Opinion classification on a social network by a novel feature selection technique

Atchara Choompol, Panida Songram, Phattahanaphong Chomphuwiset
POLAR Lab, Department of Computer Science, Faculty of Informatics, Mahasarakham University, Thailand

Article Info

Article history:

Received Feb 10, 2020
Revised Apr 8, 2020
Accepted Apr 21, 2020

Keywords:

Association rule
Feature selection
Opinion classification
Opinion mining
Social network

ABSTRACT

Most of the opinion comments on social networks are short and ambiguous. In general, opinion classification on the comments is difficult because of lacking dominant features. A feature extraction technique is therefore necessary for improving accuracy of the classification and computational time. This paper proposes an effective feature selection method for opinion classification on a social network. The proposed method selects features based on the concept of a filter model, together with association rules. Support and confidence are used to calculate the weights of features. The features with high weight are selected for classification. Unlike supports in association rules, supports in our method are normalized to 0-1 to remove outlier supports. Moreover, a tuning parameter is used to emphasize the degree of support or confidence. The experimental results show that the proposed method provides high classification efficiency. The proposed method outperforms Information Gain, Chi-Square, and Gini Index in both computational time and accuracy

Corresponding Author:

Atchara Choompol,
Department of Computer Science,
Mahasarakham University, Thailand.
Email: atchara.cho@msu.ac.th

1. INTRODUCTION

Social networking websites have nowadays become important communication tools. They allow users to share opinions and discuss various issues through message formats (such as posts and comments). Opinion messages on social networks are important information and analyzed for useful in many applications. For example, opinion messages are used to track consumers' attitudes toward products or services. Moreover, they are used to identify the satisfactions of demographics features with particular products [1, 2]. In politics, opinion messages are used for electoral predictions or to make a survey of people's opinions about political parties [3]. In education, they are exploited to analyze student sentiments to improve the efficiency of studying [4].

In opinion classification, features are extracted from messages (texts) and then learning algorithms will determine the orientation of opinions from the features. Since most of the messages on social networking websites are short and vague, features are resulted in a large feature space that contains irrelevant and redundant features for classification. The irrelevant and redundant features lead to misclassification in opinion classification task. Therefore, feature selection becomes an important phase in the task. It selects relevant features to increases the performance of the classification [5]. Various feature selection methods have been proposed for opinion classification [6-9]. The filter model is one popular feature selection model. The idea of filter model is to calculate the weight of features and then features are decided to keep or remove from determination the weight of features. The filter model is simple and effective [10].

In this paper, a feature selection method is proposed to improve the performance of opinion classification on social network. It processes based on filter model. Unlike the previous feature selection methods, confidence and support values in association rule mining are applied to calculate weight of features. In the proposed method, the confidence value indicates a percentage of class c_i among feature $f_{j.}$ The support value indicates frequency of feature f_j in class c_i. The support is normalized to 0-1 to remove outlier support. Moreover, the balance of support and confidence values are adjusted by the parameter *p*. In addition, the vertical data format is used to easily calculate the support and confidence values. Then the time for computing weights of features can be improved.

In conclusion, the contributions of this paper are the following.

a) Support and confidence values are applied to calculate weights of features that are used to select most relevant features. The selected features improve accuracy when comparing to Information Gain, Chi-Square, and Gini Index.
b) The support is normalized to 0-1 to remover outlier support. Then relevant features will be found and lead to improve classification.
c) A tuning parameter *p* is proposed to adjusted balance between support and confidence.
d) The weight of feature is easily calculated by using vertical data format that can improve the computation time

The rest of the paper is organized as follows. Related works are mentioned in Section 2. Section 3 describes the proposed method. The experimental evaluations are discussed in Section 4. Finally, Section 5 provides conclusions and future work.

2. RELATED WORKS

Currently, social network websites have become data sources for researchers because the amount of data on social networking sites has grown enormously, especially opinion messages. Opinion classification is a text mining task that try to find orientation of opinions from opinion messages. It is widely applied in many research domains, such as restaurant reviews, product reviews, movie reviews etc. Due to unstructured data, opinion messages are transformed to feature space. Most of the opinion messages on social networks are short, ambiguous and have non-dominant features. Therefore, the feature space is very large. Feature selection becomes an important process to decrease the number of features for improving the accuracy and reducing the computational burden.

Feature selection methods have been studied in many researches. For example, Alhaj *et al.* [6] presented a two-tier feature selection method to select appropriate and significant features. The subset of features is ranked based on high information gain entropy in the first tier. Then features are extended with high ability in the second tier. The analysis results showed that the selected features gave high clustering accuracy. Parla and Ozel [7] proposed a new feature selection method, called Query Expansion Ranking. The method is based on query expansion term weighting methods. The results showed that Query Expansion Ranking could improve sentiment analysis performance in terms of classification accuracy and computational time.

Pratiwi and Adiwijaya [8] proposed feature selection and classification based on Information Gain for sentiment analysis. The method performed on a movie review dataset and showed that it could reduce more than 90% of unnecessary features with 96% accuracy. Yang *et al.* [9] proposed a new feature selection algorithm based on comprehensive measurements, both inter-category and intra-category, for text categorization. Three benchmark document collections, 20-Newsgroups, Reuters-21578 and WebKZ, were classified by Naïve Bayes and Support Vector Machines. The experimental results showed that the feature selection method is significantly superior to other methods. Adeleke *et al.* [5] proposed a two-step feature selection method. In the first step, Chi-square was adopted to reduce the dimensionality of a feature set. In the second step, a wrapper correlation-based technique was employed to further select most relevant features from the reduced feature set. The results shown that the feature selection method achieved accuracy of 93.60% in 4.17 seconds. Somantri and Apriliani [11] proposed a hybrid feature selection model to solve the non-optimal process of selecting features. Hybrid feature selection models combine Information Gain and a genetic algorithm. The results showed that the proposed method gave an accuracy of 93.00%.

Rafei *et al.* [12] compared the performance of two feature selection techniques for select the relevant features for classifying biomedical text abstracts. The two feature selection techniques, Pearson's Correlation and Information Gain, are investigated for reducing the high dimensionality of data. Stroke documents were classified by Support Vector Machine. The experimental results showed that Information Gain outperformed Pearson's Correlation by 3.3%. Purnamasari [13] classified tweets that contain bullying by using Support Vector Machine. Relevant features are selected by using Information Gain. In the first step, tweets are preprocessed by using tokenizing, filtering, stemming and term weighting. In the second step, Information Gain feature selects relevant features by calculating the entropy value of each features. After that, the classification process is performed by Support Vector Machine classifier. The results showed that the best threshold of information gain is 90% with accuracy 76.66%.

From previous work, feature selection is an important process in text classification. Opinion classification is a kind of text classification that needs to reduce dimension of feature space and select relevant features. Most feature selection methods calculate the weights of features based on relationships within features and classes. Unlike the previous methods, the frequency of the relationship of features and classes is added to calculate the weight of features in our method. Furthermore, most feature selection on opinion classification retrieve relevant features from a horizontal vector, where a document is mapped to a row. The row consists of a huge number of all possible features. If a document has a small subset of the features, a large null value will be generated which has a negative impact on computational performance. In our proposed method, a vertical data format is employed to reduce computation cost.

3. THE PROPOSED METHOD

In this paper, a feature selection method is proposed to select most relevant features and improve performance of opinion classification on social network. Due to opinion classification processing on text dataset, text dataset has to be preprocessed to structured dataset. The proposed method represents the dataset as a vertical data format to easily calculate support and confidence values. The preprocessing is explained in subsection 3.1. From the vertical dataset, it will be processed to find most relevant features. The process in the proposed method consist of 4 phases; support and confidence calculation, support normalization, parameter tuning and weighing calculation, and feature ranking. All phases will be explained more details in the subsections 3.2 - 3.5.

3.1. Preprocessing

The preprocessing is the following steps: (1) text fragments, such as #, emoticons, URLs and @, are removed from the text dataset because they do not significantly designate the polarity identification, (2) stop words are removed from the dataset using a dictionary-based technique, (3) the stemming process is performed before tokenization and in the tokenization process words preceded by “no/not” are tokenized using the bi-gram technique, otherwise, uni-gram. A token is considered as a feature. Next, the dataset is transformed to the vertical dataset as an example in Table 1, where $D = \{d_1, d_2, .., d_5\}$ is the set of documents , $T = \{t_1, t_2, .., t_6\}$ is the set of features, and $C = \{c_1, c_2,\}$ is the complete set of distinctive class labels.

Table 1. Vertical data format

Feature	Set of Documents
t_1	$\{d_1, d_3, d_5\}$
t_2	$\{d_1, d_2, d_3\}$
t_3	$\{d_1, d_3, d_4\}$
t_4	$\{d_1, d_2, d_3, d_5\}$
t_5	$\{d_1, d_5\}$
t_6	$\{d_2, d_3, d_4, d_5\}$
c_1	$\{d_1, d_2, d_3\}$
c_2	$\{d_4, d_5\}$

3.2. Support and confidence calculation

Supports and confidences values of all features are easily calculated for each class in the vertical dataset. To calculate a confidence value of feature t_i in class c_k, we need to find the support of feature t_i and the support of feature t_i in class c_k. The support of feature t_i is the number of documents containing t_i that is easily obtained by counting the number of documents containing the feature t_i in the vertical dataset. For example the support of $t_3 = |\{d_1, d_3, d_4\}| = 3$, denote as $S(t_3)$.

The support of feature t_i in class c_k is the number of documents containing t_i in c_k, denoted as $|S(t_i, c_k)|$. In our work the support of feature t_i in class c_k is easily calculated from $|S(t_i,c_k)|) = |S(t_i)\cap S(c_k)|$. For example, the support of feature t_3 in class c_1 can be calculated from $|S(t_3,c_1)|) = |S(t_3)\cap S(c_1)\ | = |\{d_1, d_3, d_4\}\cap|\{ d_1, d_2, d_3\}| = |\{ d_1, d_3\}| = 2$.

As a result, the confidence is easily calculated from the calculated support. The confidence is the ratio of the number of documents that contain feature t_i in class c_k and the number of documents contain feature t_i. Therefore, the confidence is calculated from $C(t_i,c_k\ ) = |S(t_i,c_k)| / |S(t_i)|$. The confidence of feature t_3 in class c_1 can be found from $C(t_3,c_1\ ) = |S(t_3,c_1)| / |S(t_3)| = 2/3 = 0.667$. This states that if feature t_3 is in a document, the probability of the document belonging to class c_1 is 66.70%.

3.3. Support normalization

Since the support may be very small or very large, it needs to be normalized to remove outlier support. In the proposed method, the support is normalized to 0-1, the same as the unit value of confidence. In the normalization of the support process, all features in class c_k are ranked by their supports in descending order. For example, in Table 2, class c_1 consists of features t_1, t_2, t_3, t_4, t_5, and t_6. The supports of all features in class c_1 are shown in the second row ($S(t_i, c_1)$). The supports of t_2 and t_4 are the highest, so t_2 and t_4 are ranked in the first order. The second highest supports are the supports of t_1, t_3, and t_6, so t_1, t_3, and t_6 are ranked in the second order. The support of t_5 is the lowest support, so t_5 is ranked in the third order. The ranking values are shown in the third row ($R(t_i, c_1)$). After ranking values found, the normalization of support can be calculated from $NS(t_i,c_k) = R(t_i, c_k) / N$, where N is the number of all features. For example, $NS(t_3,c_1) = R(t_3, c_1) / N = 2/6 = 0.33$.

Table 2. The ranking values of features

Feature (t_i)	t_1	t_2	t_3	t_4	t_5	t_6	Class (c_k)
$S(t_i, c_1)$	2	3	2	3	1	2	c_1
$R(t_i, c_1)$	2	1	2	1	3	2	c_1

3.4. Parameter tuning and weighting calculation

First, the weight of each feature in c_k is calculated based on its normalized support and confidence. A tuning parameter (p) is introduced to balance the significance of the normalized support and confidence. p is a constant value ($0 < p < 1$). If support is more significant than confidence, p is more than 0.5, and less than 0.5 otherwise.

The weight of feature t_i in class c_k can be evaluated using $w(t_i,c_k) = p \times NS(t_i,c_k) + (1\text{-}p) \times C(t_i,c_k)$, where $w(t_i,c_k)$ is the weight feature t_i in class c_k,. For example, if $p = 0.9$, the weight of feature t_3 in c_1 and c_2 are calculated as follows.

$w(t_3,c_1) = 0.9 \times NS(t_3,c_1) + (1\text{-}0.9) \times C(t_3,c_1) = (0.9 \times 0.33) + (1\text{-}0.9) \times 0.667 = 0.427$

$w(t_3,c_2) = 0.9 \times NS(t_3,c_2) + (1\text{-}0.9) \times C(t_3,c_2) = (0.9 \times 0.33) + (1\text{-}0.9) \times 0.50 = 0.347$

Finally, the final weight of feature is decided by the maximum value of the weight of feature t_i in class c_k, $W(t_i) = \max(w(t_i,c_k))$. For example, the final weight of feature $t_3 = W(t_3) = \max\{w(t_3, c_1), w(t_3, c_2)\} = \max\{0.427, 0.347\} = 0.427$.

3.5. Feature ranking

All features in dataset are ranked by their final weights in descending order. The feature with the highest weight is the first rank and means that it is the most relevant features. The set of most relevant features can be selected from the ranked features.

4. EXPERIMENTAL EVALUATIONS

4.1. Experimental setup

The experimental datasets were collected from twitter, and comprised 10,000 instances from Standford twitter sentiment data (STS) [14], 4,000 instances from SemEval-2017 Task4A dataset (SemEval) [15], 2,600 instances from Sentiment strength twitter dataset (SS-Tweet) [16] and 1,000 instances from HCR Twitter dataset [17]. All datasets are preprocessed as in subsection 3.1. To perform feature section using Information Gain (IG), Chi-Squared (Chi2), and Gini Index (Gini), the datasets are transformed in vector space model. A document is transformed into a vector. For each vector, the value of features is 1, if the feature occurs in the document, otherwise, 0. All datasets have two classes, positive and negative. Finally, the characteristics of datasets are shown in Table 3.

Table 3. The characteristics of the datasets

Data set	Number of comments	Label		Number of features
		Positive Class	Negative Class	
1. STS	10,000	5,000	5,000	12,772
2. SemEval	4,000	2,000	2,000	9,065
3. SS-Twitter	2,600	1,300	1,300	6,845
4. HCR	1,000	500	500	1,867

For evaluating the proposed method against IG, Chi2, and Gini, the performance of classification and computation time are investigated in our experiments. For investigating performance of classification, the number of selected features is varied from 10% to 90%. Then they are investigated to classify opinion orientation by using Naïve Bayes that is a simple classifier and effective in opinion classification [18-25]. 10-fold cross-validation is used to divide dataset the experiments. The average accuracy and average F-measure for each class are used to evaluate the proposed method against IG, Chi2, and Gini. The computation time is evaluated from weighting calculation and features ranking

4.2. Experimental results and discussion

First, the values of *p* are investigated to find the best values for giving highest accuracy. In Table 4, *p* = 0.8 gives the highest accuracy on SemEval and HCR and *p* = 0.9 gives the highest support on STS and SS-Tweet. Moreover, the result is shown that the accuracy is increased when the *p* value is increased. It means that the normalized support is more important than confidence in opinion classification.

Table 4. Accuracy of the classification for each *p*

p value	Dataset			
	STS	SemEval	SS-Tweet	HCR
p = 0.1	58.93	77.71	55.04	73.41
p = 0.2	60.19	80.52	55.49	73.91
p = 0.3	62.58	83.83	56.47	75.64
p = 0.4	66.15	85.87	58.04	77.30
p = 0.5	69.30	86.82	59.33	78.98
p = 0.6	71.14	87.22	60.60	80.04
p = 0.7	72.10	87.43	61.02	80.20
p = 0.8	72.23	87.44	61.41	80.44
p = 0.9	72.28	87.41	61.42	80.22

The performances of classification on four datasets are reported in Table 5 to Table 8. In Table 5, the proposed method gives the highest average accuracy and F-measure for negative class on the STS dataset. However, the average F-measure for a positive class of the proposed method is slightly lower than those Gini, Chi2 and IG. In Table 6, the proposed method gives the highest average accuracy and F-measure for both class on the SemEval dataset. In Table 7, the proposed method gives the highest average accuracy and F-measure for the negative class on SS-Twitter dataset. However, it gives lower average F-measure for the positive class than other methods. In Table 8, the proposed method gives highest average accuracy and F-measure for the negative class on the HCR dataset. However, the average F-measure for the positive class is slightly lower than other methods. From Table 5 to Table 8, they are shown that the proposed method provides higher accuracy than Gini, Chi2 and IG. Furthermore, the proposed method results in the highest F-measure for the negative class. Table 9 reports the computational time of the proposed method against Gini, Chi2 and IG. It is shown that the proposed method outperforms Gini, Chi2 and IG.

Therefore, we can conclude that the proposed method selects most relevant features for classification with lower computational time when compared to Gini, Chi2 and IG. Moreover, support value or frequency of feature is significant for selecting relevant feature.

Table 5. Performance of the classification on the STS dataset

Number of Selected Features	Accuracy				F-Measure (Positive Class)				F-Measure (Negative Class)			
	Gini	Chi2	IG	Proposed	Gini	Chi2	IG	Proposed	Gini	Chi2	IG	Proposed
10%	71.16	71.05	71.06	71.91	72.47	72.37	72.55	72.63	70.07	69.95	69.86	71.27
20%	71.77	71.77	71.34	72.39	72.78	72.78	72.55	73.18	70.85	70.84	70.29	71.62
30%	71.43	71.45	71.18	72.13	72.71	72.72	72.62	72.71	70.44	70.48	70.11	71.59
40%	71.26	71.28	71.08	71.89	72.85	72.85	72.80	71.94	70.18	70.22	69.97	71.91
50%	71.21	71.25	71.05	71.95	72.97	72.99	72.88	71.28	70.02	70.07	69.79	72.65
60%	71.69	71.72	71.25	71.94	73.29	73.30	72.96	70.93	70.36	70.42	69.79	72.94
70%	71.66	71.66	71.54	72.27	73.11	73.11	73.12	72.11	70.29	70.29	70.03	72.50
80%	72.15	72.10	72.10	72.43	73.30	73.27	73.27	72.99	70.95	70.89	70.89	71.92
90%	72.25	72.25	72.27	72.30	73.20	73.20	73.21	73.25	71.27	71.27	71.30	71.32
Avg.	71.62	71.61	71.43	72.13	72.97	72.95	72.88	72.34	70.49	70.49	70.23	71.97

Table 6. Performance of the SemEval dataset

Number of Selected Features	Accuracy				F-Measure (Positive)				F-Measure (Negative)			
	Gini	Chi2	IG	Proposed	Gini	Chi2	IG	Proposed	Gini	Chi2	IG	Proposed
10%	86.68	86.68	86.68	87.18	86.55	86.55	86.51	87.12	86.77	86.77	86.85	87.18
20%	87.38	87.38	87.40	87.10	87.22	87.22	87.24	87.05	87.47	87.47	87.51	87.12
30%	86.95	86.95	87.38	87.38	86.82	86.82	87.27	87.36	87.06	87.06	87.47	87.35
40%	87.23	87.23	86.98	87.33	87.04	87.04	86.81	87.21	87.43	87.43	87.18	87.41
50%	87.00	87.00	86.95	87.50	86.78	86.78	86.73	87.19	87.27	87.27	87.23	87.76
60%	87.38	87.38	87.18	87.33	87.18	87.18	86.95	86.92	87.58	87.58	87.40	87.66
70%	87.48	87.50	87.23	87.63	87.24	87.26	87.00	87.34	87.67	87.70	87.42	87.86
80%	87.55	87.55	87.55	87.50	87.36	87.36	87.36	87.33	87.69	87.69	87.69	87.64
90%	87.68	87.68	87.68	87.68	87.56	87.56	87.56	87.56	87.76	87.76	87.76	87.76
Avg.	87.26	87.26	87.22	87.40	87.08	87.09	87.05	87.23	87.41	87.42	87.39	87.53

Table 7. Performance of the SS-twitter dataset

Number of Selected Features	Acc				F-Measure (Positive)				F-Measure (Negative)			
	Gini	Chi2	IG	Proposed	Gini	Chi2	IG	Proposed	Gini	Chi2	IG	Proposed
10%	59.31	59.23	58.35	60.92	63.69	63.62	63.57	62.29	55.01	54.94	53.38	59.64
20%	60.54	60.23	60.19	62.00	63.64	63.47	64.06	63.55	57.63	57.14	56.54	60.57
30%	60.19	60.23	60.38	61.77	63.75	63.82	64.71	62.50	57.06	57.04	56.61	61.19
40%	60.69	60.42	60.00	61.62	64.60	64.39	64.67	60.01	57.40	57.00	55.84	63.33
50%	61.12	61.08	60.65	61.27	65.14	65.07	65.33	58.02	57.35	57.33	56.02	64.23
60%	60.35	60.38	60.04	61.35	64.28	64.28	64.36	59.17	56.23	56.29	55.49	63.46
70%	60.73	60.58	59.88	60.85	64.25	64.03	63.88	60.94	56.90	56.83	55.52	60.91
80%	60.54	60.58	60.58	60.96	63.83	63.85	63.85	63.07	56.97	57.04	57.04	58.89
90%	60.12	60.12	60.12	60.12	63.23	63.23	63.23	63.23	56.73	56.73	56.73	56.73
Avg.	60.40	60.32	60.02	61.21	64.05	63.97	64.18	61.42	56.81	56.70	55.91	60.99

Table 8. Performance of the HCR dataset

Number of Selected Features	Acc				F-Measure (Positive)				F-Measure (Negative)			
	Gini	Chi2	IG	Proposed	Gini	Chi2	IG	Proposed	Gini	Chi2	IG	Proposed
10%	75.80	76.00	76.50	77.50	76.35	76.48	76.96	77.88	76.80	77.17	77.70	77.54
20%	78.40	78.50	77.90	78.10	78.55	78.56	78.17	78.82	78.99	79.20	78.55	78.01
30%	80.10	80.30	79.50	79.00	80.61	80.74	79.82	79.25	80.17	80.38	79.80	79.22
40%	80.40	80.50	81.20	80.10	81.06	81.16	81.78	80.56	80.10	80.16	81.01	79.82
50%	81.20	81.20	81.10	80.80	82.02	81.94	81.97	81.24	80.76	80.71	80.65	80.38
60%	81.70	81.60	81.80	82.80	82.57	82.46	82.57	82.94	81.22	81.14	81.35	82.73
70%	82.90	82.70	82.20	82.20	83.54	83.28	83.02	82.62	82.44	82.27	81.58	81.83
80%	81.60	81.70	81.50	82.90	82.44	82.51	82.36	83.32	80.68	80.80	80.58	82.55
90%	81.90	81.90	81.90	82.20	82.75	82.75	82.75	82.95	80.91	80.91	80.91	81.37
Avg.	80.44	80.49	80.40	80.62	81.10	81.10	81.04	81.06	80.23	80.31	80.24	80.38

Table 9. Computational time for feature ranking (second)

	Gini	Chi2	IG	Proposed
STS	3.149	2.923	3.512	2.844
SemEval	0.911	0.828	0.915	0.687
SS-Twitter	0.468	0.467	0.478	0.359
HCR	0.071	0.097	0.070	0.053

5. CONCLUSION

Feature selection is an important process to reduce the features and improve the performance of opinion classification. A feature selection method is proposed based on the concept of filter models together with association rule techniques. Support and confidence values are applied to calculated weight of feature. Support values are normalized to remove outliers. Moreover, a tuning parameter *p* is presented to balance the significance of the normalized support and confidence. The experimental results show that the support or frequency of feature has significant for feature selection. The proposed algorithm gives a higher accuracy than Gini, Chi2 and IG because it can find most relevant features that lead to high performance of classification. Moreover, the computation time of the proposed method is the best when comparing Gini, Chi2 and IG because the weights of features are quickly and easily calculated on vertical data format.

ACKNOWLEDGEMENTS

The authors are very grateful to the Departments of Information Technology, Mahasarakham University and Kalasin University for supporting this research.

REFERENCES

[1] S. Sangam and S. Shinde, "Sentiment classification of social media reviews using an ensemble classifier," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 16, no. 1, pp. 355-363, 2019.

[2] C. Troussas, et al., "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning," in *2013 Fourth International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1-6, 2013.

[3] M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning," in *2014 Sixth International Conference on Communication Systems and Networks (COMSNETS),* pp. 1-8, 2014.

[4] A. Ortigosa, et al., "Sentiment analysis in Facebook and its application to e-learning," *Computers in Human Behavior,* vol. 31, pp. 527-541, 2014.

[5] A. Adeleke, et al., "A two-step feature selection method for quranic text classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 730-736, 2019.

[6] T. A. Alhaj, et al., "Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation," *PloS one*, vol. 11,no. 11, p. e0166017, 2016.

[7] T. Parlar, et al., "QER: a new feature selection method for sentiment analysis," *Human-Centric Computing and Information Sciences,* vol. 8, 2018.

[8] A. I. Pratiwi and Adiwijaya, "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis," *Applied Computational Intelligence and Soft Computing,* vol. 2018, pp. 1-5, 2018.

[9] J. Yang, et al., "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," *Information Processing and Management,* vol. 48, no. 4, pp. 741-754, 2012.

[10] X. Deng, et al., "Feature selection for text classification: A review," *Multimedia Tools and Applications,* vol. 78, pp. 3797-3816, 2019.

[11] O. Somantri and D. Apriliani, "Opinion Mining on Culinary Food Customer Satisfaction Using Naïve Bayes Based-on Hybrid Feature Selection," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 15, no. 1, pp. 468-475, 2019.

[12] N. S. I. M. Rafei, et al., "Comparison of feature selection techniques in classifying stroke documents," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 14, no. 3, pp. 1244-1250, 2019.

[13] N. M. G. D. Purnamasari, et al., "Cyberbullying identification in twitter using support vector machine and information gain based feature selection," *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 18, no. 3, pp. 1494-1500, 2020.

[14] A. Go, et al., "Twitter sentiment classification using distant supervision," Stanford University, pp. 1-6, 2009.

[15] S. Rosenthal, et al., "SemEval-2017 Task 4: Sentiment Analysis in Twitter," *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017),* pp. 502-518, 2017.

[16] M. Thelwall, et al., "Sentiment strength detection for the social web," *Journal of the American Society for Information Science and Technology,* vol. 63, pp. 163-173, 2012.

[17] M. Speriosu, et al., "Twitter polarity classification with label propagation over lexical links and the follower graph," *The Proceedings of the First Workshop on Unsupervised Learning in NLP*, Edinburgh, Scotland, pp. 53-63, 2011.

[18] Y. L. Phua, "Social Media Sentiment Analysis and Topic Detection for Singapore English," Master's Thesis, Naval Postgraduate School, 2013.

[19] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artificial Intelligence Review,* vol. 52, pp. 1495-1545, 2019.

[20] M. Bilal, et al., "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques," *Journal of King Saud University - Computer and Information Sciences,* vol. 28, no. 3, pp. 330-344, 2016.

[21] K. M. A. Hasan, et al., "Opinion mining using Naïve Bayes," *2015 IEEE International WIE Conference on Electrical and Computer Engineering,* pp. 511-514, 2015.

[22] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," *International Conference on Applied and Theorical Computing and Communication Technology,* pp. 416-419, 2016.

[23] A. Goel, et al., "Real time sentiment analysis of tweets using Naive Bayes," *2016 2nd International Conference on Next Generation Computing Technologies,* pp. 257-261, 2016.

[24] R. A. Ramadhani, et al., "Comparison of Naive Bayes smoothing methods for Twitter sentiment analysis," in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 287-292, 2016.

[25] A. Prabhat and V. Khullar, "Sentiment classification on big data using Naïve bayes and logistic regression," in *2017 International Conference on Computer Communication and Informatics (ICCCI),* pp. 1-5, 2017.

BIOGRAPHIES OF AUTHORS

Atchara Choompol has been a Ph.D. student at the Faculty of Informatics, Mahasarakham University (MSU) Thailand since 2014. She obtained her master's degree in Information Technology at King Mongkut's Institute of Technology North Bangkok (KMUTNB) Thailand. Her research includes machine learning, data mining, text classification and feature selection.

Panida Songram is currently an Assistant Professor at the Faculty of Informatics, Mahasarakham University (MSU), Thailand.

Phattahanaphong Chomphuwiset is currently an Assistant Professor at the Faculty of Informatics, Mahasarakham University (MSU), Thailand.