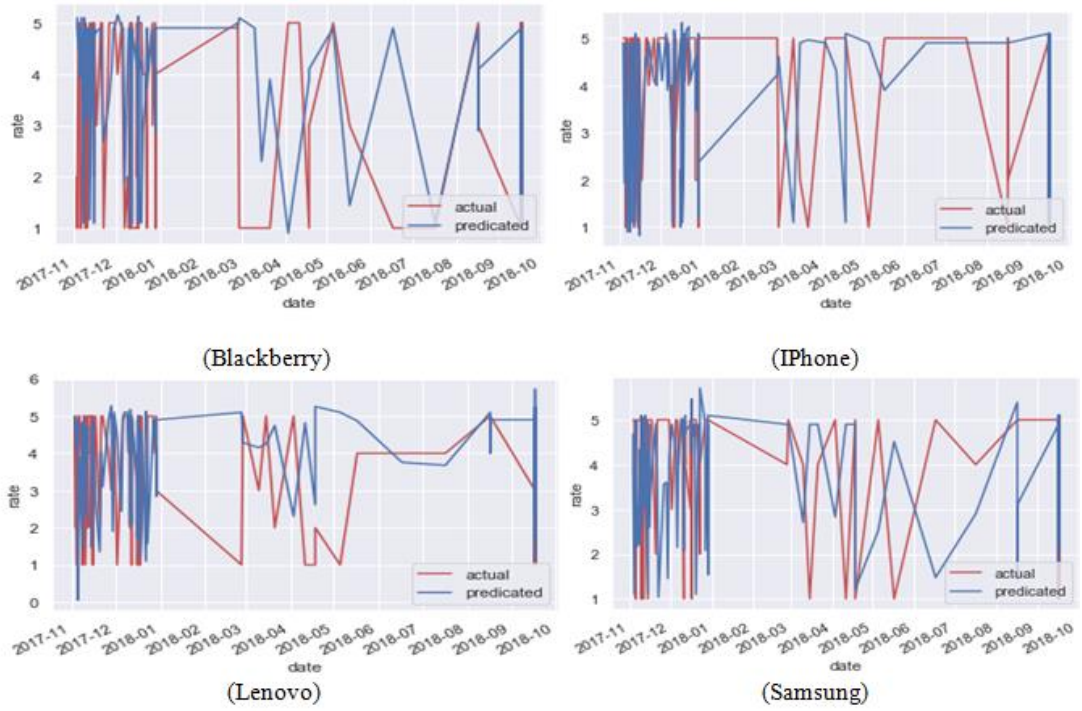

Fig 10. Comparison between actual and predicted rate using Random Forest Regression

C. Summary

In the following, we summaries the results presented in the previous sub-sections.

Prediction of future Blackberry rate: First, in Logistic Regression, we achieve 0.3006 when using mean absolute error, 0.3811 with mean squared error, and 0.0000 with median squared error. Second, in 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. Third, in 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.

Prediction of future IPhone rate: First, in Logistic Regression, we achieve 0.2304 when using mean absolute error, 0.3010 with mean squared error, and 0.0000 with median squared error. Second, in 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. Third, in 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.

Prediction of future Lenovo rate: First, in Logistic Regression, we achieve 0.4047 when using mean absolute error, 0.5216 with mean squared error, and 0.0757 with median squared error. Second, in 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. Third, in 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.

Prediction of future Samsung rate: First, in Logistic Regression, we achieve 0.4211 when using mean absolute error, 0.5765 with mean squared error, and 0.0787 with median squared error. Second,

in 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. Third, in 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.

V. CONCLUSION AND FUTURE WORK

In this work, we implemented an approach for opinion mining in social network data to improve the accuracy of the classification and predicting future products rate according to users' feedback about each product. The feedback is extracted from twitter using application programming interface (API). Then, algorithms such as Naïve Bayes, MaxEntropy, Logistic Regression, Random Forest Regression, and Support Vector Machines are used to classify and predict future products rate. Our improvements are by modifying data preprocessing phase in order to clean it by (a) using (tokenization, stemming and lemmatization), (b) convert all words to lower case, (c) removing usernames, mentions, links, repeated characters, numbers, empty tweets, punctuations, and stop words, (d) all words like isn't are converted to is not to clean the data, and (e) both (unigrams and bigrams) are used to extract the features from the data. The dataset is split with the ratio of 70:30, where 70% used to train the algorithms while 30% is used for testing. In the classification process, the accuracy achieved is 90%. In the prediction process, the Support Vector Regression model can predict future product rate with a mean squared error (MSE) of 0.4122, Logistic Regression can predict with a mean squared error of 0.4986 and Random Forest Regression 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., Ponce, J., Ochoa, A., & Álvarez, M. (2018, March). "Analyzing Students Reviews of Teacher Performance Using Support Vector Machines by a Proposed Model". *In International Symposium on Intelligent Computing Systems* (pp. 113-122). Springer, Cham.

[2] Kaewyong, P., Sukprasert, A., Salim, N., & Phang, F. A. (2015, October). "The possibility of students' comments automatic interpret using lexicon based sentiment analysis to teacher evaluation". *In 3rd International Conference on Artificial Intelligence and Computer Science (AICS2015)* (pp. 179-189).

[3] Zhang, Harry. "The optimality of naive Bayes." *AA* 1.2 (2004): 3.

[4] Colace, F., De Santo, M., & Greco, L. (2014). "SAFE: A Sentiment Analysis Framework for E-Learning". *International Journal of Emerging Technologies in Learning*, 9(6).

[5] Rahman, Nor Azziaty, Abdul, Kian Lam Tan, and Chen Kim Lim. "Supervised and unsupervised learning in data mining for employment prediction of fresh graduate students." *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 9.2-12 (2017): 155-161.

[6] Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell, J. B. (2007). "Random forest models to predict aqueous solubility". *Journal of chemical information and modeling*, 47(1), 150-158.

[7] Suthaharan, Shan. "Support vector machine." *Machine learning models and algorithms for big data classification*. Springer, Boston, MA, 2016. 207-235.

[8] Cristianini, Nello and John Shawe-Taylor, "Support Vector Machines." *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, Cambridge, 2000, pp. 93–124.

[9] Oliveira, N., Costa, J., Silva, C., & Ribeiro, B. (2018, December). "Retweet Predictive Model for Predicting the Popularity of Tweets". *In International Conference on Soft Computing and Pattern Recognition* (pp. 185-193). Springer, Cham.

[10] 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.

[11] Nigam, Kamal, John Lafferty, and Andrew McCallum. "Using maximum entropy for text classification." *IJCAI-99 workshop on machine learning for information filtering*. Vol. 1. No. 1. 1999.

[12] Suthaharan, Shan. "Support vector machine." *Machine learning models and algorithms for big data classification*. Springer, Boston, MA, 2016. 207-235.

[13] 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.

[14] Zarrinkalam, F., Fani, H., Bagheri, E., & Kahani, M. (2017, April). "Predicting users' future interests on Twitter". *In European Conference on Information Retrieval* (pp. 464-476). Springer, Cham.

[15] Esparza, G. G., de-Luna, A., Zezzatti, A. O., Hernandez, A., Ponce, J., Álvarez, M., ... & de Jesus Nava, J. (2017, June). "A sentiment analysis model to analyze students reviews of teacher performance using support vector machines". *In International Symposium on Distributed Computing and Artificial Intelligence* (pp. 157-164). Springer, Cham.