Modified TF-Assoc Term Weighting Method for Text Classification on News Dataset from Twitter

Imroatul Khuluqi Izzah, Abba Suganda Girsang

***Abstract*— Text classification is a process of locating text documents automatically into categories based on the text content. In-text classification, there is a stage that has an important role in giving the value of importance to each document, that is term weighting. In the researchers' previous study, a new supervised term weighting (TF-Assoc) was introduced with the concept of association to optimize term weighting distribution in a case of multiclass classification. To improve the performance of text categorization, this paper proposes a term weighting scheme with a modified association concept, that is mTF-IDF-Assoc. The proposed term weighting scheme considered Document Length (DL). DL was used to normalize the term frequency by dividing it by the length of the document's vector and then formulting IDF and Assoc in calculating the weight of each word. The results showed that mTF-IDF-Assoc implemented with SVM classifier and 10-fold cross-validation technique could outperform the TF-IDF, TF-ICF, and TF-Assoc weighting scheme with an average accuracy of 82.322%.**

***Index Terms*— text classification, document length, supervised term weighting, association, confidence**

I. Introduction

THE development of the utilization of online media such as social networking, microblogging, and online news media causes the number of available electronic text documents increasing in the last few years. The accessible text documents can be analyzed and then used in several cases, such as text classification [1][2], sentiment analysis [3], and automatic summarization [4] so that useful information can be obtained from a collection of the text documents. Text classification is a supervised learning task because the process of locating text documents uses information about the document membership in a predetermined class to develop classification using machine learning technique [5] [6].

For decades, text classification has been well studied and many types of classification algorithm have been developed. This algorithm includes K-Nearest Neighbor (KNN) [7], Support Vector Machines (SVM) [8][9], Decision Trees (DT) [10] and Naïve Bayes [11]. In text classification, a term weighting scheme controls the performance of classifier. That is why term weighting becomes a special concern for better text classification performance. The commonly used and effective approach to document representation is the Bag-of-Words (BoW) model. BoW is a simple but effective method to map documents into fixed-length vectors [12]. Therefore, a data preprocessing phase is needed to prepare text to be data that will be processed further and then given weight according to the level of importance.

A data preprocessing stage is needed in text classification to extract words from documents and term weighting to give weight to the words contained in documents according to the term weighting scheme that is used for the weighting [13]. There are two types of term weighting schemes, supervised term weighting and unsupervised term weighting [14]. The unsupervised term weighting scheme do not consider class labels in training documents, for example TF-IDF, Binary, and entropy. On the other hand, the supervised term weighting scheme, for example TF-RF, TF-ICF, and TF-IG consider class labels in training documents [14][15]. In the researcher' previous study [16], the researcher proposed a new supervised term weighting which considers the concept of association by using the confidence value as a parameter in measuring the strength of terms and finding patterns of relationships between one or more terms in the class. This is called TF-Assoc term weighting [16]. The use of a different term weighting scheme in text classification can affect the result of classification accuracy [17].

Term Frequency and Inverse Document Frequency (TF-IDF) is a well-known and most commonly used unsupervised term weighting in information search and many text classification tasks. However, this term weighting is not the most effective term weighting in text classification because TF-IDF ignores class labels that are available from training documents [18]. The use of supervised term weighting in text classification sometimes gives better accuracy results but not too significant the increase in the accuracy. In the researchers' previous study, TF-Assoc produced better accuracy values compared to TF-IDF but the difference in accuracy values is not very significant [16]. Therefore, it is necessary to have improvement or modification to the TF-Assoc supervised term weighting to increase the accuracy results.

Many previous studies investigated the domain of term weighting method for text classification. However, this study contributes by proposing supervised term weighting that formulates modified TF, IDF and Assoc. That is mTF-IDF-Assoc. The proposed method considers Document

Manuscript received August 31, 2020; revised January 4, 2021. This work was supported in part by Bina Nusantara Graduate program.

Imroatul Khuluqi Izzah is a master student at 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 is a lecturer at Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480 (corresponding author, e-mail: agirsang@binus.edu).

Length (DL), term frequency in document (IDF), and the value of confidence to see the strength level of terms in class and to find patterns of relationships between one or more terms in the class. The proposed method is intended to improve the performance of TF-Assoc method that was introduced previously and to improve the accuracy in the multiclass classification of the short-text news dataset from twitter.

II. Related Works

The interesting and important thing in text mining is the classification of text documents with sophisticated machine learning technique to categorize documents that are not yet structured. Thus, in the text classification there is a term weighting stage to determine the appropriate value for each word in a document. Man et al. [14] explained that there are two types of term weighting methods, namely supervised and unsupervised term weighting (also known as traditional weighting method). The unsupervised term weighting schemes are, for example, binary term weighting, TF term weighting, and TF-IDF term weighting. The unsupervised term weighting which is often used in text classification is TF-IDF term weighting.

In the previous studies, several studies were conducted on the modification of TF-IDF as a development of term weighting method to improve accuracy in text classification cases [15]. Thabit Sabbah et al. [17] proposed a weighting method by modifying TF-IDF. The term weighting was proposed based on the basis of term for document classification from the web. This scheme was proposed by considering the number of missing terms in a document in term weighting process. In the study, four frequency-based term weighting schemes were introduced that considered the number of terms missing from a document as a factor in calculating the term weight, namely mTF, mTFIDF, TFmIDF, and mTFmIDF. Issa Alsmadi et al. [19] also developed TF-IDF weighting which utilized the distribution of terms in both classes and across the corpus and combined the distribution of terms with event weight that measured the strength of terms in the document where the term appeared. The proposed scheme aimed to overcome the problem of time frequency in short text and to overcome the negative effects of sparsity and shortness in short text on the performance of classification tasks.

The characteristic of text categorization is supervised, so the term weighting in which there is a supervised learning activity is called supervised term weighting, TF-RF is an example of this term weighting [14]. The performance of an unsupervised term weighting scheme such as TF-IDF can be outperformed by the performance of TF-RF [20]. However, TF-RF does not consider the terms spread in several classes, that is why the accuracy of the multi-class classification results in the short text document is not optimal [18].

Besides, Haoying Wu, et al [21] proposed a modification of the FDCD-TF-IDF feature weighting algorithm based on the information distribution of term frequency and the information distribution of class. This algorithm modification introduced the concept of term frequency distribution and class distribution to describe term weight. Fuji Ren et al. [22] proposed term weighting based on class index that combined TF-IDF with ICF (Inverse Class Frequency). Therefore, term weighting is an interesting research object and it is widely discussed in research with the topic of text classification. Kewen Chen et al. [18] proposed TF-IGM (Inverse Gravity Moment) weighting to measure the difference in strength of terms in the text corpus precisely by measuring the concentration of term distribution in multiclass [18]. The basic idea of this weighting scheme is that the concentration of distribution between classes of terms can be reflected by the position of the center of gravity.

In the researchers' previous study, the researcher proposed supervised term weighting namely TF-Assoc term weighting which considers the association rules by using the confidence value as a parameter in measuring the strength of terms and finding patterns of relationships between one or more terms in each class [16]. In this study, TF-Assoc produced better accuracy values compared to TF-IDF, but the difference in accuracy values is not very significant. Based on the related literature review about the method of unsupervised term weighting and supervised term weighting for text classification, supervised term weighting gives higher accuracy compared to unsupervised term weighting.

III. Research Method

In this study, the researchers proposed a new supervised term weighting by formulating modified TF, IDF and Assoc to optimize the results of the multiclass classification. The contribution in this research is to modify TF-Assoc which previously introduced. Figure 1 is an overview of the research method in this study.

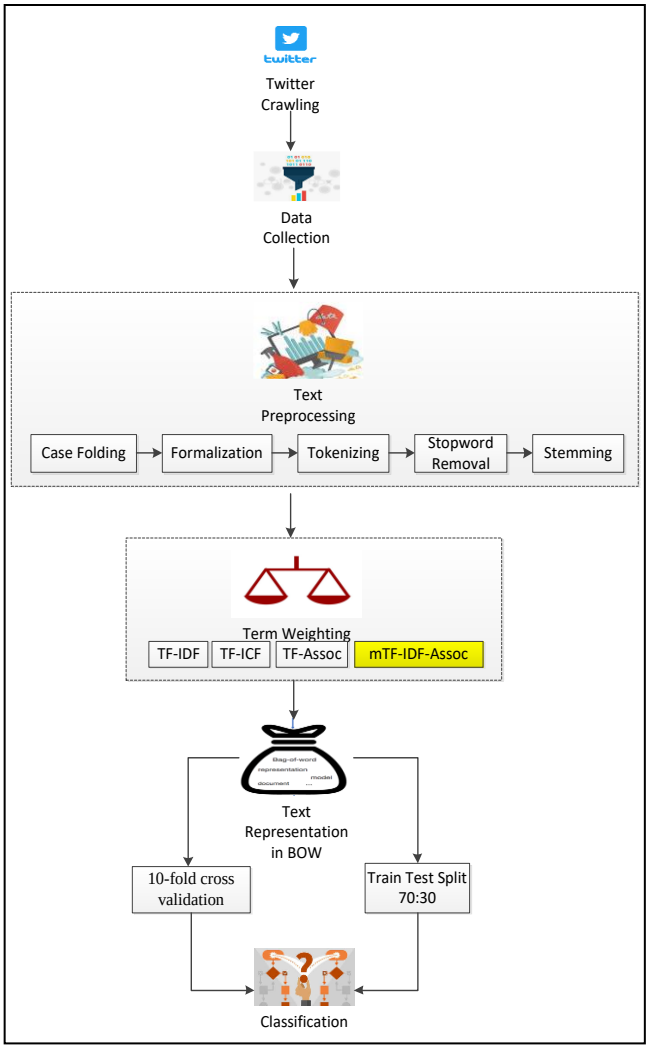

Fig. 1. Research Method

A. Data Collection

This study used a dataset collected from the Twitter account of private radio in Bandung, namely PR FM (@PRFMnews), and the dataset were short texts in Indonesian-language. The process of collecting this dataset was done using a crawler program that utilized the Twitter API service. The collected data contained information of id posts, tweet contents, tweet date, and the number of retweets. Based on 5,012 tweets on the PR FM Twitter account that had been crawled, there were 9 types of content or rubric category of news, namely economic rubric, criminal rubric, traffic rubric, political rubric, disaster rubric, health rubric, sports rubric, weather rubric, and general rubric. The data were then stored in a MySql database which then would be processed in text preprocessing stage, as illustrated in Figure 2.

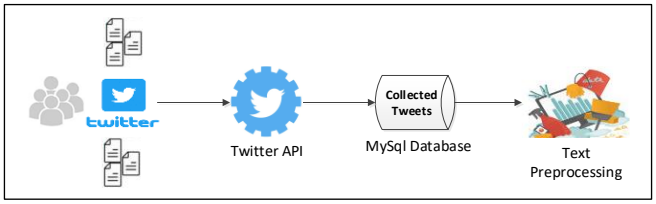

Fig. 2. Data Collection

B. Preprocessing

The short texts in each dataset was preprocessed to facilitate the processing of text document at the later stage. Preprocessing consists of several steps illustrated in Figure 3.

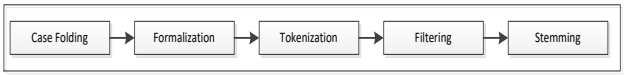

Fig. 3. Pre-processing

The first step is case folding. Case-folding is the process of equating letter format in a document. This stage is done because not all text documents are consistent in the use of capital letters. Therefore, the case-folding role is needed in converting the entire text in a document from capital letters to the standard form which is usually in the form of lowercase letters (a, ..., z). For example, the sentence "Cuacanya di desa gue, Cikoneng masih Berawan" will be processed to "cuacanya di desa gue, cikoneng masih berawan".

The second step is formalization. Dataset taken from social media sometimes contains non-standard words so at this stage terms will be formalized by changing them to standard words. For example, the sentence is "cuacanya di desa gue, cikoneng masih berawan". Every word in the sentence will be checked whether there are non-standard words using formalization dictionary. If there are non-standard words, they have to be changed to standard words. Then, the sentence will be "cuacanya di desa saya, cikoneng masih berawan ".

The third step is tokenization. Tokenization is the process of separating every word that makes up a document or conversation. This process also omits punctuation such as periods (.), commas (,), question marks (?), etc. For example, the sentence "cuacanya di desa saya, cikoneng masih berawan" will be processed to be "cuacanya", "di", "desa", "saya", "cikoneng", "masih", "berawan".

The fourth step is filtering. Filtering is the process of removing words that are not needed or words that have no meaning. For example, the sentence "cuacanya di desa saya, cikoneng masih berawan" will be processed to "cuacanya", "desa", "cikoneng", "masih", "berawan". Finally, the last step in preprocessing is stemming. Stemming is the process of finding the root word by eliminating the suffixes and prefixes. For example, the word "cuacanya" will be processed to "cuaca".

C. Term Weighting

The proposed term weighting scheme, mTF-IDF-Assoc, was developed from the modified TF, IDF and Assoc formula. The proposed method in this study was to modify TF-Assoc [16] to be mTF-IDF-Assoc for multiclass classification by considering Document Length (DL). DL was used to normalize the term frequency by dividing TF by the length of document. Inverse Document Frequency (IDF) was used to calculate the weight of each word in a document. IDF showed the availability relationship of a term in all documents. The smaller the number of documents containing the intended term, the bigger the IDF value [23]. This study also formulated Assoc which is a parameter with the concept of association to know the level of strength of terms in class by taking the confidence value found from the relationship between terms.

Firstly, in this proposed term weighting the term frequency (TF) was normalized by dividing it by document length (DL) (Equation 1). Then, the researchers calculated IDF by dividing the number of documents (N) by the number of documents containing term value t ($df_{(ti)}$) (Equation 2).

$$mTF_{(ti)} = \left(\frac{Log\ (TF_{(ti)} + 1)}{Log\ (DL)}\right) \quad (1)$$

$$\mathrm{IDF}_{(ti)} = \log\left(\frac{\mathrm{N}}{\mathrm{df}_{(ti)}}\right) \quad (2)$$

The average value of confidence was found out by calculating the confidence value ($Conf$) which fulfill certain limitation ($threshold$) of each pattern (r) containing term i (ti) (Equation 3).

$$Avg(Conf_{(ti,r)}) = \frac{\sum_{i=1}^{k} Conf_{ti,ck}}{r_{ck}} \quad (3)$$

The value of the term strength in a class can be determined by calculating the total of classes that have a term confidence value ($C_{conf(ti)}$) and also the average confidence of the relationship pattern between words in each class $Avg(Conf(ti))$ in $AssocBased_{(ti)}$ (Equation 4). $Assoc_{(ti)}$ was found out by calculating the total value of $AssocBased_{(ti)}$ and the comparison of the number of total class (C) to the total term frequency of each class ($C_{k,ti}$) with the number of classes containing the terms i ($f_{ti,ck}$) (Equation 5). $mTF - IDF - Assoc$ was formed by combining the values of $mTF_{(ti)}$, $\mathrm{IDF}_{(ti)}$ and $Assoc_{(ti)}$ (Equation 6). The proposed term weighting in this study is not the same as the previous term weighting schemes like TF-IDF, TF-ICF, TF-Assoc and another term weighting.

$$AssocBased_{(ti)} = \frac{Avg(Conf(ti))}{\sum_{j=1}^{k} C_{conf(ti)}} \quad (4)$$

$$Assoc_{(ti)} = AssocBased_{(ti)} + Log\left(\frac{C}{\sum_{j=1}^{k} \frac{C_{k,ti}}{f_{ti,ck}}}\right) \quad (5)$$

$$mTF - IDF - Assoc = mTF_{(ti)} * \mathrm{IDF}_{(ti)} * Assoc_{(ti)} \quad (6)$$

D. Classification

In the last few years, text classification algorithms have been widely studied and developed, such as Naïve Bayes [11][24], K-Nearest Neighbor (KNN) [7] and Support Vector Machines (SVM) [8][9]. That is why in the classification stage, the Naïve Bayes, KNN and SVM classifier method were used to determine the type of cotents or news rubric in the classification of posting content (tweets) from Twitter.

The segmentation technique used in this study was the Split Test Train 70:30 data and 10-fold cross-validation. For 10-fold cross-validation, the dataset of tweets was divided into training data and testing data using the 10-fold cross-validation method by dividing the dataset 10 times with cross-validation with equal distribution for training data and test data [25]. On the other hand, for the Train Test Split 70:30 the dataset (tweet) were divided into training data and testing data using the Random split 70:30 method by dividing the data set into training data and testing data with a ratio of 70% for training data and 30 % for testing data. After that, the data were used for the process of multiclass classification using the Naïve Bayes, KNN, and SVM algorithm.

E. Evaluation

The performance evaluation in this study used confusion matrix for multiclass data classification. The confusion matrix analyzed how well the classifier recognized data from more than two classes. Also, it gave information about the level of accuracy. If the accuracy value obtained in the classification were close to 100%, the error-rate obtained would get closer to the value of 0. Therefore, it can be concluded that the classification model developed is getting better. Table 1 presents the confusion matrix used for multi-class classification cases.

The equations to look for the value of accuracy, precision, recall and error rate are presented below respectively (7), (8), (9), and (10). (*TP)* is true positive, (*TN)* is true negative, (*FN)* is false negative, (*FP*) is false positive and (*C*) is the the number of classes.

$$Recall = \frac{\sum TP}{\sum (TP+FN)} \quad (7)$$

$$Precision = \frac{\sum TP}{\sum (FP+TP)} \quad (8)$$

$$Accuracy = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} / C \quad (9)$$

$$Error\ Rate = 1 - Accuracy \quad (10)$$

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. RESULTS AND DISCUSSION

A. Dataset

This study used a dataset collected from the Twitter account of private radio in Bandung, namely PR FM (@PRFMnews), and the dataset were short texts in Indonesian-language. From dataset that have been collected, there were 5,012 short-texts. The short-text is classified into 9 rubrics or classes, namely economic rubric, criminal rubric, traffic rubric, political rubric, disaster rubric, health rubric, sports rubric, weather rubric, and general rubric. The Table 2 below presents the list of classified tweets.

TABLE II
LIST OF CLASSIFIED TWEETS

Type of Rubric	Amount
Economic Rubric	163
Criminal Rubric	204
Traffic Rubric	1.145
Political Rubric	330
Disaster Rubric	337
Health Rubric	95
Sports Rubric	556
Weather Rubric	681
General Rubric	1.501

B. Preprocessing

The tweets having been crawled then got into the preprocessing stage in order that the data to be weighted was clean. The process was divided into several steps, namely case folding, formalization, tokenizing, filtering, and stemming. The library which used in the preprocessing stage is INA-NLP [26].

Tweets in the dataset must had to pass the preprocessing stage to select data in documents that will be processed in the later stage. Before executing the preprocessing stage, the researchers had carried out data cleaning by which the dataset obtained from the PR FM twitter account had been stored in the database and then the data cleaning had been done using Jython to omit the hashtags, emoticons, and hyperlinks that exist in each text document. After that the first step of preprocessing was executed that was case folding. In text documents, not all words use the same letterform, so it needed a step to equalize the letters in the document, this step is called Case-folding.

The second step is formalization that is the process of changing the non-standard words to standard words. This step was done because dataset taken from social media sometimes contain non-standard words. The third step is tokenization, that is the process of converting a document or

sentence into certain parts called tokens. The last step in preprocessing is stemming, stemming is the process of deciphering words or changing affix words into basic word forms. The method that used in the stemming process is the INANLP-Non-Deterministic Algorithm [27]. The table 3 is unpreprocessed tweets from twitter and the table 4 is preprocessed tweet. After the preprocessing stage was executed, the sentence was taken to term weighting stage. The table 4 contains the preprocessed tweets.

TABLE III
UNPREPROCESSED TEXT

Type of Rubric	Text
Economic Rubric	Itu adalah harga-harga perhiasan di pasar Cimahi
	Harga pasaran elektronik di pasar
	Elektronik-elektronik di pasar cimahi mahal
Criminal Rubric	Pencurinya hanya mencuri barang yang mahal
	Ada pencuri yang membawa kabur perhiasan
	Pencuri perhiasan yang kabur, ditangkap polisi
General Rubric	Toko-toko di pasar ramai
	Polisi membawa stempel untuk tiap toko
	Bawa vouchernya ya, untuk belanja di toko elektronik

TABLE IV
PREPROCESSED TEXT

Type of Rubric	Text
Economic Rubric	harga harga hias pasar cimahi
	harga pasar elektronik pasar
	elektronik elektronik pasar cimahi mahal
Criminal Rubric	curi curi barang mahal
	curi bawa kabur hias
	curi hias kabur tangkap polisi
General Rubric	toko toko pasar ramai
	polisi bawa stempel toko
	bawa voucher belanja toko elektronik

C. Proposed Method Implementation

The mTF-IDF-Assoc value was calculated by formulating modified TF, IDF and Assoc. The first thing that must be calculated is the mTF value which is the normalized term weighting based on the term frequency (*tk*) appearing in the document (*dj*) by dividing it by document length. mTF formula is the Equation 1 and the results of mTF calculation are presented in Table 5.

After finding the mTF value, the next step was to calculate IDF. Inverse Document Frequency (IDF) is the calculation of how the terms are widely distributed to the collection of relevant documents. IDF shows the term availability relationship in all documents. The smaller the number of documents containing the intended term, the bigger the IDF. IDF formula is the Equation 2 and the results of IDF calculation are presented in Table 6.

TABLE VI
IDF CALCULATION

Term	IDF
barang	0.954
bawa	0.352
belanja	0.954
cimahi	0.352
curi	0.477
elektronik	0.477
harga	0.653
hias	0.653
kabur	0.653
mahal	0.653
pasar	0.653
polisi	0.653
ramai	0.954
stempel	0.954
tangkap	0.954
toko	0.477
voucher	0.954

TABLE V
RESULTS OF MTF CALCULATION

Term	D1	D2	D3	D4	D5	D6	D7	D8	D9
barang	0	0	0	0.5	0	0	0	0	0
bawa	0	0	0	0	0.5	0.431	0	0.5	0.431
belanja	0	0	0	0	0	0	0	0	0.431
cimahi	0.431	0.793	0.431	0	0	0	0.5	0	0
curi	0	0	0	0.792	0.5	0.431	0	0	0
elektronik	0	0.5	0.683	0	0	0	0	0	0.431
harga	0.683	0.5	0	0	0	0	0	0	0
hias	0.431	0	0	0	0.5	0	0	0	0
kabur	0	0	0	0	0.5	0.431	0	0	0
mahal	0	0	0.431	0.008	0	0	0	0	0
pasar	0.431	0	0.431	0	0	0	0	0	0
polisi	0	0	0	0	0	0.431	0	0.5	0
ramai	0	0	0	0	0	0	0.5	0	0
stempel	0	0	0	0	0	0	0	0.5	0
tangkap	0	0	0	0	0	0.431	0	0	0
toko	0	0	0	0	0	0	0.792	0.5	0.431
voucher	0	0	0	0	0	0	0	0	0.431

The association pattern in each class can be found from a list of all terms in each class. For example, the transaction in economic rubric class is presented in Table 7.

TABLE VII
LIST OF TERMS IN ECONOMIC RUBRIC CLASS

Term	D1	D2	D3	Amount
barang	0	0	0	0
bawa	0	0	0	0
belanja	0	0	0	0
cimahi	1	0	1	2
curi	0	0	0	0
elektronik	0	1	1	2
harga	1	1	0	2
hias	1	0	0	1
kabur	0	0	0	0
mahal	0	0	1	1
pasar	1	0	1	2
polisi	0	0	0	0
ramai	0	0	0	0
stempel	0	0	0	0
tangkap	0	0	0	0
toko	0	0	0	0
voucher	0	0	0	0

From the data, the researchers carried out the searching process for item combination that met the requirement of the minimum support value in the database. Previously, the researchers determined the minimum support value. The parameters set experimentally to create association rules are 0.05 for minSupport and minConfidence. But for this example, the value of MinSupport was 0.5 as in the Equation 11. The results of support combination 1 itemset are presented Table 8 and the support combination 2 itemset are presented in Table 9.

$$Sup(X) = \frac{number\ of\ transactions\ containing\ X}{Total\ transaction} \quad (11)$$

TABLE VIII
SUPPORT VALUE FOR COMBINATION 1 ITEMSET

Itemset	Support
cimahi	1
harga	0.666667
pasar	0.666667
elektronik	0.666667

TABLE IX
SUPPORT VALUE FOR COMBINATION 2 ITEMSET

Itemset	Support
harga, cimahi	0.666667
harga, elektronik	0.666667
cimahi, elektronik	0.333333
cimahi, pasar	0.666667

The formation of association rules was carried out after all the association patterns with high-frequency number were found. This process is carried out by calculating the confidence value of the associative rule X → Y. Confidence is a measure that shows the relationship between 2 items conditionally. The formula is presented below as in the Equation 12.

$$Conf = \frac{number\ of\ transactions\ containing\ X \text{ and } Y}{Total\ transaction\ containing\ X} \quad (12)$$

The main objective in this stage was to find all association rules in each class that met the threshold (for example MinConfidence = 0.7). The implementation of finding association rules was shown in Table 10. The calculation of average confidence was carried out after all the rules were found, as shown in Table 11.

TABLE X
RULES OF ASSOCIATION

Association Rule	Support	Confidence
harga → cimahi	0.666667	1
harga → elektronik	0.666667	1
elektronik → harga	0.666667	1
elektronik → cimahi	0.666667	1

TABLE XI
AVERAGE OF CONFIDENCE

Association Rule	Average of Confidence
harga	0.75
cimahi	0.5
elektronik	0.75

$Assoc_{(ti)}$ was calculated by calculating the total comparison value of classes containing the term i ($C_{k,ti}$) with total term frequency for each class ($f_{ti,ck}$), as in the Equation 5. To consider the terms in each class, the researchers calculated the number of terms in documents that were the members of the class. The proposed term weighting can measure the level of strength of terms that were spread across classes. This measurement can be done using the *AssocBased*$_{(ti)}$ formula, as in equation 4. So that the results of the *AssocBased*$_{(ti)}$ calculation can be combined with the mTF and IDF to become mTF-IDF-Assoc. The example of calculation of proposed term weighting is presented in Table 12 and Table 13.

TABLE XII
CALCULATION OF $Assoc_{(ti)}$

Term	$AssocBased_{(ti)}$	$\sum_{j=1}^{k} \frac{C_{k,ti}}{f_{ti,ck}}$	$Log\left(\frac{C}{\sum_{j=1}^{k} \frac{C_{k,ti}}{f_{ti,ck}}}\right)$	$Assoc_{(ti)}$
barang	0	1	0.477121	0.48
bawa	0.75	0.666667	0.653213	1.40
belanja	0	1	0.477121	0.48
cimahi	0.5	0.4	0.875061	1.38
curi	0.5	0.25	1.079181	1.58
elektronik	0.75	0.5	0.778151	1.53
harga	0.75	0.333333	0.954243	1.70
hias	0	1	0.477121	0.48
kabur	0.75	0.5	0.778151	1.53
mahal	0	1	0.477121	0.48
pasar	0	0.5	0.778151	0.78
polisi	0	1	0.477121	0.48
ramai	0	1	0.477121	0.48
stempel	0	1	0.477121	0.48
tangkap	0	1	0.477121	0.48
toko	1	0.25	1.079181	2.08
voucher	0	1	0.477121	0.48

The term weighting results presented in table 13 are the results of the calculation between mTF, IDF and $Assoc_{(ti)}$. The formula is the Equation 6. The results of this calculation

was the final results that showed the importance of a term in each document. In this stage, the form of data changed from previously unstructured to structured. These results were processed to the stage, namely classification.

TABLE XIII
CALCULATION OF $mTF - IDF - Assoc$

Term	D1	D2	D3	…	D9
barang	0	0	0	…	0
bawa	0	0	0	…	0.2128
belanja	0	0	0	…	0.1961
cimahi	0.2086	0.3838	0.2086	…	0
curi	0	0	0	…	0
elektronik	0	0.3646	0.4977	…	0.3140
harga	0.7599	0.5566	0	…	0
hias	0.1343	0	0	…	0
kabur	0	0	0	…	0
mahal	0	0	0.1342	…	0
pasar	0.2189	0	0.2189	…	0
polisi	0	0	0	…	0
ramai	0	0	0	…	0
stempel	0	0	0	…	0
tangkap	0	0	0	…	0
toko	0	0	0	…	0.4272
voucher	0	0	0	…	0.1961

D. Performance Evaluation

In the classification stage, the Naïve Bayes, KNN and SVM method were used to evaluate the performance of the proposed term weighting. The data partitioned by 10 cross-fold validation and 70:30 split test train were used to measure the accuracy of the classification model made. The results of the multiclass classification were evaluated using the confusion matrix by calculating the average values of Precision, Recall, and Accuracy. Then, with the confusion matrix, the researchers analyzed how well the classifier could recognize data from more than two classes. The confusion matrix gave information about the level of accuracy, precision, and recall. It was intended to find out the proposed term weighting performance compared to the other term weighting like TF-IDF, TF-ICF, and TF-Assoc. The comparison is presented in Table 14 and Table 15.

TABLE XIV
THE COMPARISON RESULTS OF PRECISION AND RECALL VALUES WITH TF-IDF

Type of Rubric	10-Fold Cross Validation						Split 70:30					
	SVM Classifier		Naïve Bayes Classifier		KNN Classifier		SVM Classifier		Naïve Bayes Classifier		KNN Classifier	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Criminal Rubric	0.665	0.534	0.447	0.559	0.671	0.240	0.281	0.409	0.182	0.545	0.250	0.045
Disaster Rubric	0.894	0.822	0.778	0.780	0.685	0.632	0.806	0.685	0.675	0.712	0.407	0.507
Economic Rubric	0.783	0.552	0.662	0.540	0.739	0.399	0.436	0.706	0.560	0.824	0.550	0.647
General Rubric	0.728	0.856	0.744	0.744	0.646	0.726	0.821	0.785	0.859	0.705	0.755	0.614
Health Rubric	0.662	0.453	0.534	0.495	0.366	0.158	0.000	0.000	0.032	0.250	0.000	0.000
Political Rubric	0.793	0.709	0.579	0.764	0.731	0.321	0.432	0.514	0.250	0.459	0.313	0.135
Sports Rubric	0.914	0.824	0.922	0.845	0.715	0.716	0.875	0.689	0.780	0.811	0.457	0.303
Traffic Rubric	0.866	0.839	0.896	0.817	0.769	0.790	0.833	0.849	0.867	0.819	0.758	0.765
Weather Rubric	0.929	0.935	0.872	0.907	0.654	0.896	0.673	0.854	0.552	0.902	0.162	0.854

TABLE XV
THE COMPARISON RESULTS OF PRECISION AND RECALL VALUES WITH TF-ICF

Type of Rubric	10-Fold Cross Validation						Split 70:30					
	SVM Classifier		Naïve Bayes Classifier		KNN Classifier		SVM Classifier		Naïve Bayes Classifier		KNN Classifier	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Criminal Rubric	0.665	0.534	0.447	0.559	0.620	0.216	0.281	0.409	0.182	0.545	0.250	0.045
Disaster Rubric	0.894	0.822	0.778	0.780	0.676	0.632	0.806	0.685	0.675	0.712	0.425	0.507
Economic Rubric	0.783	0.552	0.662	0.540	0.774	0.399	0.436	0.706	0.560	0.824	0.611	0.647
General Rubric	0.728	0.856	0.744	0.744	0.647	0.726	0.820	0.785	0.859	0.705	0.750	0.616
Health Rubric	0.662	0.453	0.534	0.495	0.395	0.158	0.000	0.000	0.032	0.250	0.000	0.000
Political Rubric	0.793	0.709	0.579	0.764	0.730	0.312	0.432	0.514	0.250	0.459	0.250	0.108
Sports Rubric	0.914	0.824	0.922	0.845	0.726	0.714	0.875	0.689	0.780	0.811	0.463	0.303
Traffic Rubric	0.866	0.839	0.896	0.817	0.768	0.791	0.832	0.847	0.867	0.819	0.758	0.771
Weather Rubric	0.929	0.935	0.875	0.907	0.644	0.900	0.673	0.854	0.552	0.902	0.160	0.854

TABLE XVI
THE COMPARISON RESULTS OF PRECISION AND RECALL VALUES WITH TF-ASSOC

Type of Rubric	10-Fold Cross Validation						Split 70:30					
	SVM Classifier		Naïve Bayes Classifier		KNN Classifier		SVM Classifier		Naïve Bayes Classifier		KNN Classifier	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Criminal Rubric	0.665	0.534	0.445	0.554	0.620	0.216	0.281	0.409	0.179	0.545	0.250	0.045
Disaster Rubric	0.894	0.822	0.781	0.783	0.676	0.632	0.806	0.685	0.684	0.712	0.425	0.507
Economic Rubric	0.783	0.552	0.664	0.546	0.774	0.399	0.436	0.706	0.538	0.824	0.611	0.647
General Rubric	0.728	0.857	0.745	0.744	0.647	0.726	0.820	0.785	0.859	0.705	0.750	0.616
Health Rubric	0.662	0.453	0.534	0.495	0.395	0.158	0.000	0.000	0.032	0.250	0.000	0.000
Political Rubric	0.793	0.709	0.580	0.767	0.730	0.312	0.432	0.514	0.250	0.459	0.250	0.108
Sports Rubric	0.916	0.824	0.924	0.847	0.726	0.714	0.875	0.689	0.805	0.811	0.463	0.303
Traffic Rubric	0.866	0.839	0.895	0.818	0.768	0.791	0.832	0.847	0.868	0.821	0.758	0.771
Weather Rubric	0.929	0.935	0.874	0.907	0.644	0.900	0.673	0.854	0.544	0.902	0.160	0.854

TABLE XVII
THE COMPARISON RESULTS OF PRECISION AND RECALL VALUES WITH MTF-IDF-ASSOC

Type of Rubric	10-Fold Cross Validation						Split 70:30					
	SVM Classifier		Naïve Bayes Classifier		KNN Classifier		SVM Classifier		Naïve Bayes Classifier		KNN Classifier	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Criminal Rubric	0.703	0.544	0.486	0.618	0.814	0.172	0.297	0.500	0.162	0.500	0.333	0.045
Disaster Rubric	0.903	0.825	0.654	0.813	0.842	0.662	0.823	0.699	0.481	0.685	0.561	0.438
Economic Rubric	0.791	0.534	0.419	0.632	0.862	0.344	0.543	0.735	0.297	0.794	0.895	0.500
General Rubric	0.732	0.855	0.891	0.516	0.547	0.868	0.832	0.798	0.927	0.466	0.677	0.855
Health Rubric	0.712	0.495	0.273	0.663	0.821	0.242	0.000	0.000	0.042	0.750	0.000	0.000
Political Rubric	0.806	0.758	0.644	0.739	0.758	0.294	0.500	0.459	0.309	0.568	0.500	0.081
Sports Rubric	0.934	0.836	0.902	0.847	0.910	0.531	0.863	0.721	0.748	0.705	0.784	0.238
Traffic Rubric	0.855	0.843	0.868	0.814	0.848	0.783	0.831	0.862	0.841	0.799	0.859	0.734
Weather Rubric	0.940	0.938	0.612	0.910	0.827	0.875	0.685	0.902	0.183	0.854	0.366	0.829

TABLE XVIII
THE ACCURACY RESULTS FOR TESTED TERM WEIGHTING

Term Weighting	10-Fold Cross Validation			Split 70:30		
	SVM Classifier	Naïve Bayes Classifier	KNN Classifier	SVM Classifier	Naïve Bayes Classifier	KNN Classifier
TF-IDF	81.684%	77.913%	68.8348%	77.7926%	74.734%	61.5691%
TF-ICF	81.684%	77.913%	68.735%	77.7261%	74.734%	61.7686%
TF-Assoc	81.7039%	**77.9729**%	68.735%	77.7261%	**74.8005**%	61.7686%
mTF-IDF-Assoc	**82.3224**%	71.9673%	**70.3113**%	**79.3218**%	62.0346%	**70.5452**%

Table 18 shows the results of the accuracy average value for mTF-IDF-Assoc term weighting which was tested with the KNN classifier, Naïve Bayes and the SVM classifier. From these results, the researchers compared the accuracy value, so that the researchers analyzed the effect of modified TF-Assoc on the performance of the classification algorithm. The proposed method, mTF-IDF-Assoc weighting scheme used SVM classifier with 10-fold cross-validation, showed a better performance compared to Naïve Bayes classifier and KNN classifier. Moreover, the results of mTF-IDF-Assoc term weighting outperformed TF-ICF and TF-Assoc term weighting as supervised term weighting and also outperformed TF-IDF as unsupervised term weighting. It got the highest average accuracy on SVM classifier that was 82.3224%.

The proposed method, mTF-IDF-Assoc weighting scheme used SVM classifier, outperformed TF-ICF and TF-Assoc term weighting as supervised term weighting and also outperformed TF-IDF as unsupervised term weighting. The results showed that text classification accuracy average value for short text using the mTF-IDF-Assoc term weighting with the SVM classifier and 10-fold cross-validation was higher than the other term weighting, which was equal to 82.3224%. The results are presented in figure 4.

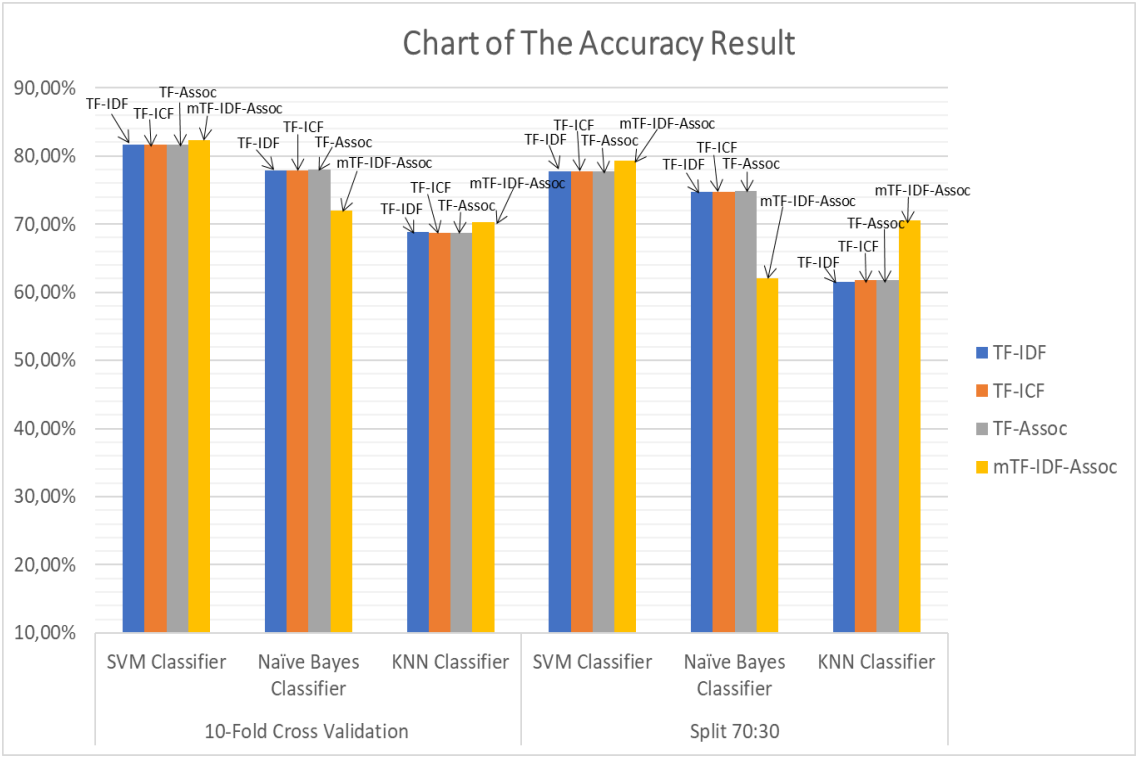

Fig. 4. Chart of The Accuracy Result

V. Conclusion

In this study, the researchers proposed a modified TF-Assoc supervised term weighting, namely mTF-IDF-Assoc. mTF-IDF-Assoc integrates the calculation of mTF value, that is a normalized term weighting based on the frequency term that appears in the document by dividing it by the length of the document (DL) then combining it with IDF and Assoc in calculating the weight of each word in the document. Therefore, it can provide information on the importance level of words in a class. This is proven by experiments that the classification process of tweet data from PR FM using mTF-IDF-Assoc weighting with SVM and 10-fold cross-validation segmentation technique has the highest level of accuracy compared to TF-IDF, TF-ICF and TF-Assoc.

In the future, the researchers will conduct experiments using different and larger dataset for text classification and also consider other combinations of parameters from term weighting.

References

[1] R. Luss and A. D'Aspremont, "Predicting abnormal returns from news using text classification," *Quant. Financ.*, vol. 15, no. 6, pp. 999–1012, 2015.

[2] J. Weissbock, A. A. A. Esmin, and D. Inkpen, "Using external information for classifying tweets," *Proc. - 2013 Brazilian Conf. Intell. Syst. BRACIS 2013*, pp. 1–5, 2013.

[3] S. Zhang, Z. Wei, Y. Wang, and T. Liao, "Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary," *Futur. Gener. Comput. Syst.*, vol. 81, pp. 395–403, 2018.

[4] A. T. Summarization, "Towards High Performance Text Mining : A TextRank-based Method for Automatic Text Summarization," vol. 8, no. 2, 2016.

[5] S. A. Yousif, V. W. Samawi, and I. Elkabani, "Arabic Text Classification : The Effect of the AWN Relations Weighting Scheme," no. July, 2017.

[6] 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.

[7] K. Gayathri, "Text Document Pre-Processing with the KNN for Classification Using the SVM."

[8] F. Sebastiani, "Machine Learning in Automated Text Categorization," vol. 34, no. 1, pp. 1–47, 2002.

[9] J. Mandowara, "Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification," vol. 6, no. 2, pp. 126–129, 2016.

[10] S. D. Jadhav and H. P. Channe, "Comparative Study of K-NN , Naive Bayes and Decision Tree Classification Techniques," vol. 5, no. 1, pp. 2014–2017, 2016.

[11] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen, "Scalable sentiment classification for Big Data analysis using Naïve Bayes Classifier," *Proc. - 2013 IEEE Int. Conf. Big Data, Big Data 2013*, pp. 99–104, 2013.

[12] M. Melucci, "Vector-Space Model," no. Encyclopedia of Database Systems, pp. 3259–3263, 2009.

[13] 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.

[14] 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.

[15] 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.

[16] I. K. Izzah and A. S. Girsang, "Association on Supervised Term Weighting Method for Classification on Data Twitter," *Int. J. Recent Technol. Eng.*, vol. 8, no. 6, pp. 859–863, 2020.

[17] T. Sabbah *et al.*, "Modified frequency-based term weighting schemes for text classification," *Appl. Soft Comput. J.*, vol. 58, pp. 193–206, 2017.

[18] 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.

[19] I. Alsmadi and G. K. Hoon, "Term weighting scheme for short-text classification: Twitter corpuses," *Neural Comput. Appl.*, vol. 8, pp. 1–13, 2018.

[20] 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.

[21] H. Wu and N. Yuan, "An Improved TF-IDF algorithm based on word frequency distribution information and category distribution information," pp. 211–215, 2018.

[22] F. Ren and M. G. Sohrab, "Class-indexing-based term weighting for automatic text classification," *Inf. Sci. (Ny).*, vol. 236, pp. 109–125, 2013.

[23] S. S. Ge, "An Improved TF-IDF Algorithm Based on Class

Discriminative Strength for Text Categorization on Desensitized Data," pp. 39–44, 2019.
[24] C. Le, P. W. C. Prasad, A. Alsadoon, L. Pham, and A. Elchouemi, "Text Classification : Naïve Bayes Classifier with Sentiment Lexicon," no. May, 2019.
[25] T. Fahrudin, J. L. Buliali, and C. Fatichah, "Enhancing The Performance of SMOTE Algorithm by Using Attribute Weighting Scheme and New Selective Sampling Method for Imbalanced Data Set," *Int. J. Innov. Comput. Inf. Control*, vol. 15, no. 2, pp. 423–444, 2019.
[26] 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.
[27] A. Purwarianti and A. I. M. Rule, "A Non Deterministic Indonesian Stemmer," no. July, pp. 1–5, 2011.