APPLICATION OF DATA MINING TO CLASSIFY HATE SPEECH ON SOCIAL MEDIA BY USING THE K NEAREST NEIGHBOR ALGORITHM

Windania Purba[1], Fando Marehitno Salim [2], Antoni[3], Yuni Suhendrik[4], Jeanie Winata[5]
Fakultas Teknologi dan Ilmu Komputer, Universitas Prima Indoneisia

E-mail: Winda.nia04@gmail.com, fandosalim@yahoo.com, antoni.duaa@gmail.com, yunsunn05@gmail.com, jeaniewinata97@gmail.com

Abstract: Social media is one of the biggest sources of information that we can get right now. However, in the use and dissemination of information, there are still many social media users who spread information or hateful words (Hate Speech). Therefore classification needs to be done to reduce the appearance of hate speech sentences with K Nearest Neighbor. K Nearest Neighbor Algorithm classifies based on the results of learning on the object being carried out. In the research carried out the KNN algorithm succeeded in classifying the Hate Speech on the given tweet data.

Keywords: *K neares neighbor, Classification, Data Mining, Hate Speech, Social Media*

Abstrak: Sosial media fungsinya pada saat ini merupakan salah satu sumber informasi terbesar yang dapat kita dapatkan. Namun, dalam penggunaan dan penyebaran informasinya, masih banyak pengguna sosial media yang menyebarkan informasi atau kata-kata berbau kebencian (Hate Speech). Untuk itu klasifikasi perlu dilakukan untuk mengurangi munculnya kalimat berbau hate speech dengan K Nearest Neighbor. Algoritma K Nearest Neighbor mengklasifikasikan berdasarkan hasil dari pembelajaran terhadap objek yang dilakukan. Pada penelitian yang dilakukan algoritma KNN berhasil melakukan klasifikasi Hate Speech pada data tweet yang diberikan.

Kata Kunci: *K Nearest Neighbor, Klasifikasi, Data Mining, Ujaran Kebencian, Sosial Media*

INTRODUCTION

In the current global era, information is very easy to obtain. Social media is one of the latest sources of information because it can be accessed from anywhere as long as you have an internet network. Different from conventional information sources, such as radio and newspapers. Social media in the delivery of information provides an opportunity for recipients of information to provide opinions regarding the information provided. Although information pertaining to social media has a big positive impact on society. However, in the use and dissemination of information, there are still many social users who disseminate information or hateful words (Hate Speech) whether they smell racially, or because of differences of opinion. Hate Speech can be the root of an unfavorable atmosphere. To maintain a conducive atmosphere in carrying out activities in using social media, of course things that smell of Hate Speech must be prevented from appearing in information received on social media. To prevent the emergence of hate speech on social media, of course it needs to be classified in advance what information includes hate speeches.

p-ISSN: 2355 - 1739
e-ISSN: 2407 - 6295

Classification with K Nearest Neighbor works based on the learning done where the object with the distance closest to the learning result. In general, the K Nearest Neighbor algorithm classifies based on the results of learning on the object carried out. In the study entitled "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Indonesian Giving to The Poor Determining the Classification of Healthy Cards", said that "K Nearest Neighbor has a higher level of accuracy than the Naive Bayes classifier model" (Yofi, 2018). The advantage of usage is that this method only requires a small amount of training data and has a fast time in processing.

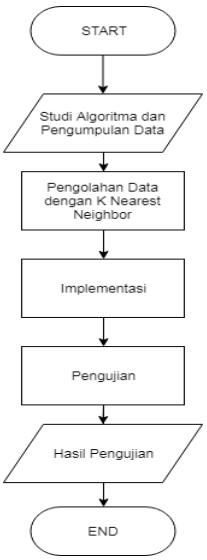
Previously in the study using the Naïve Bayes method in the study entitled "Hate Speech Classification on Twitter Using Nïve Bayes Based on Gram with the Information Gain Feature Selection" the percentage of Naïve Bayes classification was 84% (M.Hakiem, 2019).

Then in the study entitled "Analysis of Sentemien Hate Speech on Twitter with Naïve Bayes Classifier Method and Support Vector Machine" by Ghulam Asrofi Buntoro, 2016. The results of Naïve Bayes Classification are 75.8%.

Based on the description above, the author wishes to compile a thesis entitled "**Application of Data Mining for Classification of Hate Speech on Social Media by Using K Nearest Neighbor Algorithm**".

METHOD

In this study the method used is the K Nearest Neighbor Method. The following are the research methods that we will use in research.



Picture 1. Plot of Research Methods

The research methodology carried out in this study are as follows:

1. Study of Algorithms and data collection, where at this stage learning will be carried out on algorithms, and conducting data collection related to research
2. Manually processing data, then processing data with the help of rapidminer.
3. Then do the testing, and draw conclusions from the results of the tests performed.

K-Nearest Neighbor Algorithm is a classification method that classifies new data based on the distance of the new data to some of the closest data / neighbors (5). K-Nearest Neighbor technique by doing steps namely [5], starting input: Training data, data traning label, k, data testing

a. For all testing data, calculate the distance to each training data

p-ISSN: 2355 - 1739
e-ISSN: 2407 - 6295

b. Determine the training data k which is the closest to the data
c. Testing
d. Check the label of this data
e. Determine the label with the most frequency
f. Enter testing data to the class with the most frequency
g. Stop Labels for all testing data obtained.

To calculate the distance between two x and y points, you can use the Euclidean distance as follows:

$$d\ (X_1, Y_2) = \sum \left| \frac{n_{1i}}{n1} - \frac{n_{2i}}{n2} \right|$$

Which X1, l = 1, 2, are category attributes, and n1j, n1 represent the appropriate frequency. K-Nearest algorithm Neighbor is a supervised learning algorithm where the results of newly classified instances are based on the majority of the closest k-neighbor categories.

The purpose of this algorithm is to classify new objects based on attributes and samples from training data. The k-Nearest Neighbor algorithm uses Neighborhood Classification as the predictive value of the value of the new instance.

The data that has been collected and compiled in the excel file will then be determined by the weight of each word after manual stemming and stopword removal. Examples are as follows:

Result

For example, from the data that has been collected and arranged in the table, the data is as follows:

Table 1. Example Data

Tanggal	Isi Pesan / Berita	Kelas
25/05/2019	liga 1 musim ini banyak pemain asing berkualitas... bakalan seru nih persaingan liga 1	Positif
25/05/2019	2 menit yg merubah pertandingan, Hari Nur dari musim lalu bagus dan layak masuk timnas @psisfcofficial	Positif
25/05/2019	Hahaha di bulan Ramadhan masih aja cari dosa pak2 kasihan keluarga anda makan uang haram.	Negatif
25/05/2019	wasit *n**ng,mafia masih bermain t*ik!	Negatif
25/05/2019	Wasit bodo.	Negatif
25/05/2019	Tnp mafia wasit kalian lemah dan kalah...	Negatif

Then from the example data above, a stopword removal is done where the word "link" will be removed manually. After that the words that have been obtained from deletion of the conjunctions will be removed. So that in the example data above, the available words for given weights are as follows:

Table 2. Giving weight to the word

Kata	D1	D2	D3	D4	D5	D6	Total	Bobot Log (jumlah data/t

p-ISSN: 2355 - 1739
e-ISSN: 2407 - 6295

								otal)
liga	1	0	0	0	0	0	1	0,845
musim	1	1	0	0	0	0	2	0,544
banyak	1	0	0	0	0	0	1	0,845
pemain	1	0	0	0	0	0	1	0,845
asing	1	0	0	0	0	0	1	0,845
kualitas	1	0	0	0	0	0	1	0,845
bakal	1	0	0	0	0	0	1	0,845
seru	1	0	0	0	0	0	1	0,845
saing	1	0	0	0	0	0	1	0,845
menit	0	1	0	0	0	0	1	0,845
ubah	0	1	0	0	0	0	1	0,845
tanding	0	1	0	0	0	0	1	0,845
bagus	0	1	0	0	0	0	1	0,845
layak	0	1	0	0	0	0	1	0,845
masuk	0	1	0	0	0	0	1	0,845
timnas	0	1	0	0	0	0	1	0,845
bulan	0	0	1	0	0	0	1	0,845
Ramadhan	0	0	1	0	0	0	1	0,845
dosa	0	0	1	0	0	0	1	0,845
kasihan	0	0	1	0	0	0	1	0,845
keluarga	0	0	1	0	0	0	1	0,845
makan	0	0	1	0	0	0	1	0,845
duit	0	0	1	0	0	0	1	0,845
haram	0	0	1	0	0	0	1	0,845
wasit	0	0	0	1	1	1	3	0,368
a*****	0	0	0	1	0	0	1	0,845
mafia	0	0	0	1	0	1	2	0,544
main	0	0	0	1	0	0	1	0,845
t**	0	0	0	1	0	0	1	0,845
bodoh	0	0	0	0	1	0	1	0,845
lemah	0	0	0	0	0	1	1	0,845
kalah	0	0	0	0	0	1	1	0,845

Then each weight is put back into the dataset, where in the table below each weight in the word is then put back into the dataset.

Table 3. Data Collection with words that have been given weights

Kata	D1	D2	D3	D4	D5	D6
liga	0,845	0,000	0,000	0,000	0,000	0,000
musim	0,845	0,845	0,000	0,000	0,000	0,000
banyak	0,845	0,000	0,000	0,000	0,000	0,000
pemain	0,845	0,000	0,000	0,000	0,000	0,000
asing	0,845	0,000	0,000	0,000	0,000	0,000
kualitas	0,845	0,000	0,000	0,000	0,000	0,000
bakal	0,845	0,000	0,000	0,000	0,000	0,000
seru	0,845	0,000	0,000	0,000	0,000	0,000
saing	0,845	0,000	0,000	0,000	0,000	0,000
menit	0,000	0,845	0,000	0,000	0,000	0,000
ubah	0,000	0,845	0,000	0,000	0,000	0,000
tanding	0,000	0,845	0,000	0,000	0,000	0,000
bagus	0,000	0,845	0,000	0,000	0,000	0,000
layak	0,000	0,845	0,000	0,000	0,000	0,000
masuk	0,000	0,845	0,000	0,000	0,000	0,000
timnas	0,000	0,845	0,000	0,000	0,000	0,000
bulan	0,000	0,000	0,845	0,000	0,000	0,000
Ramadhan	0,000	0,000	0,845	0,000	0,000	0,000
dosa	0,000	0,000	0,845	0,000	0,000	0,000
kasihan	0,000	0,000	0,845	0,000	0,000	0,000
keluarga	0,000	0,000	0,845	0,000	0,000	0,000
makan	0,0	0,0	0,8	0,0	0,0	0,0

p-ISSN: 2355 - 1739
e-ISSN: 2407 - 6295

	00	00	45	00	00	00
duit	0,000	0,000	0,845	0,000	0,000	0,000
haram	0,000	0,000	0,845	0,000	0,000	0,000
wasit	0,000	0,000	0,000	0,845	0,845	0,845
a*****	0,000	0,000	0,000	0,845	0,000	0,000
mafia	0,000	0,000	0,000	0,845	0,000	0,845
main	0,000	0,000	0,000	0,845	0,000	0,000
t**	0,000	0,000	0,000	0,845	0,000	0,000
bodoh	0,000	0,000	0,000	0,000	0,845	0,000
lemah	0,000	0,000	0,000	0,000	0,000	0,845
kalah	0,000	0,000	0,000	0,000	0,000	0,845

Next in the sample data we enter a new word, namely: "How much is the referee paid?" Which consists of the words Pay, How much, Referee. Then the table is obtained as follows:

Table 4. The weight value of the new document (D0) and Test document

Kata	D0	D1	D2	D3	D4	D5	D6
Bayar	0	0	0	0	0	0	0
Berapa	0	0	0	0	0	0	0
Wasit	0,368	0	0	0	0,368	0,368	0,368

Then look for similarities where using the formula:

$$D_i = \frac{X}{\sqrt{Y^2} - \sqrt{Z^2}}$$

Information :

X = is the weight of the document to i during the test

Y = is the weight value of the document being tested

Z = is the weight value of the document when giving weight

From the weighting table above, the results of classifying a new document can be taken with the content "Paid how many referees" are negative sentences based on test data that has been tested and given its weight. This is because of all the test data there are 3 out of 4 negative test data that have weight values. From the search results, the similarity of the 5th data is data that has the value most similar to the test data, so that the test data can be ascertained as Hate Speech based on the dataset we have.

Table 5. Result of Clarification

Isi Pesan / Berita	Kelas	Bobot
liga 1 musim ini banyak pemain asing berkualitas... bakalan seru nih persaingan liga 1	Positif	0
2 menit yg merubah pertandingan, Hari Nur dari musim lalu bagus dan layak masuk timnas @psisfcofficial	Positif	0
Hahaha di bulan Ramadhan masih aja cari dosa pak2 kasihan keluarga anda makan uang haram.	Negatif	0
wasit *n**ng,mafia masih bermain t*ik!	Negatif	0,237
Wasit bodo.	Negatif	0,592
Tnp mafia wasit kalian lemah dan kalah...	Negatif	0,296

p-ISSN: 2355 - 1739
e-ISSN: 2407 - 6295

CONCLUSION

After completing research on the application of the Nearest Neighbor K method to Data Mining in classifying the Hate Speech, the drafting team can draw the following conclusions:

1. K Nearest Neighbor can classify sentences or words that contain utterances of hatred.
2. There is a Tweet that is done by Classification and Data Mining processing, there are still many tweets that smell of hatred.

BIBLIOGRAPHY

Aditya Kresna, dkk. 2018. "Identifikasi Ujaran Kebencian Pada Facebook Dengan Metode Ensemble Feature Dan Support Vector Machine". Universitas Brawijaya.

Ghulam. 2016. "Analisis Sentimen Hatespeech Pada Twitter Dengan MetodeNaïve Bayes Classifier Dan Support Vector Machine". Universitas Muhammadiyah Ponorogo.

Hakiem.M , Ali Fauzi, Moh. 2019. "Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur Information Gain". Universitas Brawijaya.

Hastuti, K. (2012, Juni). Analisis Komparasi Algoritma Klasifikasi Data Mining V. Seminar Nasional Teknologi Informasi & Komunikasi Terapan(979 - 26 - 0255 - 0), 241- 249.

Ian H. Witten, f. E. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3 ed.). (A. S. Burlington, Ed.) United States of America: Morgan Kaufmann.

Kamber, H. &. (2006). Data Mining Concept and Tehniques. San Fransisco: Morgan Kauffman.

M. Choirul Anam dan MuhammadHafiz, SE . 2015. Kapolri tentang Penanganan Ujaran Kebencian (Hate Speech) dalam Kerangka Hak Asasi Manusia.

Meri Febriani. 2018. "Analisis Faktor Penyebab Pelaku Melakukan Ujaran Kebencian (Hate Speech) Dalam Media Sosial". Universitas Lampung, Bandar Lampung.

Santoso, B. (2007). Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis (1 ed.). Yogyakarta: Graha Ilmu.

Tahyudin, I. (2013, December). Comparing Clasification Algorithm Of Data Mining to Predict the Graduation Students on Time. Information Systems International Conference(ISICO).

p-ISSN: 2355 - 1739
e-ISSN: 2407 - 6295