International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Association on Supervised Term Weighting Method for Classification on Data Twitter

Imroatul Khuluqi Izzah, Abba Suganda Girsang

Abstract*: Term weighting is a preprocessing phase that has an important role in the text classification by giving the appropriate weight for each term in all documents. In previous research, many supervised term weighting methods have been introduced, but most of the supervised term weighting only considers the distribution of terms in the two classes so that it is not optimal for the multi-class classification. This paper introduces a new supervised weighting with association concept to optimize term weighting distributions in multi-class cases by considering terms that exist in each class and paying attention to the number of terms in the document belonging to the class, also considering the relationship pattern between one or more items with association concept in a dataset to measure the strength of terms in a class by using confidence values. The dataset used are the data twitter taken from the PR FM twitter account. The proposed supervised term weighting method implemented with SVM classifier can outperform unsupervised weighting schemes such as TF-IDF with the average accuracy 81.704%.

Keywords: Term-weighting, classification, twitter, association, confidence.

I. INTRODUCTION

The advancement of information technology is currently causing the development of use for social networks such as Twitter as a communication media for people to post messages, share information, communicate ideas and establish friendships between social media users from around the world to produce a large number of short electronic documents. Thus, this type of document must be classified so that it can be useful for the development of various applications related to social networks. Text classification is defined as an automatic process for assigning a text document to one or more predetermined topic categories or classes [1].

One of the main problems in work with text documents is unstructured documents form so that appropriate representations must be used in the automatic classification process. The most widely used approach to textual data representation is Vector Space Model (VSM) [2], where each document is represented as a vector formed from its index term value, also known as bag of words. Preprocessing phase of data are needed to extract words from the document and to weight according to the level of importance in each document in accordance with the chosen weighting scheme [3]. So that the appropriate term weighting is a basic problem in text classification and directly affects the classification accuracy.

Lan et al. [4] categorizes the term weighting method into two types, namely supervised and unsupervised. One of the most frequently used unsupervised term weighting methods (also known as traditional weighting methods) is TF-IDF which has proven effective in information retrieval. But this method is not very effective for text classification because it does not consider class labels in training documents so that they cannot fully reflect important terms in text classification [5]. For supervised term weighting, three methods are often used, namely TF-CHI, TF-IG, and TF-GR by replacing the IDF function in TF-IDF with such as feature selection function as $\chi 2$ statistic (CHI), information gain (IG) and gain ratio (GR) [6]. However, most of the supervised term weighting only considers the distribution of terms in the two classes so that it is not optimal for the multi-class classification of more than two classes. Therefore, the term weighting is a hot research topic on text classification [5].

Many existing works are found in the domain of term weighting methods for text classification but there is no weighting scheme that includes the association rule in its calculations. Whereas association rule can be used to find associative rules between a combination of items and analyze high-frequency patterns (frequent pattern mining) [7]. So, it can be used to get associative rules between a combination of words in each class in the text classification. By considering the pattern of relationships between one or more items in a dataset in a class, it can be measured the strength of terms in the class using the value of confidence. The association pattern is formed from every class that exists because class information on word weighting will be able to improve the accuracy of text classification [8].

Hence to overcome the weaknesses mentioned above, in this paper, a new supervised term weighting method is proposed which can optimize the multi-class classification taking into account the terms that exist in each class and pay attention to the number of terms in the document that are members of the class. Besides, it also considers the pattern of relationships between one or more items (association concepts) in a dataset to measure the strength of terms in the class by using confidence values. The proposed method is intended to improve accuracy in the multi-class classification process of data twitter.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Imroatul Khuluqi Izzah, Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480. Email : imroatul.izzah@binus.ac.id
Abba Suganda Girsang, Computer Science Department, BINUS Graduate Program, Bina Nusantara University, Jakarta, Indonesia 11480. Email : agirsang@binus.edu

II. RELATED WORK

One of the main steps in improving performance of text classification is the term weighting phase because at this stage the process of presenting text is done by setting the appropriate value for each term to improve the performance of text classification [9]. Man et al. [4] categorizes the term weighting method into two types, namely supervised and unsupervised (also known as traditional weighting methods).

Traditional word weighting schemes are known as unsupervised term weighting. These schemes include binary (boolean), TF and TF-IDF weighting [4]. One of the most frequently used unsupervised term weighting methods is TF-IDF which has proven effective in information retrieval. In several previous studies, a lot of research was done on the development of TF-IDF as a new weighting method based on frequency terms for classification of documents [10]. Then modified TF-IDF schemes are introduced which consider the count of terms lost from the document as a factor in calculating the existing term weights namely mTF, mTFIDF, TFmIDF, and mTFmIDF [1]. Jie Chen, et al. [11]proposes an adaptive weighting of keyword positions following the traditional TF-IDF algorithm, called the TF-IDF-AL Algorithm. This algorithm can utilize the internal characteristics of key documents.

Supervised term weighting is a weighting method that involves class labels on training documents. An example of this weighting method is TF-RF weighting which weighs a term based on the frequency of its relevance (RF), namely the frequency of the term documents in the positive and negative classes [4]. TF-RF showed that the more high-frequency terms are concentrated in the positive class than in the negative class, the more contributions they make in choosing positive text from negative text [4]. TF-RF gave better performance compared to traditional term weighting schemes [12]. However, TF-RF is not an optimal term weighting scheme for multi-class text classification because it ignores the distribution of this term in various text classes [5].

III. PROPOSED METHOD

In this section, we propose a new supervised term weighting method by incorporating association concept to optimize multi-class classification. The detail of the proposed method can be seen in the next paragraph. Fig.1 is an overview of the research methods in this study.

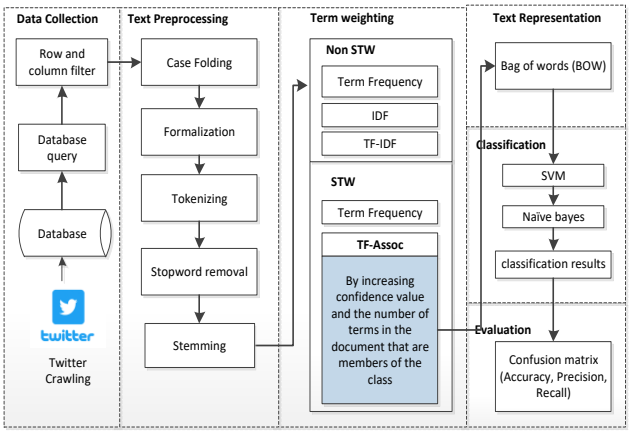

Fig. 1. Research Method

At Data collection stage, dataset is collected from the Indonesian-language content of posts (tweets) from PR FM's Twitter account, @PRFMnews. Based on 5,012 tweets that has been crawled, there are 9 types of tweet, namely disaster information, weather information, economic information, health information, criminal information, traffic information, sports information, political information, and general information.

The next stage is the text preprocessing stage, preprocessing consists of several stages namely case folding, formalization, tokenization, filtering and stemming to find the root of word. In term weighting stage**,** we propose a new supervised term weighting method by incorporating the concept of association to optimize the multi-class classification by considers the pattern of relationships between one or more items (association concepts) in a dataset by using confidence values.

To find out the pattern of relationships between one or more items on the concept of association in a class first is to calculate the confidence value (*Conf*) of each pattern (*r*) in each class. Next, calculate the average value of confidence of the resulting pattern containing term *i* (*ti*) as in Eq. (1).

$$Avg(Conf_{(ti,r)}) = \frac{\sum_{i=1}^{k} Conf_{ti,ck}}{r_{ck}} \quad (1)$$

Calculation of the total class that contains the value of confidence term ($C_{conf(ti)}$), as well as the average confidence of the words in each class $Avg(Conf(ti))$ in $AssocBased_{(ti)}$ is intended to calculate the value of the term strength in a particular class as in Eq. (2) $Assoc_{(ti)}$ is calculated by calculating the total value of $AssocBased_{(ti)}$ and the comparison of the total class number (*C*) to the total term frequency of each class ($C_{k,ti}$) with the number of classes containing the terms *i* ($f_{ti,ck}$), as in Eq. (3).

$$AssocBased_{(ti)} = \frac{Avg(Conf(ti))}{\sum_{j=1}^{k} C_{conf(ti)}} \quad (2)$$

$$Assoc_{(ti)} = AssocBased_{(ti)} + Log\left(\frac{C}{\sum_{j=1}^{k} \frac{C_{k,ti}}{f_{ti,ck}}}\right) \quad (3)$$

$$TF - Assoc = f_{(ti,dj)} * Assoc_{(ti)} \quad (4)$$

Then $TF - Assoc$ is formed by combining the values of *TF* and $Assoc_{(ti)}$ as in Eq. (4). TF is the calculation of the frequency of occurrence of the term (t_i) in the document (d_i).

At classification stage, this research use Naïve Bayes and SVM algorithms. And for evaluation step, the confusion matrix method is used for see the level of accuracy, precision, recall, and error-rate. An example of a confusion matrix for multi-class classification is shown in (Table-I).

The formula for calculate Accuracy, Precision, Recall and Error Rate are defined by Eq. (5), Eq. (6), Eq. (7), and Eq. (8). With information that (*TP*) is true positive, (*TN*) is true negative, (*FN*) is false negative, (*FP*) is false positive and (*C*) is count of classes.

$$Recall = \frac{\sum TP}{\sum(TP+FN)} \quad (5)$$

$$Precision = \frac{\sum TP}{\sum(FP+TP)} \quad (6)$$

$$Accuracy = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} / C \quad (7)$$

$$Error\ Rate = 1 - Accuracy \quad (8)$$

Table- I: Confusion Matrix Model

		Predicted Class				
		A	***B***	***C***	***…***	***I***
Actual Class	***A***	$tp_{A,A}$	$fn_{A,B}$	$fn_{A,C}$	…	$fn_{I,D}$
	B	$fp_{B,A}$	$tp_{B,B}$	$fn_{B,C}$	…	$fn_{I,D}$
	C	$fp_{C,A}$	$fp_{C,B}$	$tp_{C,C}$	…	$fn_{I,D}$
	…	…	…	…	…	…
	I	$fp_{I,A}$	$fp_{I,B}$	$fp_{I,C}$	…	$tp_{I,D}$

IV. DATA ANALYSIS AND RESULT

A. Dataset

The dataset used in this study is collected from the Indonesian-language content of posts (tweets) from PR FM's Twitter account, @PRFMnews. From 5,012 tweets, then data are classified into 9 classes as shown in (Table-II).

Table- II: List of The Amount Classified Tweets

Class	Amount
Disaster Information	337
Weather Information	681
Economic Information	163
Health Information	95
Criminal Information	204
Traffic Information	1.145
Sport Information	556
Political Information	330
General Information	1.501

B. Preprocessing

The preprocessing stages including case folding, formalization, tokenizing, filtering and stemming. This step is done using the INA-NLP library [13]. Case-folding is done because not all text documents are consistent in the use of capital letters. Formalization will be carried out by changing from non-standard words to standard words. Tokenization stage is carried out by separating each word that composes a document or conversation. The last one is stemming, with transforming words obtained from filtering results into basic word forms using the INANLP-Non-Deterministic algorithm [14]. The unpreprocessed tweets and preprocessed tweets that have cleaned can be seen at (Table-III) and (Table-IV).

Table- III: Unpreprocessed Tweets

Class	Tweet
Traffic Information	Macet macetan di jalan belokan pasteur
	Macet parah di belok belokan
	Parah nih parah belokan pasteur merayap
Weather Information	Hujan hujan gini dago merayap
	hujan di jalan cikoneng baru
	suporter persib kehujanan di cikoneng baru
Sport Information	penonton penonton padati belokan
	suporter penonton dari pemain baru
	parah penonton bola baru rusuh

$$Sup(A) = \frac{number\ of\ transactions\ containing\ A}{Total\ transaction} \quad (10)$$

Table- IV: Preprocessed Tweets

Class	Tweet
Traffic Information	macet macet jalan belok pasteur
	macet parah belok belok
	parah belok pasteur rayap
Weather Information	hujan hujan dago rayap
	hujan jalan cikoneng baru
	suporter persib hujan cikoneng baru
Sport Information	tonton tonton padat belok
	suporter tonton main baru
	parah tonton bola baru rusuh

C. Proposed Method Implementation

The first process in TF-Assoc is calculate TF. TF is a term weighting method based on the frequency of term (*tk*) appearing in the document (*dj*). TF defined by Eq. (9). The next step is to look for patterns of association in each class with list the transaction in each class, for example transaction in traffic information class shown in (Table-V).

$$TF_{(tk,dj)} = f_{(tk,dj)} \quad (9)$$

Table-V: Transaction in Traffic Information Class

Term	D1	D2	D3	Amount
macet	1	1	0	2
jalan	1	0	0	1
pasteur	1	1	1	3
belok	1	0	1	2
parah	0	1	1	2
hujan	0	0	0	0
dago	0	0	0	0
rayap	0	0	1	1
cikoneng	0	0	0	0
baru	0	0	0	0
persib	0	0	0	0
tonton	0	0	0	0
padat	0	0	0	0
suporter	0	0	0	0
main	0	0	0	0
rusuh	0	0	0	0
bola	0	0	0	0

Then looking for a combination of items that meet the minimum requirements of the support value in the database. Previously we can determine the minimum support, the parameters set experimentally for making association rules in this research are determined by taking the minimum value of minSupport equal to minConfidence which is 0.05. But for this example, the value of MinSupport = 0.5. As for the Eq.(10), for support combination 1 itemset implementation is shown in (Table-VI) and for support combination 2 itemset implementation is shown in (Table-VII).

Table- VI: Table for Support Combination 1 Itemset

Itemset	Support
macet	0.666667
pasteur	1
belok	0.666667
parah	0.666667

Table- VII: Table for Support Combination 1 Itemset

Itemset	Support
macet, pasteur	0.666667
macet, parah	0.666667
pasteur, belok	0.666667
pasteur, parah	0.333333

After all high frequency patterns have been found, then the association rules that meet the minimum requirements for confidence are searched by calculating the associative rule confidence A → B, as for the calculation using Eq. (11).

$$Conf = \frac{number\ of\ transactions\ containing\ A \text{ and } B}{Total\ transaction\ containing\ A} \quad (11)$$

Furthermore, this stage will be useful to find all the rules of association in each class that meets the threshold (for example MinConfidence = 0.7). For implementation is shown in (Table-VIII). After find all the rule, then calculate the average of confidence which was shown in (Table-IX).

Table- VIII: Rules of Association

Association Pattern	Support	Confidence
macet → pasteur	0.666667	1
macet → parah	0.666667	1
parah → macet	0.666667	1
parah→pasteur	0.666667	1

Table-IX: Average of Confidence

Association Pattern	Confidence
macet	0.75
pasteur	0.5
parah	0.75

$AssocBased_{(ti)}$ is intended to measure the strength of the term in the class formulated in Eq.(2), so the results can be combined with the calculation of Frequency Term (TF) to TF-Assoc. For example, the result of form term weighting calculation (TF-Assoc) is shown in (Table-X).

Table-X: Transaction in Traffic Information Class

Term	D1	D2	D3	…	D9
macet	3.4085	1.7042	0	…	0
jalan	0.4771	0	0	…	0
pasteur	1.3750	2.7501	1.3751	…	0
belok	0.7781	0	0.7782	…	0
parah	0	1.5282	3.0563	…	1.5282
hujan	0	0	0	…	0
dago	0	0	0	…	0
rayap	0	0	0.4771	…	0
cikoneng	0	0	0	…	0

D. Performance Evaluation

The vectors obtained from each document are then classified using two classifiers, Naïve Bayes and SVM. Existing data partitioned with 10 cross fold validation is used to measure the accuracy of the classification model created. In the evaluation step, the confusion matrix method is used to measure the performance of each classifier using the new supervised term weight with the value of confidence as additional information, and the performance of TF-IDF term weighting. Then the calculation of the value of accuracy, precision, and recall are done by calculating the average value of accuracy, precision and recall in each class as shown in (Table-XI) And (Table-XII).

Table-XI: The Comparison results With TF-IDF Term Weighting

Class	SVM Classifier		Naïve Bayes Classifier	
	Precision	*Recall*	*Precision*	*Recall*
Disaster Information	0.894	0.822	0.778	0.780
Weather Information	0.929	0.935	0.872	0.907
Economic Information	0.783	0.552	0.662	0.540
Health Information	0.662	0.453	0.534	0.495
Criminal Information	0.665	0.534	0.447	0.559
Traffic Information	0.866	0.839	0.896	0.817
Sport Information	0.914	0.824	0.922	0.845
Political Information	0.793	0.709	0.579	0.764
General Information	0.728	0.856	0.744	0.744

Table-XII: The Comparison results With TF-Assoc Term Weighting

Class	SVM Classifier		Naïve Bayes Classifier	
	Precision	*Recall*	*Precision*	*Recall*
Disaster Information	0.894	0.822	0.781	0.783
Weather Information	0.929	0.935	0.874	0.907
Economic Information	0.783	0.552	0.664	0.546
Health Information	0.662	0.453	0.534	0.495
Criminal Information	0.665	0.534	0.445	0.554
Traffic Information	0.866	0.839	0.895	0.818
Sport Information	0.916	0.824	0.924	0.847
Political Information	0.793	0.709	0.580	0.767
General Information	0.728	0.857	0.745	0.744

The results of average value of accuracy for the tested term weighting with SVM classifier and Naïve Bayes classifier in Table-XIII.

Table-XIII: The Results for Tested Term Weighting

Term Weighting	Metric	SVM Classifier	Naïve Bayes Classifier
TF-IDF	Accuracy	81.684%	77.913%
TF-Assoc	Accuracy	**81.704%**	**77.972%**

From these results, our proposed method, TF-Assoc weighting scheme uses SVM classifier,

show better performance compared to Naïve Bayes classifier. And result with TF-Assoc term weighting, outperform TF-IDF term weighting as unsupervised term weighting which can be seen from the highest average accuracy on SVM classifier (81.704%) and Naïve Bayes classifier (77.972%).

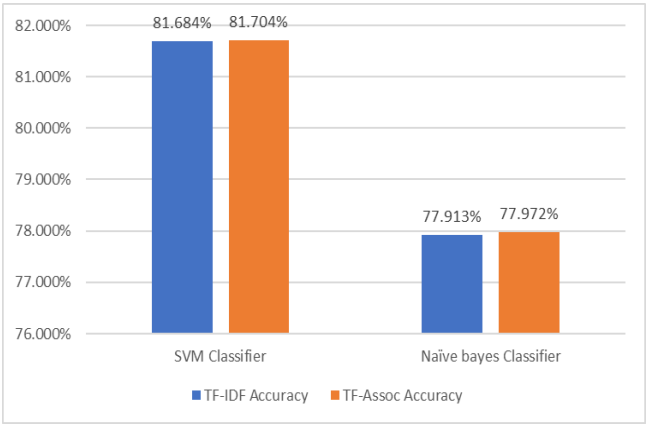

Fig. 2. Chart of the Results For TF-IDF and TF-Assoc Term Weighting

For our proposed method, TF-Assoc weighting scheme uses SVM classifier, outperforms TF-IDF term weighting. The average value of accuracy based on the SVM classifier in TF-Assoc is higher (81.704%) compared to TF-IDF term weighting (for 81.684%) as shown in Fig.2.

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed TF-Assoc supervised term weighting, TF-Assoc integrates calculations with the concept of association using a priori algorithms by taking confidence values from the resulting pattern. So that it can characterize the distribution of words in each class and measure the strength of a term in the class. This is proven by experiments conducted with data sets obtained from the PR FM data twitter account and it is known that TF-Assoc with the SVM classifier outperforms unsupervised term weighting such as TF-IDF.

In the future, we will conduct experiments with a broader scope of experiments for text classification, such as experiments in text document classification (not short text from social media).

REFERENCES

1. T. Sabbah *et al.*, "Modified frequency-based term weighting schemes for text classification," *Appl. Soft Comput. J.*, vol. 58, pp. 193–206, 2017.
2. M. Melucci, "Vector-Space Model," no. Encyclopedia of Database Systems, pp. 3259–3263, 2009.
3. G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, "A Study on Term Weighting for Text Categorization: A Novel Supervised Variant of tf.idf," *Proc. 4th Int. Conf. Data Manag. Technol. Appl.*, no. July, pp. 26–37, 2015.
4. M. Lan, C. L. Tan, S. Member, J. Su, and Y. Lu, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," vol. 31, no. 4, pp. 721–735, 2009.
5. K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Syst. Appl.*, vol. 66, pp. 1339–1351, 2016.
6. F. Debole and F. Sebastiani, "Supervised Term Weighting for Automated Text Categorization," *Symp. A Q. J. Mod. Foreign Lit.*, no. Ml, pp. 784–788, 2003.
7. A. Bhandari, A. Gupta, and D. Das, "Improvised apriori algorithm using frequent pattern tree for real time applications in data mining," *Procedia - Procedia Comput. Sci.*, vol. 46, no. Icict 2014, pp. 644–651, 2015.
8. M. Haddoud, A. Mokhtari, T. Lecroq, and S. Abdeddaïm, "Combining supervised term-weighting metrics for SVM text classification with extended term representation," *Knowl. Inf. Syst.*, 2016.
9. G. Feng, S. Li, T. Sun, and B. Zhang, "A probabilistic model derived term weighting scheme for text classification," *Pattern Recognit. Lett.*, vol. 110, pp. 23–29, 2018.
10. F. Sebastiani, "Machine Learning in Automated Text Categorization," vol. 34, no. 1, pp. 1–47, 2002.
11. J. Chen, C. Chen, and Y. Liang, "Optimized TF-IDF Algorithm with the Adaptive Weight of Position of Word," vol. 133, pp. 114–117, 2016.
12. H. Altınçay and Z. Erenel, "Analytical evaluation of term weighting schemes for text categorization," *Pattern Recognit. Lett.*, vol. 31, no. 11, pp. 1310–1323, 2010.
13. A. Purwarianti, A. Andhika, A. F. Wicaksono, I. Afif, and F. Ferdian, "InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification," *4th IGNITE Conf. 2016 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2016*, pp. 5–9, 2016.
14. A. Purwarianti and A. I. M. Rule, "A Non Deterministic Indonesian Stemmer," no. July, pp. 1–5, 2011.

AUTHORS PROFILE

Imroatul Khuluqi Izzah, received her bachelor degree at informatics engineering from Universitas Muhammadiyah Sidoarjo (UMSIDA), Sidoarjo, Indonesia. Currently she is entering the second year of magister student in computer science at Bina Nusantara University, Jakarta, Indonesia. Her research interest includes text mining and data mining.

Abba Suganda Girsang, is currently a lecturer at Master in Computer Science, Bina Nusantara University. He got Ph.D. in the Institute of Computer and Communication Engineering, Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, He graduated bachelor from the Department of Electrical Engineering, Gadjah Mada University (UGM), Yogyakarta, Indonesia, in 2000. He then continued his masters degree in the Department of Computer Science in the same university in 2006–2008. He was a staff consultant programmer in Bethesda Hospital, Yogyakarta, in 2001 and also worked as a web developer in 2002–2003. He then joined the faculty of Department of Informatics Engineering in Janabadra University as a lecturer in 2003-2015. His research interests include swarm, intelligence combinatorial optimization, and decision support system.