NEWS TOPIC CLASSIFICATION ON TRIBUNNEWS ONLINE MEDIA USING K-NEAREST NEIGHBOR ALGORITHM

Herman

Makassar Center of Human Resources Development and Research on Communication and Informatics
herman@kominfo.go.id

Abstract— **Online media journalists like tribunnews journalists usually determine the news category when make news input. Unfortunately, often the topic submitted is not in accordance with what is expected by the editor. These errors will make it difficult for news searches by customers. To eliminate these errors, editors can be assisted by an application that able to classify topics. Thus, editors is no longer too dependent on journalist input. This study aims to design application that able to classify topics based on the texts contained in the news. The method used is the K-Nearest Neighboor algorithm. This design has produced a system that able to classify news topics automatically. To measure the accuracy of the application, several test were carried out by comparing between its results and the results of manual classification by the editor. The tests those carried out with several scenarios produce an accuracy rate of 82%.**

*Keywords***: Classification; K-Nearest Neighboor; News category.**

I. INTRODUCTION

Tribun Timur Makassar is a daily newspaper that was first published in 2004. This newspaper is a member of the Kompas group. In addition to print media, this media company, through PT. Digital Online Tribune, also manages online media. This online service is named tribunnews. Tribunnews is supported by nearly 500 journalists in 22 important cities in Indonesia. Every reporter has the access to an account that allow them to write the news they get. After a news is input, it will be audited by the editor before being published on the tribunnews.com online news page [1].

In news management of every online media, the news topic classification must always be made. Both reporters and editors of Tribunnews.com also did it. Unfortunately, there often appears to be a less appropriate categorization between news content and the topic of those were done by the reporter. By the amount of news that is very much, the editor may sometimes be negligent, due to not realize the reporter's mistakes, then publish the news that confusing to the reader, because the search results may not display the desired news [2], [3].The establishment of a content text based automated news topic classification is expected to be a solution of these problem [4].

A method that is commonly used in conducting text-based classification is the K-Nearest Neighbor algorithm. Its advantages are easy-use and self-learning. It is able to study the data structure and do its own categorization [4]. It is better in terms of accuracy results compared to the Jst-lvq and C45 algorithms, although it requires a longer computation time than other algorithms [5]. It has a strong consistency, due to it

searches for cases by calculating the similiarity between the current and the previous cases [6]. The statement is reinforced by the results of Adeniyi's (2016) study which shows that the K-Nearest Neighbor method is more transparent, consistent, easy, simple to understand, high tendency to have the desired quality, and easy to implement than most other machine learning techniques, especially when only there is no or only a little prior knowledge about the data distribution [7].

II. METHOD

The design of this system begins with the system architecture as seen in Figure 1. It uses the K-Nearest Neighbor method to determines news topics and headline news classification. The classification is based on previous news database records. There are two actors in it, namely the administrators (journalists and the editor), and the editor in chief. Administrator are users who input the data (news) that has been collected by reporter. The editor in chief is an actor who will receive a classified list of news reports. The system is designed with two parts, namely the front end and back end. The front And is an application that shows the news that has been published, while the back end is an applications that can only be accessed by administrators to do the news input.
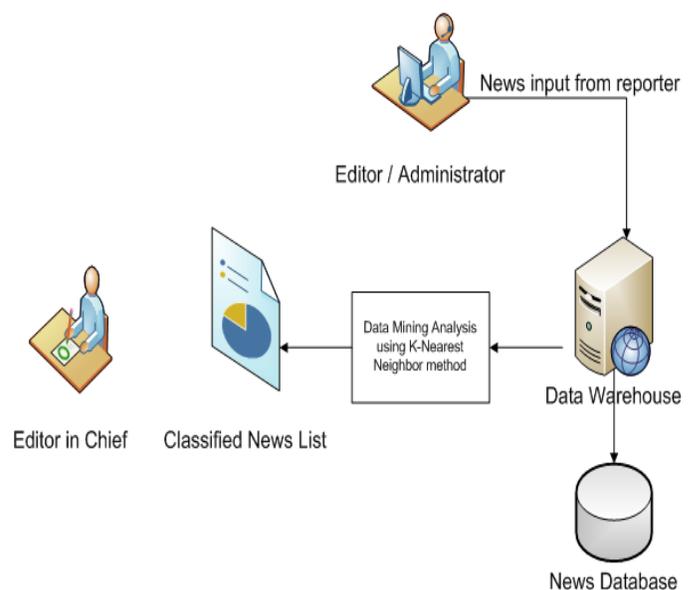

Fig. 1. Architecture of the news classification system using K-Nearest Neighbor algorithm.
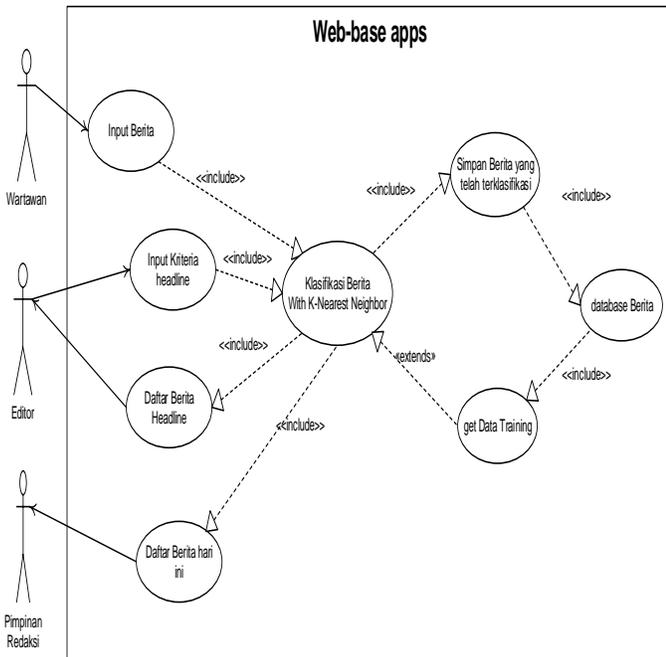
Fig. 2. Use case diagram to verify the news by using the K-Nearest Neighbor algorithm
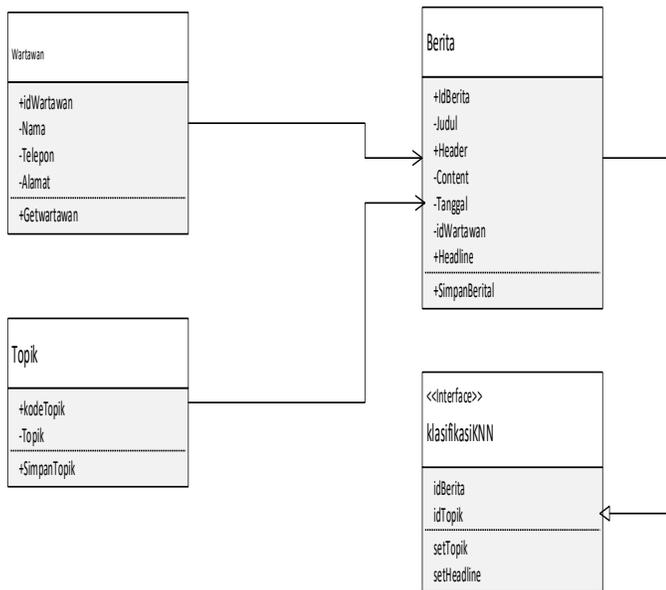

Fig. 3. Class news classification diagram by using the K-Nearest Neighbor algorithm

The user interacts with the system by inputting the news and topic description, running the K-Nearest Neighbor, and publishing the news.

In the training phase, this algorithm only stores feature vectors and classification of sample training data. In the classification phase, the same features are calculated for testing data (the classification is unknown). The distance from this new vector to the entire training sample vector is calculated, and the closest number of K pieces is taken. The new classification point is predicted to be included in the highest classification of these points. To define the distance between two points, namely

the point on the training data (x) and point in the testing data (y), the Euclidean formula is used, as g is shown in the equation:

$$D(x,y) = \sqrt{\sum_{k-1}^{n}(x_k - Y_k)^2} \qquad (1)$$

The accuracy of the KNN algorithm is determined by the presence or absence of irrelevant features, or if the weight of the feature is equivalent to its relevance to classification. K-Nearest Neighbor Algorithm has the advantage of being able to produce strong and clear data, and effective for used on large data. [5]. Beside of these advantages, K-Nearest Neighbor also has several disadvantages, such as: requiring a K value as a parameter; the distance from the experimental data cannot be clear with the type of distance used. To obtain the best results, all attributes or only one definite attributes; and price calculations are very high, because this experiment requires calculating distances from several queries for all experimental data [8]. The algorithm of KNN is shown in Figure 4.
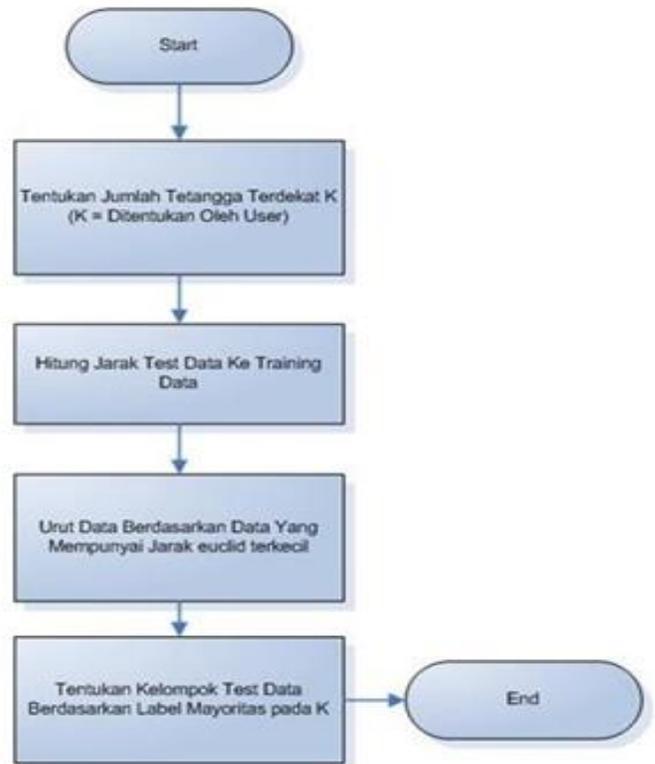

Fig. 4. T the K-Nearest Neighbor method Flowchart

The KNN method calculating steps flowchart starts from determining the number of closest neighbors, then calculating the closest distance, subtracting based on distance and determining the data group .

III. RESULT AND DISCUSSION

Figure 5 and 6 show the intervace view of the application that has been designed, where there are interactions between the user and the system.

Fig. 5. List of News Topics

In the list of news topics, various kinds of news topics are inputted according to media needs, as shown in Figure 5. Each topic is then given the training data as references for the classification. The more is inputted into the training data, the more accurate the results of the classification. [8]

All the news that will be published on tribunnews.com, should be first inputted in a form as Figure 6 shows, then the system will perform the classification process automatically. The classification process dependen on the perviously introduced training data, as shown in figures 8 and 9.

Fig. 6. The News Input Form

There are 4 stages in the classification process of news topics with the K-Nearest Neighboor algorithm, namely:
1. The first step is to select the news that will be used as training data. In this case, 15 training data are given for each category, after that, the data is grouped based on the words that appear most often, then taken the 20 words that appear most often, as shown in table 1. The last line is data that will be calculated using the K-Nearest Neigboard Method to determine the category of the news.
2. Calculating Distance Using Euclidean
The First Training Data Distance with the test data is as follows:

$$D(x,y) = round(sqrt(23-49)^2 + (22-20)^2 + (21-19)^2 + (18-17)^2 + (10-8)^2 + (9-8)^2 + (8-8)^2 + (8-8)^2 + (8-8)^2 + (7-8)^2 + (7-8)^2 + (7-8)^2 + (7-4)^2 + (6-4)^2 + (6-4)^2 + (6-4)^2 + (5-3)^2 + (4-3)^2 + (4-3)^2 + (4-3)^2)) = 26.93$$

So, the Distance of Training Data 1 to 15 Training Data is as follows:

Data Distance with Sample No. 1	26.93
Data Distance with Sample No. 2	29.33
Data Distance with Sample No. 3	22.74
Data Distance with Sample No. 4	44.06
Data Distance with Sample No. 5	49.54
Data Distance with Sample No. 6	48.52
Data Distance with Sample No. 7	55.76
Data Distance with Sample No. 8	47.00
Data Distance with Sample No. 9	51.99
Data Distance with Sample No. 10	55.5
Data Distance with Sample No. 11	51.48
Data Distance with Sample No. 12	55.31
Data Distance with Sample No. 13	56.17
Data Distance with Sample No. 14	37.26
Data Distance with Sample No. 15	51.26

Fig. 7. List of news

K-NEAREST NEIGHBOR

1. K-NEAREST NEIGHBOR

Sampel No.	Kategori	Isi Berita												
1	News	rawa [23]	pening [22]	danau [21]	yang [18]	wisata [10]	menjadi [9]	dunia [8]	juga [8]	iklan [7]	untuk [7]	quot [7]	pquot [7]	bisa [6]
2	News	yang [22]	freeport [14]	karyawan [14]	timika [12]	tidak [9]	sabtu [9]	mereka [8]	petrosea [7]	quot [6]	pquot [6]	massa [6]	papua [6]	indonesia [6]
3	News	yang [32]	bendera [17]	desain [14]	desainer [14]	grafis [13]	indonesia [11]	saya [11]	tidak [11]	tersebut [9]	pada [9]	dari [9]	menjadi [8]	event [8]
4	Bisnis	makassar [11]	tribunnews [7]	title [7]	http [7]	href [7]	text-align [6]	style [6]	pameran [5]	makassarmakassar [4]	quot [4]	anbsp [3]	berlian [3]	mitsubishimitsubishi [3]
5	Bisnis	ekonomi [7]	syariah [6]	pengembangan [5]	dalam [5]	event [4]	keuangan [4]	fesyar [4]	serta [4]	quot [3]	pquot [3]	yang [3]	sulsel [3]	makassar [3]
6	Bisnis	tipe [8]	unit [7]	mamiri [6]	rumah [4]	anging [4]	yang [4]	quot [3]	bank [3]	makassar [3]	dari [3]	residence [3]	sekitar [3]	dengan [3]
7	PSM	steven [4]	latihan [4]	paulle [3]	robert [2]	pamit [2]	gonna [2]	miss [2]	makassar [2]	urus [1]	masih [1]	kitas [1]	kartu [1]	sementara [1]
8	PSM	makassar [9]	quot [6]	pemain [6]	pssi [5]	saya [5]	dengan [5]	sudah [4]	ingin [4]	href [4]	untuk [4]	tribunnews [4]	http [4]	title [4]
9	PSM	tidak [7]	seharusnya [4]	robert [4]	yang [3]	terjadi [3]	pertandingan [2]	dengan [2]	asisten [2]	pemain [2]	bisa [2]	keadilan [2]	karena [2]	fair [2]
10	Super Ball	robert [4]	kuning [3]	wiljan [3]	kartu [2]	komdis [2]	pssi [2]	karena [2]	mendampingi [2]	tidak [2]	rapat [2]	pekan [2]	dari [2]	hari [2]

Fig. 8. Normalization of learning data

2. Hitung Distance menggunakan ecludian distance

Jarak Data Dengan Sample No 1 ==> 26.93
Jarak Data Dengan Sample No 2 ==> 29.33
Jarak Data Dengan Sample No 3 ==> 22.74
Jarak Data Dengan Sample No 4 ==> 44.06
Jarak Data Dengan Sample No 5 ==> 49.54
Jarak Data Dengan Sample No 6 ==> 48.52
Jarak Data Dengan Sample No 7 ==> 55.76
Jarak Data Dengan Sample No 8 ==> 47
Jarak Data Dengan Sample No 9 ==> 51.99
Jarak Data Dengan Sample No 10 ==> 55.5
Jarak Data Dengan Sample No 11 ==> 51.48
Jarak Data Dengan Sample No 12 ==> 55.31
Jarak Data Dengan Sample No 13 ==> 56.17
Jarak Data Dengan Sample No 14 ==> 37.26
Jarak Data Dengan Sample No 15 ==> 51.26

3. Urutkan Data Berdasarkan Jarak [10 Besar Data]

Jarak Data Dengan Sample No 13 ==> 56.17==>Kampus
Jarak Data Dengan Sample No 7 ==> 55.76==>PSM
Jarak Data Dengan Sample No 10 ==> 55.5==>Super Ball
Jarak Data Dengan Sample No 12 ==> 55.31==>Super Ball
Jarak Data Dengan Sample No 9 ==> 51.99==>PSM
Jarak Data Dengan Sample No 11 ==> 51.48==>Super Ball
Jarak Data Dengan Sample No 15 ==> 51.26==>Kampus
Jarak Data Dengan Sample No 5 ==> 49.54==>Bisnis
Jarak Data Dengan Sample No 6 ==> 48.52==>Bisnis
Jarak Data Dengan Sample No 8 ==> 47==>PSM

4. Tentukan klasifikasi menggunakan kategori Mayoritas

Kategori ==> [PSM] ==> 3
Kategori ==> [Super Ball] ==> 3
Kategori ==> [Bisnis] ==> 2
Kategori ==> [Kampus] ==> 2
Maka Kategori Untuk Berita Ini Adalah ==> PSM

Fig. 9. The stages of the process of determining classification by the K-Nearest Neighbor

3. Determining K, in this case, K = 10. After that, the Distances are sequenced based on the smallest distance, os, the data becomes as follows:

Data Distance with Sample No. 3	22.74	News
Data Distance with Sample No. 1	26.9	News
Data Distance with Sample No. 2	29.33	News
Data Distance with Sample No. 14	37.26	Kampus
Data Distance with Sample No. 4	44.06	Bisnis
Data Distance with Sample No. 8	47	PSM
Data Distance with Sample No. 6	48.52	Bisnis
Data Distance with Sample No. 5	49.54	Bisnis
Data Distance with Sample No. 15	51.26	Kampus
Data Distance with Sample No.11	51.48	Super Ball

4. Determining Classification Based on Majority. Since the calculation of the category was as follows:

Category	[News]	3
Category	[Bisnis]	3
Category	[Kampus]	2
Category	[PSM]	1
Category	[Super Ball]	1

Then, it can be determined that the tested data is included in the News NEWS category

The classification system that uses K-Nearest Neighbors has successfully classified news with an accuracy rate up to 82 %, as shown in Table 1. The tests were conducted by comparing between the results of the system and the editor's manual classification.

TABLE 1
Accuracy Test Results

Trial	Training Data	Test Data	Accuracy
I	15	15	67%
II	30	15	71%
III	45	15	78%
IV	60	15	82%
V	75	15	82%

The test results show that large training data is needed for improving accuracy to achieve the saturation point (a condition when the treatment of adding training data to subsequent experiments is no longer able to provide increased accuracy). This result also indicate the similarity of the level of accuracy to the research conducted by Andreas Daniel and Suwanto. Daniel has conducted clustering for the text category with K-Nearest Neighbor and resulted in 85% accuracy [9], while Suwanto conducted document grouping using winowing fingerprint method with K-Nearest Neighbor and produces 80% grouping accuracy value to 10 test documents [10].

IV. CONCLUSION

This research has established a system that able to classify topics automatically based on the news input. The results and the testing of the design shows that that the K-Nearest Neighbor algorithm can do news classification according to the topic categories those have been previously determined with an accuracy rate up to 82%.

The tribunnews that has been running for several years certainly has a lot of digital documents. Those data can be used as the data mining of training data. The more of the trainng data the more the accuracy will increase.

For further development, two or more algorithms might be combined as a method to improve accuracy in classifying topics.

V. ACKNOWLEDGMENT

VI. REFERENCES

[1] "tribunnews.com," 2018. [Online]. Available: http://www.tribunnews.com/. [Accessed: 25-Oct-2018].

[2] S. F. E. P. Dimas Bagus Prasetyo, Freddy Arviando, Muhammad Farhan Mubarak, "Klasifikasi berita berdasarkan pendekatan semantik," in *Prosiding Seminar Ilmiah Nasional Komputer dan sistem Intelejen (KOMMIT)*, 2014, vol. 8, no. Kommit, pp. 201–206.

[3] E. Junianto and D. Riana, "Penerapan PSO Untuk Seleksi Fitur Pada Klasifikasi Dokumen Berita Menggunakan NBC," vol. 4, no. 1, pp. 38–45, 2017.

[4] A. Rachmat, "Implementasi Metode K-Nearest Neighbor dengan Decision Rule untuk Klasifikasi Subtopik Berita," *J. Inform.*, vol. 10, no. Juni, pp. 1–15, 2014.

[5] R. W. Muhammad Fakhurrifqi, "Perbandingan Algoritma Nearest Neighbour, C4.5 dan LVQ untuk Klasifikasi Kemampuan Mahasiswa," *Ijccs*, vol. 7, no. July, pp. 145–154, 2013.

[6] E. T. L. Kusrini, *Algoritma Data Mining*. Yogyakarta: Andi Offset, 2009.

[7] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Appl. Comput. Informatics*, vol. 12, no. 1, pp. 90–108, 2016.

[8] S. K. Lidya, O. S. Sitompul, and S. Efendi, "Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (Svm)," *Semin. Nas. Teknol. dan Komun. 2015*, vol. 2015, no. Sentika, pp. 1–8, 2015.

[9] A. D. Arifin, "Implementasi Algoritma K-Nearest Neighbour Yang Berdasarkan One Pass Clustering Untuk Kategorisasi Teks, Implementation of K-Nearest Neighbour Algorithm Based on One Pass Clustering Algorithm for Text Categorization," *Pap. Present. Informatics Eng. RSIf 518.1 Ari i, 2012*, pp. 1–7, 2012.

[10] S. Sanjaya and E. A. Absar, "Pengelompokan Dokumen Menggunakan Winnowing Fingerprint dengan Metode K - Nearest Neighbour," *J. CoreIT*, vol. 1, no. 2, pp. 50–56, 2015.