Indonesian News Classification using Weighted K-Nearest Neighbour

Muhammad Ihsan Amien Ismandiya [1], Yuliant Sibaroni [2]

[1,2,]Fakultas Informatika, Universitas Telkom, Bandung
[1]mrihsan@students.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id

Abstract

News is one means of information for the general public. Today, the number of news articles that reach 2 million articles per day can make it difficult for users to find news articles they want to read. In order to make it easier for users, most Indonesian newspapers classify their articles into certain categories, but there are also many blogs, or amateur articles that have not classified the news they circulated. Therefore this paper aims to categorize Indonesian language news using the weighted k-nearest neighbor method. In this paper there are several stages in classifying the news, namely preprocessing, feature extraction, and classification using wK-NN. The study used the wK-NN method where K = 6. In this study feature extraction was carried out in *unigram* and *bigram* which resulted in accuracy that was not much different. So it is recommended to use *unigram* because it is more efficient

Keywords: Classification, Prepocessing, feature extraction, weighted K-nearest neighbor

1. Introduction

News is new information about something that is happening, conveyed through different media such as print, broadcast, internet, or by word of mouth to a third person or many people[1]. Since the advent of the internet or the world wide web applications the world has never been more connected than ever.

News updates via the internet are increasing far more than ordinary media, because news via the internet can be published by anyone not only from journalistic organizations. At present every day there are more than 2 million news articles that are released every day on the internet[2]. News articles from non-journalistic organizations or individuals,u are not categorized, which makes searching for certain categories or avoiding certain categories more difficult for common internet users.

Most of these articles are in text form with a variety of different categories. With text classification that automarically assign text documents into predetermined categories managing the number of articles every day will be easier. The existence of a text classification system allows users to find articles more easily by having them search for the desired category.

Uncategorized news on the internet limit the search range for internet users, because most people can only read a limited amount of articles and sometimes they only want a certain category and do not want to waste their time searching for that category.

Topics and Limitations

The problem of categorizing news is always faced by journalists or newspaper editors. This happens because every day the number of news to be processed can exceed hundreds of articles, therefore categorizing news automatically using software will help the newspaper. At present many methods have been created to classify news in foreign languages, but for news in Indonesian the number is still small.

This research takes the topic of classifying news in Indonesian automatically using the WK-NN method from S. Sahara who carries out research in text classification for sentiment analysis [3]. Another research that is also a refered to in this study is the classification of texts that focus on how to use WK-NN in classifying data [5]. Another research that focus on news classification using KNN was by Winarko [4] Therefore it can be concluded that WK-NN classification method of has not been widely used in the classification of Indonesian language news.

This can be understood because most news articles uploaded to the internet are written in English. So that the implementation of the text classification used still refers to English. Therefore it is necessary to identify features specifically for the classification of Indonesian texts.

Features selection is a well-defined problem for text classification. The aim is to increase classification effectiveness, computational efficiency and or both. It is common to use n-gram method to extract these features, n is the length of the number of words used to classify the text or document. The study was conducted for 6 (six) news categories namely economics, sports, technology, automotive, lifestyle, and entertainment. The number of sample is 100 (one hundred) articles taken at random.

Objective

The objective of this research is to analyze and design a news article classification system using the wK-NN classification method which is commonly used to classify many labels. This study determines the accuracy and feature with the biggest weight to differentiate one category from the others. The journal is organized in the following order : Chapter 2 describes a review of relevant previous research that has been carried out up to now. Followed by chapter 3 with an explanation of the classification system built. Chapter 4 describes the evaluation of the research that has been done. Finally, chapter 5 presents conclusions and suggestions.

2. Related Studies

There are several journals and related studies that discuss the classification of texts using K-nearest negihbour, one of which is the classification of texts using K-nearest neighbors by Gongde [6] which discusses text classification using not only KNN but also using Rocchio classifier. In this study, KNN and Rocchio are used to help each other in classification so that classification accuracy is greater. The results of this research show that KNN alone is more accurate than the combination of KNN and Rocchio methods.

Other study that does text classification using K-nearest neighbor is Classifying news using KNN with K-means clustering by Gupta and Akansha [7]. K-means are used to classify keywords in each news article and KNN are used to classify the articles themselves. This research achieve a classification accuracy of 93.28%.

The classification using KNN outside the news is performed on sentiment analysis research on reviews on android by Sahara, Sucitra [3]. The purpose of this research is to examine the results of comments on the app store for Android games to find out what keywords make a good game or not. Thhis research found that

the accuracy of a positive review is 75.50%

Classification using KNN and sentiment analysis is also used in Twitter sentiment analysis research [8]. The purpose of this research is to find the biggest opinion on Twitter and made into positive, negative and neutral segments. KNN is used as a sentiment orientation classification while lexicon is used to identify neutral sentences in the preprocessing query stage and also evaluating negative sentences. It was found that greater accuracy than the combination of the two methods is for the combination of the lexicon approach as the process of determining the negation, evaluation and KNN algorithm method produces an accuracy of 78% while, the combination of the lexicon approach as the process of determining the negation, evaluation and KNN algorithm as a neutral sentiment determination obtained 82% accuracy.

Text categorization using KNN was also done by Han et al. using weighted k-nearest neighbor [9]. This research uses weight adjusted KNN to get better computational optimization because feature selection has been optimized to produce great accuracy and fast computation.

Research using the same topic was also carried out by Hechenbichler et al. [5]. This research aims to use the KNN and also LOESS (Local Regression technique) in order to get a second positive impact from using the methods mentioned.

Other study that also use the KNN method combined with other methods is carried out by E. Winarko in the Use of KNN for Classification of Non-Grouped News Texts [4]. This research found that by using Stc in the initial classification which only grouped in the thematic news and then continued by KNN can group the news to the existing groups with high accuracy.

The KNN method is also used by Yong [10]. This study uses a compressed training set and samples that are close to the category boundaries are removed, which results in the disappearence multi peak effect. The training sample clustering is done using k-means clustering and the results of the cluster center are used as a new training set. And also use the weight value to determine the importance of the training sample. The findings of this research are not only reducing the number of training sets but also increasing the accuracy of the KNN.

The KNN classification method for texts is also carried out by Khamar who focuses on classifying short texts [11]. This study discussesthe use of KNN in short texts and compares the KNN method with SVM and Naïve Bayes. It was found that KNN has a higher accuracy than the two methods mentioned.

The KNN method for text classification is also carried out in Arabic by Al-Shalabi and Obeidat [12]. This study not only discusses the Arabic text classification with the KNN method but also with the N-grams method. The results of this study is that text classification will get higher accuracy if using N-grams, namely unigram and bigram before doing classification using KNN. Therefore the KNN method is the most commonly used method in articles which have more than 2 (two) labels. The KNN method can also be used in various languages.

3. System Built

The system built is presented in the form of stages of classification that starts from raw data into data that has been given a classification according to the system being built. The classified news is Indonesian language news. This design refers to the commonly used KNN method.

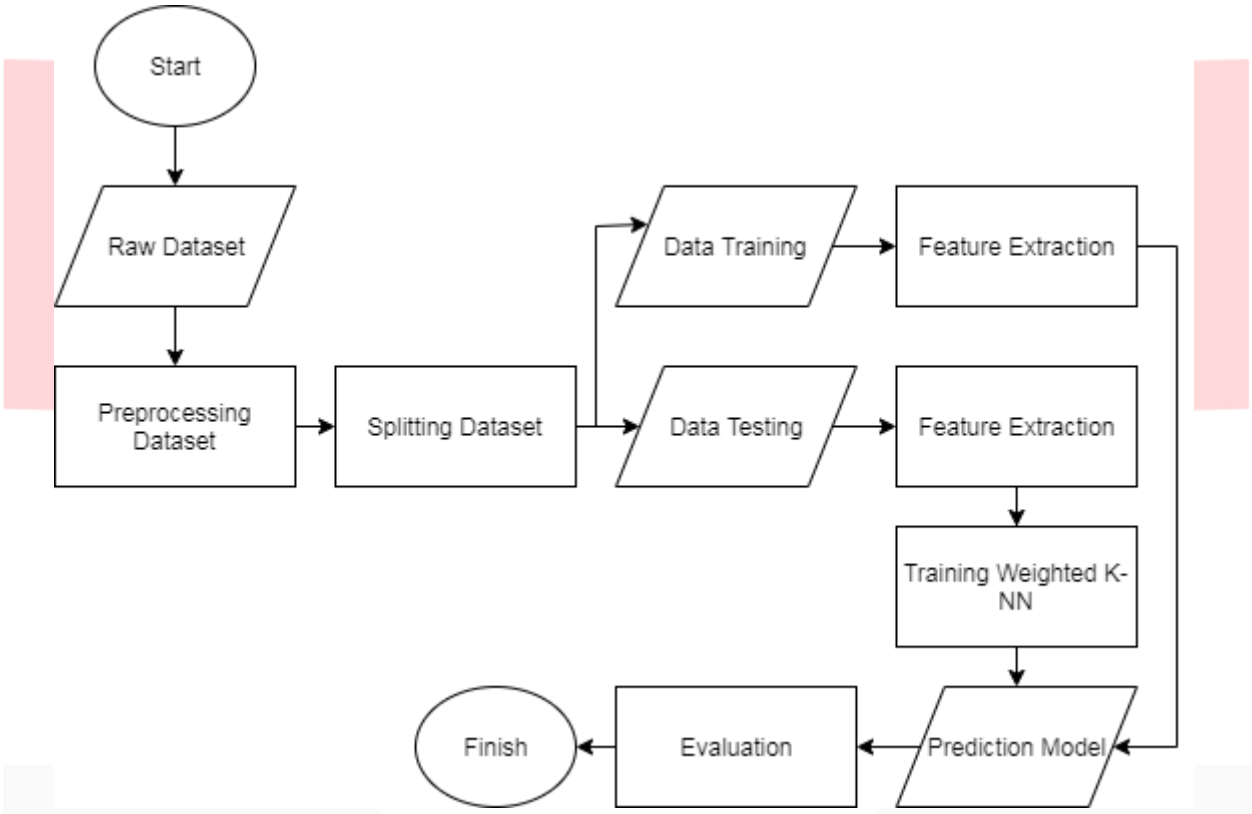

Image 1 System Diagram

3.1 Dataset

The data used in this study has 6 labels, namely economics, sports, technology, automotive, lifestyle, and entertainment, totaling 1000 articles. Most of the data is taken from kompas.com which already has their respective classifications. This dataset is used to build a classification system using weighted k-nearest neighbors. Based on the existing label, the data is divided into 6 classes according to the initial classification obtained from Kompas.com from 2019 from May to June. The dataset is complemented with articles from tribunnews.com and liputan6.com. Data from these two websites is used so that the dataset is not entirely from 1 (one) source.

3.2 Preprocessing Dataset

Preprocessing is the process carried out to transform raw data into more structured data. The first stage carried out in Preprocessing is the process of transforming data. Text preprocessing is the process of changing structured and semi-structured text into structured vector models[13]. This process is done to reduce noise or unimportant data as well as the vocabulary size. This step must be done before text mining can be performed. Preprocessing is an initial step through a dataset to change data that is not in accordance with the desired format. The processes carried out are case folding, tokenization, stopwords removal and stemming to simplify further data processing.

- Case folding: change all capital letters to lowercase, case folding also removes non-letter characters. For example like (, , !,_,?,&, dll.)
- Tokenization: solving sentences or phrases in discrete form
- Stopword removal : eliminating words that are not important but often exist [14]. Examples of stopwords that often appear in Indonesian are (“aku”, “yang”, “dan”, dll.), The stopword used is derived from the results of Jelita's research [15]
- Stemming : the process of turning words into root words by eliminating prepositions and prefixes, suffixes and infixs. The stemming used comes from the 'Sastrawi' library based on the library from Adriani's research [16]

Table 1 The data before preprocessed

No Article	ARTICLE CONTENTS	CATEGORY
160	Kejadian kecelakaan lalu lintas kembali terjadi, kali ini melibatkan Nissan Grand Livina yang hilang kendali menabrak Apotek Senopati di Jakarta...	AUTOMOTIVE
242	Bank Indonesia (BI) memproyeksikan pertumbuhan ekonomi Indonesia pada kuartal III-2019 akan relatif stagnan...	ECONOMY
404	Salah satu member girlband Girls' Generation atau SNSD, Taeyeon, akan menyanyikan ulang soundtrack film Frozen 2 dalam bahasa Korea....	ENTERTAINMENT
665	Kebiasaan yang kita jalankan setiap hari ternyata bisa membantu mencegah terjadinya stroke...	LIFESTYLE
803	Laga akbar Liverpool vs Arsenal di pentas Piala Liga Inggris bukan hanya untuk mencari kemenangan....	SPORTS
993	Oppo Reno Ace dijadwalkan akan meluncur di China 10 Oktober mendatang. Dalam teaser yang dirilis Oppo...	TECHNOLOGY

Table 2 Data after being preprocessed

No Article	ARTICLE CONTENTS	CATEGORY
160	jadi celaka lalu lintas jadi kali libat nissan grand livina hilang kendali tabrak apotek senopati jakarta…	AUTOMOTIVE
242	bank indonesia bi proyeksi tumbuh ekonomi indonesia kuartal ii relatif stagnan...	ECONOMY
404	salah satu member girlband girls generation snsd taeyeon nyanyi ulang soundtrack film frozen bahasa korea...	ENTERTAINMENT
665	biasa kita jalan hari nyata bantu cegah jadi stroke...	LIFESTYLE
803	laga akbar liverpool vs arsenal pentas piala liga inggris bukan cari menang...	SPORTS
993	oppo reno ace jadwal luncur china oktober datang teaser rilis oppo...	TECHNOLOGY

3.3 Splitting Dataset

The dataset is separated into 2 parts, namely testing and training data. Testing data consist of 100 (one hundred) articles randomly selected and taken out from the entire set. The remaining data is called training data which will be used to train to the classification model.

3.4 Feature Extraction

Feature extraction is the process of taking the characteristics that can describe the object [15]. In this case the features are sequence of words that distinguish a class of articles from the others. By using the n-gram and the followed by TF-IDF methods, we will get a sequence of words that characterize a label or a class of articles.

3.5 N-gram

N-gram is a collection of N words in a paragraph. The N-gram is formed by moving the boundary one word at a time forward. It is used as a space vector value to characterize the class or label of a document or text[12]. For example if there is a sentence "Bapak meminjam mobil ibu" then the n-gram obtained is:

Table 3 an example of using ngram in sentences

N = 1 (Unigram)	N = 2 (Bigram)
Bapak	Bapak meminjam
meminjam	meminjam mobil
mobil	mobil ibu
ibu	-

N-gram used in this study are unigram and bigram. The higher the n value in n-gram, the more term variations obtained and the higher the computational resources needed. In this research, a computer with an Intel I7 computer with 8 gigabyte RAM memory is only able to do n-grams to n = 2.

3.6 TF-IDF

Term Frequency - Inverse Document Frequency or TF-IDF is an algorithm method that is useful for calculating the weight of each commonly used word. The terms of the Term Frequency and Inverse Document Frequency components are explained below.

TF (Term Frequency) is the frequency of occurrence of a term in the relevant document. The greater the number of occurrences of a term (high TF) in the document, the greater the weight or the greater the value of conformity.

IDF (Inverse Document Frequency) is a calculation of how the terms are widely distributed in the collection of documents concerned. IDF shows the availability of relationship of a term in all documents. The smaller the number of documents containing the intended term, the greater the IDF value. Whereas the Inverse Document Frequency (IDF) is calculated using the following formula[18]

$$IDF_j = \log(D/df_j) \quad (1)$$

Where D is the number of all documents in the collection while df_j is the number of documents containing the term (t_j). The formula generally used for TF-IDF calculations is

$$w_{ij} = tf_{ij} \times idf_j \quad (2)$$

$$w_{ij} = tf_{ij} \times \log(D/df_j) \quad (3)$$

Where w_{ij} is the term weight (tj) of the document (d_i). Whereas tf_{ij} is the number of occurrences of term (t_j) in the document (d_i). D is the total number of documents in the database and df_j is the number of documents containing term (t_j) (there is at least one word, term (t_j)).

Table 4 Example of unigram term frequency calculation in the article

Term	Article 160	Article 242	Article 404	Article 665	Article 803	Article 993
alat	4	0	0	0	0	0
bagi	1	1	0	1	0	2
celana	0	0	1	0	0	0
hingga	0	4	1	0	1	0
jalan	5	0	0	0	2	0
lain	3	3	1	0	0	1

Table 5 An example of unigram TF-IDF calculation in the article

Term	Article 160	Article 242	Article 404	Article 665	Article 803	Article 993
alat	0.126326	0	0	0	0	0
bagi	0.024832	0.03361	0	0.040367	0	0.060317
celana	0	0	0.082639	0	0	0
hingga	0	0.080202	0.035571	0	0.050759	0
jalan	0.07584	0	0	0	0.100587	0
lain	0.059947	0.081136	0.040918	0	0	0.040979

Table 6 Example of calculation of the term frequency bigram in the article

Term	Article 160	Article 242	Article 404	Article 665	Article 803	Article 993
alat komunikasi	1	0	0	0	0	0
bagi file	0	0	0	0	0	1
celana warna	0	0	1	0	0	0
hingga faktor	0	1	0	0	0	0
jalan tanding	0	0	0	0	2	0
lain beda	1	0	0	0	0	0

Table 7 Examples of TF-IDF bigram calculations in the article

Term	Article 160	Article 242	Article 404	Article 665	Article 803	Article 993
alat komunikasi	0.048254	0	0	0	0	0
bagi file	0	0	0	0	0	0.06686
celana warna	0	0	0.070901	0	0	0
hingga faktor	0	0.064632	0	0	0	0
jalan tanding	0	0	0	0	0.15271	0
lain beda	0.048254	0	0	0	0	0

3.7 Weighted K-Nearest Neighbor Classification

Weighted K-Nearest Neighbor (wK-NN) is one of the data classification methods developed from ordinary K-Nearest Neighbor The difference is on wK-NN weight is added to the search for learning data with the nearest object [5].

The wK-NN algorithm is quite simple. If K-NN is only based on the shortest distance from the query to the training data to show the K-NN, the wK-NN will not only take into account the query to the training data, but the training data is projected into a multi-dimensional space, where each dimension represents a feature from data.

This space is divided into sections based on the classification of training data. A point on this space is marked as C. Class C is the most common classification found on the neighboring K of that point. To find out the closest data distance in general can be calculated via the euclidian distance formula as follows[5].

$$d(x_i, x_j) = \sqrt{\sum_{i=1}^{n} (x_i - x_j)^2} \quad (4)$$

Where the matrix $d(x_i, x_j)$ is the scalar distance from the vector $x_i$and vector x_j. And $x_i$is the testing data while $x_j$is the training data. In the training phase this formula is used only to store feature vectors and classify training data.

In the classification phase the same features are used to calculate and test the testing data. K distances of this new vector to all training data vectors are calculated and taken. The new point classification is included in the class classification with the heaviest weight of these points.

At this stage the results of preprocessing training data after feature extraction will be classified in classes that have been determined using the wK-NN method.

3.8 Prediction Model

The prediction model is the result of the classification of testing data that has been labeled. The classification of this testing data will be compared to the training data to produce the accuracy of the classification program. This prediction model is trained from the training data so that the classification prediction will have high accuracy.

3.9 Evaluation

In the Weighted K-Nearest Neighbor process, 100 sample articles will be taken randomly to be used as test data. Training data is the remainder of the total number of articles. This classification program will be run five times. Each time the experiment uses testing data that are different from the total number of articles to produce new predictive results.

To find out the optimal K and the effect of weighting to the results from wK-NN, the variable weight used is ‘uniform’ and ‘distance’. Uniform means all points in each neighbor have equal weights. While 'distance' means weighting value is the inverse of distance. In this case neighbors who are closer to the query point will have a greater influence than neighbors who have greater distances.

4. Experiments & Evaluation Results

After going through the testing process, an analysis of the results of the tests is carried out. In this section we will explain the results of the test in the form of measurement metrics, namely Accuracy. And also features that stand out for each category as well as the optimal amount of K in the classification process.

4.1 Test Results

In this research several testing stages have been carried out. The test aims to determine the accuracy of the results of the classification of articles using the wK-NN. The first testing phase is to build a classification model wK-NN using features taken from unigram and bigram. The features obtained from the n-gram will be weighted again using the TF-IDF method and weighting of each term is done according to the weight of each article. In this process, testing was carried out five times using testing data taken randomly totaling 100 articles.

Using the TF-IDF method will increase accuracy because the weight of each term will be more related to the article categories in the training data. TF-IDF will produce terms that can be used to accurately label an article according to the trained categories. This wK-NN based classification process will be carried out several

times to study the impact of K and weight variable. The K values used are 3,6,12,31 and the weight variable used is 'uniform' and 'distance'.

Table 8 The results of the accuracy calculation where the variable weight = uniform

	Weight = uniform							
	Unigram				**Bigram**			
	K =3	**K = 6**	**K = 12**	**K = 31**	**K =3**	**K = 6**	**K = 12**	**K = 31**
Test 1	98.0%	96.0%	94.0%	95.0%	94.0%	95.0%	95.0%	93.0%
Test 2	97.0%	97.0%	94.0%	88.0%	95.0%	96.0%	91.0%	87.0%
Test 3	95.0%	97.0%	94.0%	96.0%	95.0%	94.0%	95.0%	92.0%
Test 4	98.0%	92.0%	95.0%	94.0%	95.0%	92.0%	90.0%	91.0%
Test 5	93.0%	92.0%	95.0%	90.0%	90.0%	93.0%	94.0%	90.0%
Average	96.2%	94.8%	94.4%	92.6%	93.8%	94.0%	93.0%	90.6%

Table 9 The results of the accuracy calculation where the variable weight = distance

	Weight = distance							
	Unigram				**Bigram**			
	K =3	**K = 6**	**K = 12**	**K = 31**	**K =3**	**K = 6**	**K = 12**	**K = 31**
Test 1	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Test 2	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Test 3	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Test 4	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Test 5	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Average	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

4.2 Analysis of Test Results

Based on the tables 8 and 9 above, the predicted results for weight = uniform unigram produce an average value higher than the average value of bigram where the most optimal k is k = 3. While the predicted results with weight = distance produced accuracy of 100%. This is because if you use uniform variable classification, you will choose the label most found in K, while using the classification distance variable will give more weight to neighboring points that are closer than all the specified K number.

The difference between unigram and bigram can be clearly seen by observing the produced features fro each category. Features that have a heavy weight in each category are as listed in tables 10 and 11 below. Table 10 is for unigram cases, and table 11 is for bigram cases.

Table 10 Unigram features that have large weight values

Category	Term
Automotive	'automobile', 'toyota', 'transmisi', 'h'ybrid', 'mesin'
Economy	'transaksi', 'pt', 'bank', 'aset', 'usaha'
Entertainment	'ahmad', 'baim', 'peran', 'artis', 'film'
Lifestyle	'sakit', 'tubuh', 'sehat', 'usia', 'desainer'
Sports	'liverpool', 'liga', 'vs', 'ac', 'gol'
Technology	'modem', 'mobile', 'huawei', 'ponsel', 'chipset'

Table 11 Bigram features that have large weight values

Category	Term
Automotive	'automobile center', 'toyota astra', 'transmisi otomatis', 'engine hybrid', ' mesin mobil'
Economy	'transaksi uang', 'pt sbs', 'bank bjb', 'aset indonesia', 'usaha milik'
Entertainment	'raffi ahmad', 'baim wong', 'artis peran', 'artis kondang', 'tanjak film'
Lifestyle	'tidak sakit', 'tubuh bakar', 'lebih sehat', 'usia tahun', 'desainer muda'
Sports	'vs liverpool', 'liga italia', 'vs chelsea', 'ac milan', 'gol sebut'
Technology	'lengkap modem', 'mobile banking', 'huawei mate', 'ponsel lipat', 'chipset kirin'

5. Conclussions

The design of wK-NN with unigram and bigram features produces satisfactory accuracy performance. To get the most optimal value wK-NN the value K is 3 and variable weight is equal to 'distance' because the results and efficiency of the program will produce high values.

The features that have the greatest weight will vary depending on the testing data used. Therefore it is possible that the same features have the same weight, but in other testing data will be different. The bigram feature will produce less weight compared to the unigram feature, this is because the terms that have a heavy weight on the unigram will be broken up or separated and the weight will be reduced at bigram.

In order to increase program efficiency, a deeper analysis must be carried out on other weighting methods or data preprocessing may be added in addition to Indonesian. This is because in the article there are terms or words that are not Indonesian. Therefore, for further research should use data processing not only for Indonesian but also for other languages.

From this experiment it can be seen that the differences between unigram and bigram accuracy are not significant. It also understood that bigram produces more terms and therefore requires more processing resources. Therefore generally it is recommended to use unigram because it has better efficiency with acceptable accuracy.

Bibliography

[1] M. Stephens, *A history of news*. Oxford University Press, 2007.
[2] Puranjay Singh, "2 Million Blog Posts Are Written Every Day, Here's How You Can Stand Out : MarketingProfs Article," 2015. [Online]. Available: http://www.marketingprofs.com/articles/2015/27698/2-million-blog-posts-are-written-every-day-heres-how-you-can-stand-out. [Accessed: 15-Sep-2018].
[3] S. Sahara, "Penerapan Metode K-Nearest Neighbors Untuk Analysis Sentiment Review Game Pada Android," *J. Evolusi*, vol. 4, pp. 38–44, 2016.
[4] E. Winarko, "Penggunaan Knn (K-Nearst Neighbor) Untuk Klasifikasi Teks Berita Yang Tak-Terkelompokkan Pada Saat Pengklasteran Oleh Stc (Suffix Tree Clustering)," vol. IX, no. 1, 2015.
[5] K. Hechenbichler and K. Schliep, "Weighted k-Nearest-Neighbor Techniques and Ordinal Classification," *Mol. Ecol.*, vol. 399, no. January 2004, p. 17, 2004.
[6] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Using kNN model for automatic text categorization," *Soft Comput.*, vol. 10, no. 5, pp. 423–430, 2006.
[7] A. Gupta, "Non-Probabilistic K-Nearest Neighbor for Automatic News Classification Model with K-Means Clustering," vol. 2, pp. 1–6.
[8] D. Rosdiansyah, "Analisis Sentimen Twitter Menggunakan Metode K-Nearest Neighbor dan Pendekatan Lexicone," *Tugas Akhir Jur. Tek. Inform.*, pp. 1–15, 2014.
[9] E.-H. Han, G. Karypis, and V. Kumar, "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification," pp. 53–65, 2001.
[10] Z. Yong, L. Youwen, and X. Shixiong, "An Improved KNN Text Classification Algorithm Based on Clustering - Yong, Youwen, Shixiong - 2009.pdf," vol. 4, no. 3, pp. 230–237, 2009.

[11] K. Khamar, "Short Text Classification Using kNN Based on Distance Function," *Ijarcce.Com*, vol. 2, no. 4, pp. 1916–1919, 2013.
[12] R. Al-Shalabi and R. Obeidat, "Improving KNN Arabic Text Classification with N-Grams Based Document Indexing," *Proc. Sixth ...*, pp. 108–112, 2008.
[13] D. Chicco, "Ten quick tips for machine learning in computational biology.," *BioData Min.*, vol. 10, no. 35, p. 35, 2017.
[14] E. Dragut, F. Fang, and C. Yu, "Stop Word and Related Problems in Web Interface Integration," *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 349–360, 2009.
[15] A. Jelita, "Effective Techniques for Indonesian Text Retrieval," *Ph.D Thesis*, pp. 1–286, 2007.
[16] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. Williams, "Stemming Indonesian: A confix-stripping approach.," *ACM Trans. Asian Lang. Inf. Process.*, vol. 6, Jan. 2007.
[17] Adiwijaya, "APLIKASI MATRIKS DAN RUANG VEKTOR," *GRAHA ILMU*, 2014.
[18] D. Sierra, "Algoritma TF — IDF - Delta Sierra - Medium." [Online]. Available: https://medium.com/@dltsierra/algoritma-tf-idf-633e17d10a80. [Accessed: 20-Dec-2019].

Attachment

Table 12 example of a unigram prediction table

doc#	true class	predicted class		doc#	true class	predicted class		doc#	true class	predicted class
665	3	5		309	1	1		191	1	1
803	4	4		878	5	5		361	2	2
242	1	1		552	3	3		563	3	3
160	0	0		306	1	1		741	4	4
286	1	1		498	2	2		285	1	1
541	3	3		896	5	5		900	5	5
625	3	3		661	3	3		481	2	2
743	4	4		817	4	4		354	2	2
569	3	1		397	2	2		775	4	4
404	2	2		40	0	0		113	0	0
16	0	0		231	1	1		360	2	2
993	5	5		685	4	4		505	3	3
961	5	5		863	5	1		615	3	3
11	0	0		166	0	0		963	5	5
39	0	0		364	2	2		490	2	2
860	5	5		366	2	2		659	3	3
818	4	4		622	3	3		737	4	4
95	0	0		138	0	0		57	0	0
600	3	3		770	4	4		499	2	2
272	1	1		714	4	4		714	4	4
435	2	2		960	5	5		583	3	3
503	3	3		662	3	3		931	5	5
656	3	3		520	3	3		682	4	4
807	4	4		297	1	1		669	4	4
76	0	0		898	5	5		999	5	5
178	1	1		688	4	4		102	0	0
18	0	0		177	1	1		66	0	0
626	3	3		314	1	1		328	1	1
391	2	3		988	5	5		742	4	4
102	0	0		931	5	5		724	4	4
219	1	1		865	5	5		940	5	5
245	1	0		793	4	4		148	0	0
469	2	2		955	5	5		629	3	3
210	1	1		232	1	1		258	1	5
296	1	1		255	1	1		128	0	0
143	0	0		173	1	1		788	4	4
884	5	1		651	3	3		302	1	1
896	5	5		717	4	4		94	0	0
411	2	2		337	2	2		654	3	3
96	0	0		171	1	1		637	3	3
189	1	1		113	0	0		441	2	2
71	0	0		870	5	5		982	5	5
635	3	3		721	4	4		604	3	3

doc#	true class	predicted class		doc#	true class	predicted class		doc#	true class	predicted class
437	2	2		401	2	2		372	2	2
427	2	2		120	0	0		461	2	2
410	2	2		33	0	0		986	5	5
677	4	4		90	0	0		363	2	2
683	4	4		440	2	2		185	1	1
306	1	1		63	0	0		303	1	1
424	2	2		199	1	1		329	1	1
105	0	0		119	0	0		548	3	3
363	2	2		122	0	0		837	5	5
291	1	1		858	5	5		171	1	1
14	0	0		537	3	3		680	4	4
484	2	2		689	4	4		269	1	5
202	1	1		828	4	4		428	2	2
592	3	3		947	5	5		506	3	3
817	4	4		849	5	5		765	4	0
808	4	4		887	5	5		213	1	1
62	0	0		802	4	4		715	4	4
830	4	4		620	3	3		785	4	4
963	5	5		43	0	0		448	2	2
834	5	5		357	2	2		76	0	0
207	1	1		624	3	3		990	5	5
873	5	5		918	5	5		572	3	3
687	4	4		114	0	0		539	3	3
737	4	4		860	5	5		399	2	2
947	5	5		670	4	4		215	1	1
153	0	0		623	3	3		286	1	1
226	1	1		282	1	1		130	0	0
243	1	1		182	1	1		801	4	4
268	1	1		404	2	2		333	1	1
443	2	2		644	3	3		139	0	0
378	2	2		700	4	4		390	2	2
26	0	0		545	3	3		560	3	3
459	2	2		587	3	3		497	2	2
81	0	0		542	3	3		158	0	0
380	2	2		563	3	3		322	1	5
908	5	5		358	2	2		674	4	4
333	1	1		54	0	0		774	4	4
342	2	2		866	5	5		307	1	1
549	3	3		869	5	5		951	5	5
402	2	2		635	3	3		105	0	0
22	0	0		239	1	1		479	2	3
517	3	3		411	2	2		161	0	0
786	4	4		942	5	1		962	5	5
619	3	3		183	1	1		947	5	5
167	0	0		946	5	5		985	5	5
420	2	2		719	4	0		621	3	3

doc#	true class	predicted class		doc#	true class	predicted class		doc#	true class	predicted class
6	0	0		141	0	0		419	2	2
281	1	1		116	0	0		515	3	3
746	4	4		394	2	2		809	4	3
470	2	2		925	5	5		811	4	4
806	4	4		468	2	2		978	5	5
527	3	3		367	2	1		16	0	0
678	4	4		627	3	2		555	3	3
516	3	3		547	3	3		854	5	5
939	5	5		915	5	5		675	4	4
539	3	3		155	0	0		611	3	3
194	1	1		665	3	5		194	1	1

doc#	true class	predicted class		doc#	true class	predicted class
531	3	3		861	5	5
574	3	3		506	3	3
642	3	3		585	3	3
225	1	1		10	0	0
738	4	4		189	1	1
626	3	3		104	0	0
813	4	4		415	2	2
250	1	1		140	0	0
273	1	1		888	5	5
427	2	2		1	0	0
459	2	2		65	0	0
279	1	1		555	3	3
100	0	0		430	2	2
587	3	3		454	2	2
592	3	3		276	1	1
433	2	2		958	5	5
527	3	3		601	3	3
926	5	5		926	5	5
401	2	2		577	3	3
804	4	4		25	0	0
664	3	3		492	2	2
944	5	5		310	1	1
502	3	3		893	5	5
407	2	2		757	4	4
389	2	2		909	5	5
294	1	1		862	5	5
741	4	4		635	3	3
684	4	4		479	2	3
602	3	3		345	2	2
717	4	4		167	0	0
443	2	2		607	3	3
939	5	5		755	4	0
561	3	3		534	3	3

doc#	true class	predicted class		doc#	true class	predicted class
18	0	0		734	4	4
103	0	0		977	5	5
656	3	3		264	1	1
75	0	0		940	5	5
665	3	5		937	5	5
596	3	3		396	2	2
518	3	3		306	1	1
87	0	0		72	0	0
319	1	1		713	4	0
721	4	4		285	1	1
314	1	1		850	5	5
904	5	5		134	0	0
233	1	1		541	3	3
11	0	0		211	1	1
824	4	4		858	5	5
934	5	5		29	0	0
845	5	5		596	3	3
135	0	0		675	4	4
120	0	0		86	0	0
85	0	0		263	1	3
479	2	3		202	1	1
455	2	2		70	0	0
952	5	5		255	1	1
304	1	1		543	3	3
619	3	3		477	2	2
761	4	4		998	5	5
669	4	4		764	4	4
166	0	0		268	1	1
940	5	5		608	3	3
975	5	5		309	1	1
676	4	4		730	4	4
48	0	0		866	5	5
767	4	4		304	1	1
683	4	4		586	3	3
790	4	4		198	1	1
950	5	5		97	0	0
872	5	5		475	2	2
538	3	3		748	4	4
107	0	0		867	5	5
967	5	5		128	0	0
938	5	1		947	5	5
151	0	0		811	4	4
207	1	1		216	1	1
723	4	4		8	0	0
903	5	5		224	1	1
28	0	0		908	5	5

doc#	true class	predicted class		doc#	true class	predicted class
795	4	4		647	3	3
897	5	5		417	2	2
793	4	4		315	1	5
76	0	0		388	2	2
943	5	5		843	5	5
526	3	3		201	1	1
848	5	5		173	1	1
161	0	0		337	2	2
484	2	2		521	3	3
277	1	1		96	0	0
825	4	4		677	4	4
397	2	2		84	0	0
10	0	0		69	0	0
328	1	1		821	4	4
257	1	1		833	5	5
357	2	2		527	3	3
597	3	3		766	4	4
663	3	3		44	0	0
377	2	2		803	4	4
456	2	2		860	5	5
276	1	1		598	3	3

Table 13 example of a bigram prediction table

doc#	true class	predicted class		doc#	true class	predicted class		doc#	true class	predicted class
665	3	0		309	1	1		191	1	1
803	4	4		878	5	5		361	2	2
242	1	1		552	3	3		563	3	3
160	0	0		306	1	1		741	4	4
286	1	1		498	2	2		285	1	1
541	3	3		896	5	5		900	5	5
625	3	3		661	3	3		481	2	2
743	4	4		817	4	4		354	2	2
569	3	1		397	2	2		775	4	4
404	2	2		40	0	0		113	0	0
16	0	0		231	1	1		360	2	3
993	5	5		685	4	4		505	3	3
961	5	5		863	5	1		615	3	3
11	0	0		166	0	0		963	5	5
39	0	0		364	2	2		490	2	2
860	5	5		366	2	0		659	3	3
818	4	4		622	3	3		737	4	4
95	0	1		138	0	0		57	0	0
600	3	3		770	4	4		499	2	2
272	1	1		714	4	4		714	4	4
435	2	2		960	5	5		583	3	3
503	3	3		662	3	3		931	5	5

doc#	true class	predicted class		doc#	true class	predicted class		doc#	true class	predicted class
656	3	3		520	3	3		682	4	4
807	4	4		297	1	1		669	4	4
76	0	0		898	5	5		999	5	5
178	1	1		688	4	4		102	0	0
18	0	0		177	1	1		66	0	0
626	3	5		314	1	1		328	1	1
391	2	2		988	5	3		742	4	4
102	0	0		931	5	5		724	4	4
219	1	1		865	5	5		940	5	5
245	1	0		793	4	4		148	0	0
469	2	2		955	5	5		629	3	3
210	1	1		232	1	1		258	1	1
296	1	1		255	1	1		128	0	0
143	0	0		173	1	1		788	4	4
884	5	5		651	3	3		302	1	1
896	5	5		717	4	4		94	0	0
411	2	2		337	2	2		654	3	3
96	0	1		171	1	1		637	3	3
189	1	1		113	0	0		441	2	2
71	0	0		870	5	5		982	5	5
635	3	3		721	4	4		604	3	2
437	2	2		401	2	2		372	2	2
427	2	2		120	0	0		461	2	2
410	2	2		33	0	0		986	5	5
677	4	4		90	0	0		363	2	2
683	4	4		440	2	2		185	1	1
306	1	1		63	0	0		303	1	1
424	2	2		199	1	1		329	1	1
105	0	0		119	0	0		548	3	3
363	2	2		122	0	0		837	5	5
291	1	1		858	5	5		171	1	1
14	0	0		537	3	3		680	4	4
484	2	2		689	4	4		269	1	1
202	1	1		828	4	4		428	2	2
592	3	3		947	5	5		506	3	3
817	4	4		849	5	5		765	4	0
808	4	4		887	5	5		213	1	1
62	0	0		802	4	4		715	4	4
830	4	4		620	3	3		785	4	4
963	5	5		43	0	0		448	2	2
834	5	5		357	2	2		76	0	0
207	1	1		624	3	3		990	5	5
873	5	1		918	5	5		572	3	3
687	4	4		114	0	0		539	3	3
737	4	4		860	5	5		399	2	2
947	5	5		670	4	4		215	1	1

doc#	true class	predicted class		doc#	true class	predicted class		doc#	true class	predicted class
153	0	0		623	3	3		286	1	1
226	1	1		282	1	1		130	0	0
243	1	1		182	1	1		801	4	4
268	1	1		404	2	2		333	1	1
443	2	2		644	3	3		139	0	0
378	2	2		700	4	1		390	2	2
26	0	0		545	3	3		560	3	2
459	2	2		587	3	3		497	2	2
81	0	0		542	3	5		158	0	0
380	2	2		563	3	3		322	1	5
908	5	5		358	2	2		674	4	4
333	1	1		54	0	0		774	4	4
342	2	2		866	5	5		307	1	1
549	3	3		869	5	5		951	5	5
402	2	0		635	3	3		105	0	0
22	0	0		239	1	1		479	2	2
517	3	3		411	2	2		161	0	0
786	4	4		942	5	1		962	5	5
619	3	3		183	1	1		947	5	5
167	0	0		946	5	5		985	5	5
420	2	2		719	4	4		621	3	3
6	0	0		141	0	0		419	2	2
281	1	1		116	0	0		515	3	3
746	4	4		394	2	2		809	4	4
470	2	2		925	5	5		811	4	4
806	4	4		468	2	2		978	5	5
527	3	3		367	2	1		16	0	0
678	4	4		627	3	3		555	3	3
516	3	3		547	3	3		854	5	5
939	5	5		915	5	5		675	4	4
539	3	3		155	0	0		611	3	3
194	1	1		665	3	0		194	1	1

doc#	true class	predicted class		doc#	true class	predicted class
531	3	3		861	5	5
574	3	3		506	3	3
642	3	3		585	3	3
225	1	1		10	0	0
738	4	4		189	1	1
626	3	5		104	0	0
813	4	4		415	2	2
250	1	1		140	0	0
273	1	1		888	5	5
427	2	2		1	0	0
459	2	2		65	0	0
279	1	1		555	3	3

doc#	true class	predicted class		doc#	true class	predicted class
100	0	0		430	2	2
587	3	3		454	2	2
592	3	3		276	1	1
433	2	2		958	5	5
527	3	3		601	3	3
926	5	5		926	5	5
401	2	2		577	3	3
804	4	4		25	0	0
664	3	3		492	2	2
944	5	5		310	1	1
502	3	3		893	5	5
407	2	2		757	4	4
389	2	2		909	5	5
294	1	1		862	5	5
741	4	4		635	3	3
684	4	4		479	2	2
602	3	3		345	2	2
717	4	4		167	0	0
443	2	2		607	3	3
939	5	5		755	4	4
561	3	3		534	3	3
18	0	0		734	4	4
103	0	0		977	5	5
656	3	3		264	1	1
75	0	0		940	5	5
665	3	0		937	5	5
596	3	3		396	2	2
518	3	3		306	1	1
87	0	0		72	0	4
319	1	3		713	4	4
721	4	4		285	1	1
314	1	1		850	5	1
904	5	5		134	0	0
233	1	1		541	3	3
11	0	0		211	1	1
824	4	4		858	5	5
934	5	5		29	0	0
845	5	5		596	3	3
135	0	0		675	4	4
120	0	0		86	0	0
85	0	0		263	1	1
479	2	2		202	1	1
455	2	2		70	0	0
952	5	5		255	1	1
304	1	1		543	3	3
619	3	3		477	2	2

doc#	true class	predicted class		doc#	true class	predicted class
761	4	4		998	5	5
669	4	4		764	4	4
166	0	0		268	1	1
940	5	5		608	3	3
975	5	5		309	1	1
676	4	4		730	4	4
48	0	0		866	5	5
767	4	4		304	1	1
683	4	4		586	3	3
790	4	4		198	1	1
950	5	5		97	0	0
872	5	5		475	2	2
538	3	3		748	4	4
107	0	0		867	5	5
967	5	5		128	0	0
938	5	1		947	5	5
151	0	0		811	4	4
207	1	1		216	1	1
723	4	4		8	0	0
903	5	5		224	1	1
28	0	0		908	5	5
795	4	4		647	3	3
897	5	5		417	2	2
793	4	4		315	1	5
76	0	0		388	2	2
943	5	5		843	5	5
526	3	0		201	1	1
848	5	5		173	1	1
161	0	0		337	2	2
484	2	2		521	3	3
277	1	1		96	0	1
825	4	4		677	4	4
397	2	2		84	0	0
10	0	0		69	0	0
328	1	1		821	4	4
257	1	1		833	5	5
357	2	2		527	3	3
597	3	3		766	4	4
663	3	3		44	0	0
377	2	2		803	4	4
456	2	2		860	5	5