SinkrOn : Jurnal dan Penelitian Teknik Informatika
Volume 4, Number 2, April 2020
DOI : https://doi.org/10.33395/sinkron.v4i2.10549
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Optimizing Genetic Algorithms for Sentiment Analysis of Apple Product Reviews Using SVM

Elly Indrayuni
Universitas Bina Sarana Informatika
Bogor, Indonesia
elly.eiy@bsi.ac.id

Acmad Nurhadi
Universitas Bina Sarana Informatika
Jakarta, Indonesia
achmad.ahh@bsi.ac.id

Submitted: Mar 15, 2020
Accepted: Apr 1, 2020
Published: Apr 15, 2020

Abstract—Online reviews have the potential to provide buyers with insights about products such as quality, performance and recommendations. Website is one of the media that contains information or reviews provided by individuals, groups or organizations about an object or topic, one of which is Apple products. This study analyzes consumer sentiment reviews of Apple product users consisting of 200 reviews which will be classified into positive opinions and negative opinions using the Support Vector Machine algorithm and the application of genetic algorithms (GA) to obtain optimal accuracy values. The stages of this research are, *firstly* collecting a dataset, *the second* is preprocessing data. *Third*, the sentiment analysis process uses SVM and GA as optimization techniques. *Fourth*, do the validation process on the accuracy results obtained using the Confusion Matrix and ROC Curve. The results of this study indicate that Apple product review sentiment analysis produces the best accuracy of 70.00% and AUC 0.924 for SVM algorithm. Whereas the SVM + GA algorithm produces 85.76% accuracy and AUC 0.945, so that the accuracy value increases by 15.76% and the AUC 0.021 on the SVM model when compared before optimization with genetic algorithms (GA) is performed.

Keywords—Sentiment Analysis; Review Product, SVM; GA

I. INTRODUCTION

In recent years, the dramatic increase in smartphone and tablet applications has allowed users to comment on various service platforms at any time through mobile internet, social media, cloud computing, and etc (Zhang et al., 2014). Apple Inc. is a multinational technology company that designs, develops and sells consumer electronics, computer software and online services. Online reviews have the potential to provide buyers with insights about products such as quality, performance and recommendations. Thus providing a clear picture of the product to buyers in the future. The use of online reviews for manufacturers to realize customer needs by analyzing helpful reviews is one of the unrealized potentials. Positive and negative reviews play a big role in determining customer requirements and extracting consumer feedback about products faster (Z. Singla, 2017)). Sentiment analysis via Twitter offers organizations a quick and effective way to monitor public feelings about brands, businesses, their directors, etc. Various features and methods for training the sentiment classification of Twitter data sets have been studied in recent years with mixed results. For each entity extracted (eg iPhone) from a tweet, we add the semantic concept (eg "Apple products") as an additional feature, and measure the correlation of representative concepts with negative / positive sentiments (Saif H., He Y., 2012)Sentiment analysis or opinion mining is a computational study of a person's opinions, behavior and emotions towards an entity. The entity can describe an individual, event or topic (Medhat et al., 2014). Opinion mining does not pay attention to the topic of the text but rather focuses on the expressions depicted

SinkrOn : Jurnal dan Penelitian Teknik Informatika
Volume 4, Number 2, April 2020
DOI : https://doi.org/10.33395/sinkron.v4i2.10549
e-ISSN : 2541-2019
p-ISSN : 2541-044X

from the opinion text. This determines comments in online forums, blogs, or comments related to certain topics (products, books, films, etc.) including positive, negative or neutral opinions. (Martinez et al., 2014). Therefore, sentiment analysis is one solution to overcome the problem of grouping reviews into positive or negative opinions automatically.

Support Vector Machines (SVM) have become a popular classification and regression method for linear and nonlinear problems. This method tries to find the optimal linear separator between data with a maximum margin that allows positive values above the margin values and negative ones below. The specialty of Support Vector Machines comes from the ability to apply linear separation to high-dimensional non-linear data input, and this is obtained by using the necessary kernel functions. The effectiveness of Support Vector Machines is greatly influenced by the type of kernel functions chosen and implemented based on data characteristics (Haddi et al., 2013)). Support Vector Machines have a solid theoretical foundation and do a more accurate classification than most other algorithms in many applications (Moraes et al., 2013). According to Liu in Moraes (Moraes et al., 2013), many researchers have reported that SVM is the most accurate method for classification texts. According to Tsytsarau and Palpanas in (Moraes et al., 2013), SVM is also widely used in sentiment classification. However, SVM has a weakness for the selection of parameters or appropriate features (D. A. Kristiyanti and M. Wahyudi, 2017). Therefore we need the right feature selection to optimize the performance of the SVM algorithm to get the best accuracy.

Genetic Algorithms (GA) is usually used for optimization problems that can take values that are not clear. The main objective of GA is to find optimal solutions in a series of potential solutions. Each set of solutions is called a population. The population consists of vectors, namely chromosomes or individuals (Babaoglu et al., 2010). Genetic Algorithms have the potential to produce better features and become optimal parameters at the same time (Putri, 2015).

In this study the Support Vector Machines (SVM) algorithm is used to classify text, and the application of Genetic Algorithms (GA) optimization techniques to the Support Vector Machines (SVM) algorithm to find the best accuracy by comparing the results of the accuracy of all models applied.

II. LITERATURE REVIEW

Some of the previous studies that have been conducted are the research on sentiment classification that has been done is the document level sentiment classification using the Support Vector Machines algorithm and Artificial Neural Networks (Moraes et al., 2013). The results of this study indicate that Artificial Neural Networks obtain superior results or at least comparable to Support Vector Machines. This research also tells us that there are some limitations of the two models which are rarely discussed in the sentiment of text classification.

The next sentiment classification research is opinion mining for film review using Particle Swarm Optimization Support Vector Machine method (Basari et al., 2013). The results of this study indicate that Particle Swarm Optimization can improve the level of accuracy by selecting the best parameters. The value of accuracy produced in this study increased from 71.87% to 77%.

Research conducted by (D. A. Kristiyanti and M. Wahyudi, 2017) regarding opinion mining or analysis of cosmetic product sentiment reviews by selecting Genetic Algorithms and Particle Swarm Optimization and Main Component Analysis features. In this study, researchers made improvements to previous studies using a combination of methods in selecting features in SVM by comparing three feature choices; Genetic Algorithms, Optimization of Particle Flocks, and Principal Component Analysis to determine which one of the best feature choices increases the accuracy of SVM classification. The results of this study were obtained the accuracy of the SVM algorithm with an average of 82.00% and an average AUC of 0.988. After integration of the SVM algorithm and feature selection, the Genetic algorithm showed the best results with an average accuracy of 94.00% and an average of AUC of 0.984. Particle Swarm Optimization shows the best results with an average accuracy of 97.00% and an average AUC of 0.988. While the Principal Component Analysis shows the best results with an average accuracy of 83.00% and an average AUC of 0.809. Thus in this study it can be concluded that SVM has increased the best accuracy of feature selection integrated with Particle Swarm Optimization with an increase in accuracy from 82.00% to 97.00%.

III. PROPOSED METHOD

This study analyzes the sentiment words contained in Apple product user reviews. The method proposed in this study uses the Support Vector Machine algorithm which is then followed by the

SinkrOn : Jurnal dan Penelitian Teknik Informatika
Volume 4, Number 2, April 2020
DOI : https://doi.org/10.33395/sinkron.v4i2.10549
e-ISSN : 2541-2019
p-ISSN : 2541-044X

application of GA on SVM to optimize the accuracy results. *First*, it begins with collecting a dataset of 200 reviews consisting of 100 reviews for positive opinions and 100 reviews for negative opinions taken from the site www.trustpilot.com. *Second*, the preprocessing stage includes tokenization, stopword filter. and generate n-grams so that text that is noise or HTML tags, symbols or punctuation is removed. In this study, the N-gram used was bigram, N = 3. *Third*, the sentiment analysis process is carried out on every word contained in the product review followed by the training process and testing the model using Rapidminer. The sentiment analysis process in this study is to determine a review as a member of a positive opinion class or a negative opinion class based on the weighting calculation value of the SVM formula which is 0. If the value of the weight is greater than 0 then it is classified into the positive opinion class and if the value weights less than 0 can be classified into negative opinion classes. The research began by conducting data testing experiments using the SVM method to produce the best accuracy value. Then the second model continues with the application of GA on SVM, which begins with the determination of the value of the Genetic Algorithm with the population size value that is at the default value, then continues with a multiple of 5 to the population size value which results in the highest accuracy taken for further research. After that, determine the value of *Population Size*, *P Crossover*, *P Generate* to get the best model. *Finally*, do the validation process on the accuracy results obtained using the Confusion Matrix and ROC Curve. Then conclude the best algorithm model for analyzing Apple product review sentiment based on the best accuracy value produced. The stages of this research method can be seen in Figure 1.

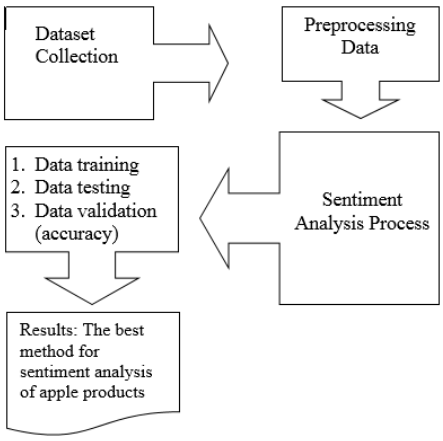

Figure 1. Stages of Research

IV. Result and Discussion

The training data used in this text classification consists of 200 review data taken from the www.trustpilot.com website. Training data collection is done by copying each review to notepad so as to produce a file with a .txt extension. After that the file is separated, the positive opinion review data is put in a positive folder, the negative opinion review data is put in the negative folder. Before conducting experiments and testing data in this study, the preprocessing stage is tokenization, stopword filter, and generate n-gram.

The core in the process of classifying product reviews is determining a review as a member of a positive opinion class or a negative opinion class. In this sentiment classification several words are used as attributes to determine the product review data, including the category of positive opinion or negative opinion. Attributes that represent positive opinions include best, amazing, love, like, excellent, perfectly, friendly, and good. While the attributes that represent negative opinions are bad, awful, disappointed, horrible, terrible, worst, useless, and complained.

4.1. Sentiment Text Classification Phase Using SVM

The value of training cycles in the experiment using the Support Vector Machine algorithm is determined by conducting a trial by entering the C and epsilon values in the SVM parameter. The results of the experiments that have been conducted to determine the value of the experimental training cycles can be seen in Table I.

TABLE I. EXPERIMENT DETERMINATION TRAINING CYCLES VALUE (SVM)

Parameter		*SVM*	
C	*Epsilon*	*Accuracy*	*AUC*
0.0	0.0	69.50%	0.922
0.0	**0.5**	**70.00%**	**0.924**
0.5	0.5	70.00%	0.923
0.5	1.0	50.00%	0.500
1.0	0.0	69.50%	0.922
1.0	1.0	50.00%	0.500

Based on experiments that have been done, the determination of the parameter values for C and epsilon can affect the resulting accuracy values. The highest accuracy value in the study with the Support

SinkrOn : Jurnal dan Penelitian Teknik Informatika
Volume 4, Number 2, April 2020
DOI : https://doi.org/10.33395/sinkron.v4i2.10549
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Vector Machine algorithm is 70.00% and the AUC value is 0.924 with the determination of the value of C = 0.0 and epsilon = 0.5. To find out the accuracy, sensitivity, specificity, ppv and npv values are generated from true positive, false negative, true negative, and false positive values through the confusion matrix table in Table II.

TABLE II. CONFUSION MATRIX MODEL SVM

Accuracy : 70.00%			
	True Positive	True Negative	Class Precision
Positive Prediction	98	58	62.82%
Negative Prediction	2	42	95.45%
Class Recall	98.00%	42.00%	

Table II is a confusion matrix table that shows that of the 200 reviews that were processed using Rapidminer there were 98 reviews that were true positive (tp), and as many as 2 reviews were false negative (fn). Next there are 42 reviews that are true negative (tn) and 58 reviews as false positive (fp). To find out the accuracy, sensitivity, specificity, ppv and npv values, the results can be seen in Table VI.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$= \frac{98 + 42}{98 + 42 + 58 + 2} = 0.7$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$= \frac{98}{98 + 2} = 0.98$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$= \frac{42}{42 + 58} = 0.42$$

$$ppv = \frac{TP}{TP + FP} \quad (4)$$

$$= \frac{98}{98 + 58} = 0.6282$$

$$npv = \frac{TN}{TN + FN} \quad (5)$$

$$= \frac{42}{42 + 2} = 0.9545$$

TABLE III. VALUE OF ACCURACY, SENSITIVITY, SPECIFICITY, PPV, AND NPV (SVM)

	% (in percent)
Accuracy	70.00
Sensitivity	98.00
Specificity	42.00
Ppv	62.82
Npv	95.45

Table III is the result of calculating the accuracy, sensitivity, specificity, ppv and npv values using the formula. It can be seen that the calculation results are valid according to the confusion matrix table produced on Rapidminer.

Figure 2 is the ROC curve from processing 200 training data using SVM algorithm with AUC value of 0.924 including Excellent Classification.

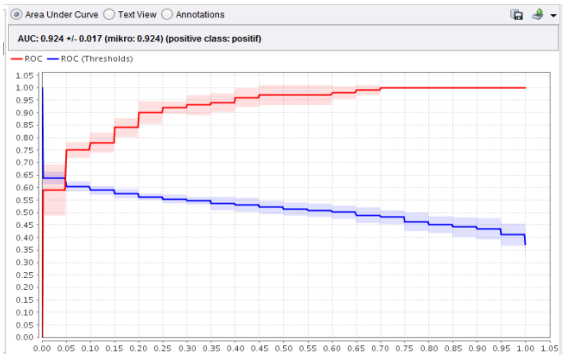

Figure 2. ROC Curve of SVM

4.2. Sentiment Text Classification Phase Using SVM + GA

Genetic Algorithm (GA) is used as optimization in the sentiment classification process. The difference in the Support Vector Machine algorithm model based on Genetic Algorithm (GA) is the process of producing to enter the value of *Population Size*, *P Crossover* and *P Generate* which will affect the resulting accuracy value. In this experiment the value of *P crossover = 0.9* and *P generate* = 1.0 are used. The results of the experiments that have been conducted to determine the value of this experimental training cycle can be seen in Table IV.

SinkrOn : Jurnal dan Penelitian Teknik Informatika
Volume 4, Number 2, April 2020
DOI : https://doi.org/10.33395/sinkron.v4i2.10549
e-ISSN : 2541-2019
p-ISSN : 2541-044X

TABLE IV. EXPERIMENT DETERMINATION TRAINING CYCLES VALUE (SVM + GA)

Parameter			SVM + GA	
C	Epsilon	*Population Size*	*Accuracy*	AUC
0.0	0.0	5	77.26%	0.888
0.0	0.5		79.16%	0.893
0.5	0.5		78.29%	0.816
0.5	1.0		50.29%	0.500
1.0	0.0		75.66%	0.846
1.0	1.0		50.29%	0.500
0.0	0.0	10	79.21%	0.896
0.0	0.5		82.24%	0.898
0.5	**0.5**		**85.76%**	**0.945**
0.5	1.0		50.29%	0.500
1.0	0.0		82.13%	0.876
1.0	1.0		50.29%	0.500
0.0	0.0	15	77.21%	0.872
0.0	0.5		84.74%	0.908
0.5	0.5		82.68%	0.897
0.5	1.0		50.29%	0.500
1.0	0.0		85.16%	0.907
1.0	1.0		50.29%	0.500

Based on experiments that have been done, the addition of optimization techniques with Genetic Algorithm on the Support Vector Machine algorithm affects the resulting accuracy. C values, epsilon and population size entered produce different accuracy values. The highest accuracy value in the study with the Support Vector Machine algorithm based on Genetic Algorithm (GA) is 85.76% and the AUC value is 0.945 with the determination of C = 0.5, epsilon = 0.5 and population size of 10.

To find out the accuracy, sensitivity, specificity, ppv and npv values can be generated from true positive, false negative, true negative, and false positive values through the GA-Based Support Vector Machine algorithm in Table V.

TABLE V. Confusion Matrix Model SVM + GA

Accuracy : 85.76%			
	True Positive	True Negative	Class Precision
Positive Prediction	95	25	79.17%
Negative Prediction	3	74	96.10%
Class Recall	96.94%	74.75%	

Table VIII shows that the number of true positive (tp) is 95 opinions, false negative (fn) is 3 opinions. Next are 74 opinions for true negative (tn) and 25 opinions for false positive (fp). The accuracy, sensitivity, specificity, ppv and npv values of the results can be seen in Table IX.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$= \frac{95 + 74}{95 + 74 + 25 + 3} = 0.8576$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$= \frac{95}{95 + 3} = 0.9694$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$= \frac{74}{74 + 25} = 0.7475$$

$$ppv = \frac{TP}{TP + FP} \quad (4)$$

$$= \frac{95}{95 + 25} = 0.7917$$

$$npv = \frac{TN}{TN + FN} \quad (5)$$

$$= \frac{74}{74 + 3} = 0.9610$$

TABLE VI. VALUE OF ACCURACY, SENSITIVITY, SPECIFICITY, PPV, AND NPV (SVM + GA)

	% (in percent)
Accuracy	85.76
Sensitivity	96.94
Specificity	74.75
Ppv	79.17
Npv	96.10

Table VI is the result of calculating the accuracy, sensitivity, specificity, ppv and npv values using the formula. It can be seen that the calculation results are valid according to the confusion matrix table produced by Rapidminer for the Support Vector

SinkrOn : Jurnal dan Penelitian Teknik Informatika
Volume 4, Number 2, April 2020
DOI : https://doi.org/10.33395/sinkron.v4i2.10549
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Machine algorithm model based on Genetic Algorithm (GA).

From the experimental processing of 200 training data using the Support Vector Machine algorithm using optimization techniques, namely Genetic Algorithm (GA), it produces an AUC value of 0.945 which includes Excellent Classification. From the results of experimental processing using the Support Vector Machine algorithm model only, the AUC value on the Support Vector Machine algorithm model based on Genetic Algorithm (GA) has increased. Display the ROC curve for the SVM algorithm model based on GA as follows:

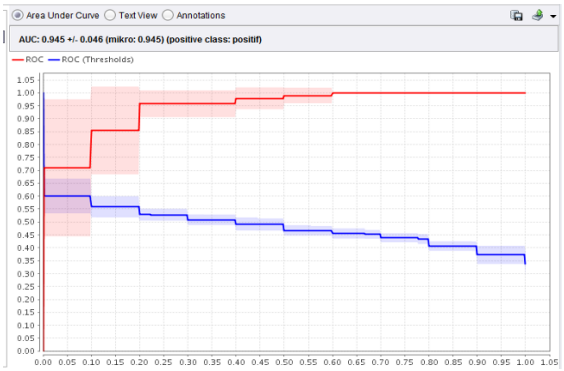

Figure 3. ROC Curve of SVM + GA

Figure 3 is the ROC curve of the experimental results of processing 200 training data using the Support Vector Machine algorithm by using an optimization technique namely Genetic Algorithm (GA) with an AUC value of 0.945 which includes Excellent Classification. From the results of experimental processing using the Support Vector Machine algorithm model only, the AUC value on the Support Vector Machine algorithm model based on Genetic Algorithm (GA) has increased.

4.3. Analysis of Model Evaluation and Validation Results

After the classification of Apple product reviews has been successfully grouped into categories of positive opinions and negative opinions, the resulting accuracy is tested to see the performance of the experimental results with the two algorithm models applied. Based on evaluations using confusion matrix and ROC curve it is proven that the application of optimization techniques using Genetic Algorithm (GA) on the Support Vector Machine (SVM) algorithm can increase the resulting accuracy. The results of applying the Support Vector Machine (SVM) algorithm model using Genetic Algorithm (GA) have a higher accuracy value compared to the Support Vector Machine (SVM) algorithm alone.

The accuracy value for the Support Vector Machine algorithm model for classification of Apple product review sentiments is 70.00% with an AUC value of 0.924. While the accuracy value for the Support Vector Machine algorithm model with Genetic Algorithm (GA) results in a higher accuracy value of 85.76% and AUC value of 0.945. Then it can be stated that the difference in the accuracy value of the two algorithm models applied in this study has a difference in the accuracy value of 15.76% and the difference in the AUC value of 0.021. For the results of testing all algorithms in detail can be seen in Table X.

TABLE VII. COMPARISON OF ACCURACY RESULTS (SVM AND SVM + GA)

	SVM	SVM + GA
Accuracy	70.00%	85.76%
AUC	0.924	0.945

Table VII shows the results of testing the two algorithms applied in this study. When compared as a whole, the Support Vector Machine algorithm using Genetic Algorithm (GA) is the method that has the best performance with an accuracy value reaching 85.76% and AUC 0.945, including Excellent Classification

V. CONCLUSION AND SUGGESTION

The results of tests conducted for Apple product reviews using the Support Vector Machine algorithm as a popular algorithm are proven to produce a high accuracy value of 70.00% and an AUC value of 0.924. The AUC value generated in this algorithm includes Excellent Classification. In the next experiment an optimization technique was applied using Genetic Algorithm (GA) on the Support Vector Machine algorithm which was proven to increase the accuracy value by 15.76% with a difference in AUC value of 0.021. The accuracy value after the application of the Genetic Algorithm (GA) optimization technique becomes 85.76% with an AUC value of 0.945 and is included in the Excellent Classification. Therefore, the application of Genetic Algorithm (GA) optimization techniques to the Support Vector Machine algorithm is the best algorithm testing model in text classification. For further research it is hoped that other researchers can develop this research using methods, algorithms and other optimization techniques such as Chi Square to compare the resulting accuracy values.

VI. REFERENCES

Babaoglu, I., Findik, O., & Ülker, E. (2010). A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine. *Expert Systems with Applications*, *37*(4), 3177–3183. https://doi.org/10.1016/j.eswa.2009.09.064

Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. *Procedia Engineering*, *53*, 453–462. https://doi.org/10.1016/j.proeng.2013.02.059

D. A. Kristiyanti and M. Wahyudi. (2017). Feature selection based on Genetic algorithm, particle swarm optimization and principal component analysis for opinion mining cosmetic product review. *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, *1–6*.

Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *First International Conference on Information Technology and Quantitative Management*, *17*, 26–32. https://doi.org/10.1016/j.procs.2013.05.005

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*. https://doi.org/10.1016/j.asej.2014.04.011

Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621–633. https://doi.org/10.1016/j.eswa.2012.07.059

Putri, D. A. (2015). Algoritma Support Vector Machine Berbasis Algoritma Genetika Untuk Analisis Sentimen Pada Twitter. *Konferensi Nasional Ilmu Pengetahuan Dan Teknologi (KNIT)*, 1–8.

Saif H., He Y., A. H. (2012). Semantic Sentiment Analysis of Twitter. *Cudré-Mauroux P. et Al. (Eds) The Semantic Web – ISWC 2012. ISWC 2012*, *7649*, 508–524.

Z. Singla, S. R. and S. J. (2017). Statistical and sentiment analysis of consumer product reviews. *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6.

Zhang, L., Hua, K., Wang, H., Qian, G., & Zheng, L. (2014). Sentiment analysis on reviews of mobile users. *Procedia Computer Science*, *34*, 458–465. https://doi.org/10.1016/j.procs.2014.07.013