Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 897-910
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Analisis Sentimen Media Sosial Twitter Dengan Algoritma K-Nearest Neighbor Dan Seleksi Fitur Chi-Square (Kasus Omnibus Law Cipta Kerja)

Aditiya Yoga Pratama[1], Yuyun Umaidah[2], Apriade Voutama[3]
[1,2,3]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
Ds. Paseurjaya, Kec. Telukjambe Timur, Kab Karawang, Prof. Jawa Barat
aditiya.yoga17017@student.unsika.ac.id[1], yuyun.umaidah@staff.unsika.ac.id[2] ,
apriade.voutama@staff.unsika.ac.id[3]

Abstract

The use of information technology is growing rapidly, marked by public opinion that can be conveyed indefinitely through social media. One of the social media that is Twitter. Twitter is considered easier to retrieve information related to existing opinions and sentiments due to the limited character in a tweet made by users and there is the hashtag "#" which can be searched easily regarding the current hotly discussed situation. Some time ago, there was a lot of discussion regarding the ratification of the omnibus work copyright law. Tweets in the form of lively sentiments adorn the hashtag "#omnibuslaw" and other related hashtags. This study discusses reviews in the form of tweets related to omnibus law with the Chi Square selection feature and the K-Nearest Neighbor algorithm using R Studio. Data were taken as many as 500 tweets related to the omnibus law. The methodology used is Knowledge Discovery in Databases. Data labeling is carried out by experts who are divided into positive and negative sentiments. The results of the modeling using K-Fold Cross Validation, the highest accuracy is obtained with a 25% feature use scheme (Chi Square feature selection), and the value of k = 5 in KNN is 81.4%. Testing on the model was carried out using 100 random data and obtained 83% accuracy, 100% precision, 15% recall and 26.08% F-Measure value. Of the 500 data taken, the word "people" is the most dominating word. Of the 500 data taken, 78.8% negative sentiments and 21.2% positive sentiments.

Keywords: *Chi-Square, K-Nearest Neighbor, Sentiment Analysis, Text Mining.*

Abstrak

Penggunaan teknologi informasi berkembang dengan pesatnya ditandai dengan opini masyarakat yang dapat disampaikan tanpa terbatas waktu melalui media sosial. Media sosial yang umum digunakan dalam penyampaian opini serta sentimen salah satunya twitter. Twitter dianggap lebih mudah dalam diambil informasinya terkait opini serta sentimen yang ada dikarenakan keterbatasan karakter dalam suatu tweet yang dilakukan oleh pengguna serta terdapat hastag "#" yang dapat dilakukan pencarian dengan mudah terkait keadaan yang sedang hangat diperbincangkan. Beberapa waktu lalu ramai diperbincangkan terkait pengesahan omnibus law cipta kerja. Tweet berupa sentimen ramai menghiasi tagar "#omnibuslaw" serta tagar lain yang berkaitan. Tweet berupa sentimen yang ada dilakukan pengambilan informasi dengan analisis sentimen. Penelitian ini membahas ulasan berupa tweet terkait omnibus law dengan fitur seleksi Chi Square dan algoritma K-Nearest Neighbor dengan menggunakan R Studio. Diambil data sebanyak 500 tweet berkaitan dengan omnibus law. Metodologi yang digunakan adalah Knowledge Discovery in Databases yaitu dengan Data Selection, Pre-processing Data, Transformation, Data Mining, Evaluation. Pelabelan data dilakukan oleh pakar yang terbagi menjadi sentimen positif dan negatif. Hasil permodelan dengan menggunakan K-Fold Cross Validation, akurasi tertinggi diperoleh dengan skema 25% penggunaan fitur (seleksi fitur Chi Square), dan nilai k = 5 pada KNN yaitu sebesar 81,4%. Pengujian terhadap model dilakukan dengan menggunakan data acak sebanyak 100 data dan diperoleh akurasi sebesar 83%, precission sebesar 100%, recall sebesar 15% dan nilai F-

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 897-910
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Measure sebesar 26,08%. Dari 500 data yang diambil, kata "rakyat" merupakan kata yang paling mendominasi. Dari 500 data yang diambil sebanyak 78,8% merupakan sentimen negatif dan 21,2% sentimen positif.

***Kata kunci**: Analisis Sentimen Twitter,* Chi-Square, *K*-Nearest Neighbor*, Teks Mining*

1. PENDAHULUAN

Banyaknya penggunaan media sosial saat ini, menandakan bahwa sosial media saat ini bukan lagi sebagai tempat untuk mencari pertemanan, namun juga bisa dijadikan sebagai tempat menyampaikan aspirasi. Sosial media adalah satu set baru komunikasi dan alat kolaborasi yang memungkinkan banyak jenis interaksi yang sebelumnya tidak tersedia untuk orang biasa (1). Beberapa waktu lalu, terjadi perbincangan publik mengenai pengesahan omnibus law. Kebijakan pemerintah mengenai mencetusnya omnibus law cipta kerja tersebut menimbulkan banyak hal pro dan kontra pada kebijakan tersebut. beriringan dengan disahkannya kebijakan tersebut, masyarakat saat ini mulai memberikan komentarnya baik itu secara demonstrasi langsung dan menggunakan media sosial sebagai media penyampaian aspirasinya terutama twitter.

Twitter digunakan karena penyampaian tweet dari twitter kata-katanya terbatas hanya beberapa karakter sehingga seseorang yang menyampaikan tweet dapat secara langsung mengarah ke topik yang ingin disampaikan (2). Namun data mentah dari twitter tersebut harus dilakukan sebuah analisa yaitu analisa sentimen. Analisa sentimen digunakan agar mendapatkan informasi mengenai pemahaman publik terhadap suatu hal yang disampaikan secara subjektif (3). Tweet yang dilakukan bisa saja merupakan tanggapan dari sebuah kebijakan yang dilakukan oleh pemerintah/instansi tertentu.

Pada penelitian sebelumnya, dilakukan sebuah perbandingan antara algoritma K-NN dan SVM dalam melakukan analisis sentimen twitter. Digunakan variabel berupa akurasi serta kecepatan pemrosesan. Hasil yang didapatkan adalah metode K-NN lebih baik dalam waktu pemrosesan namun lebih buruk secara akurasi (4). Dalam penelitian lain dilakukan analisis sentimen dengan menggunakan algoritma SVM dan dengan fitur seleksi Chi Square didapatkan hasil bahwa fitur seleksi Chi Square dapat membuat performa dan meningkatkan akurasi dalam melakukan pengklasifikasiannya(5) . Maka dari itu akan digunakan algoritma KNN yang memiliki kecepatan pemrosesan yang baik, namun dengan akurasi yang kurang baik dengan dikombinasikan terhadap fitur seleksi Chi Square untuk meningkatkan akurasi pada algoritma KNN.

Sentimen yang ada di twitter mengenai kebijakan pemerintah berkenaan dengan kebijakan omnibus law cipta kerja akan dilakukan sebuah analisa agar mengetahui berapa persentase masyarakat yang memiliki argumen positif maupun negatif mengenai kebijakan yang dilakukan. Untuk itu pada penelitian kali ini digunakanlah sebuah metode KNN dengan

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 897-910
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

selection feature Chi-Square guna memaksimalkan melakukan analisis pada sentimen twitter tersebut.

2. METODOLOGI PENELITIAN

2.1. *Data Mining*

Data mining merupakan pengolahan suatu pengetahuan yang informatif dan menarik yang didasarkan pada pola pada sebuah data besar (6). Data-data yang digunakan bisa saja diperoleh dari database, warehouse data, web, dan lain sebagainya yang dapat diolah menjadi sebuah informasi. Data mining dapat digunakan dalam berbagai sektor dan bertujuan untuk berbagai hal yaitu meningkatkan pengetahuan, untuk beberapa sektor bisa digunakan dalam meningkatkan penjualan, dan lain sebagainya (7). Salah satu bagian dari data mining adalah klasifikasi. Klasifikasi adalah proses menemukan model melalui analisis terhadap sekumpulan data pelatihan yang menggambarkan dan membedakan kelas label atau konsep data (6).

2.2. *Text Mining*

Text mining merupakan proses pengelolaan data berupa teks yang dijadikan pola untuk mendapatkan sebuah informasi penting (8). Dalam kata lain, text mining ini seperti data mining, hanya saja data yang diolah oleh text mining adalah sebuah data berupa tekstual. Sebelum melakukan text mining, terdapat tahap text preprocessing. Tahapan dari text preprocessing yaitu *case folding*, *cleansing*, *tokenizing*, *stemming*, dan *stopword removal* (9).

2.3. *Chi Square*

Fitur-fitur yang kurang relevan terhadap proses pengklasifikasian sebaiknya ditanggulangi oleh seleksi fitur. Seleksi fitur Chi Square menggunakan teori statistika untuk menguji independensi sebuah term dengan kategori atau labelnya. Seleksi fitur Chi Square berdasar pada teori statistika yang didasarkan pada dua peristiwa diantaranya adalah kemunculan dari fitur dan kemunculan dari kategorinya yang didasarkan pada perhitungan persamaan (10):

$$x^2\,(D,t,c) = \sum_{et\in\{1,0\}} \sum_{ec\in\{1,0\}} \frac{(N_{et\ el} - E_{et\ el})}{E_{et\ el}} \qquad (1)$$

2.4. K-*Nearest Neighbor*

Algoritma K-*Nearest Neighbor* (KNN) merupakan salah satu algoritma yang sering digunakan dalam pengklasifikasian pada *machine learning*. Tujuan dari algoritma KNN adalah mengklasifikasikan objek ke dalam salah satu kelas yang telah ada dalam data sampel yang sebelumnya telah ditetapkan (11). Metode KNN mengelompokkan data ke dalam suatu kelas yang telah ditetapkan berdasarkan jarak terdekat atau kemiripan terhadap data set atau data latih yang ada (12). Tahapan dalam proses K-NN ialah (13):

a) Menentukan nilai K.

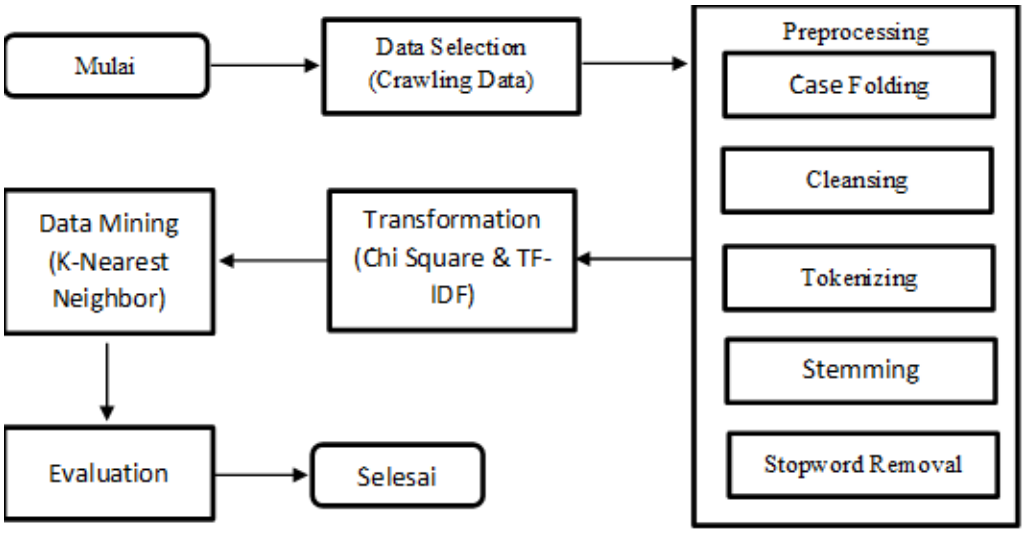
b) Menghitung jarak antara data yang akan di klasifikasikan terhadap data label.
c) Menentukan nilai k yang paling kecil.
d) Klasifikasikan data dengan berdasarkan kepada metrik jarak.

Dalam menghitung kedekatan dengan menggunakan metrik jarak, Kedekata bisa dihitung dengan jarak Eucklidean dengan rumus pada persamaan (14):

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2} \quad (2)$$

2.5. Metodologi Penelitian

Pada penelitian kali ini, tahapan dalam penyelesaian penelitian menggunakan tahapan KDD (*Knowledge Discovery in Database*) :



Gambar 1. Alur penelitian

KDD adalah metode dan cara untuk mendapatkan sebuah informasi melalui basis data yang telah tersedia (15). Dalam proses penyelesaiannya, tahap penyelesaiannya adalah (16):

a) Data Selection
Data selection adalah tahap pengambilan dan pemilihan sample data yang ingin diolah untuk dijadikan sebuah informasi penting.
b) Data Preprocessing
Data preprocessing adalah tahap pembersihan data dari data yang mengganggu (noise) dan data yang tidak konsisten.
c) Transformation
Tahap transformasi merupakan proses penyesuaian data. Data yang tadinya sudah dilakukan pembersihan akan di proyeksikan dan

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 897-910
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

disesuaikan terhadap metode atau algoritma yang akan digunakan dalam data mining.

d) Data Mining
Proses data mining merupakan proses penggalian data dan pola dari data yang telah diproses pada tahapan sebelumnya untuk menjadi sebuah informasi dengan metode dan algoritma tertentu.

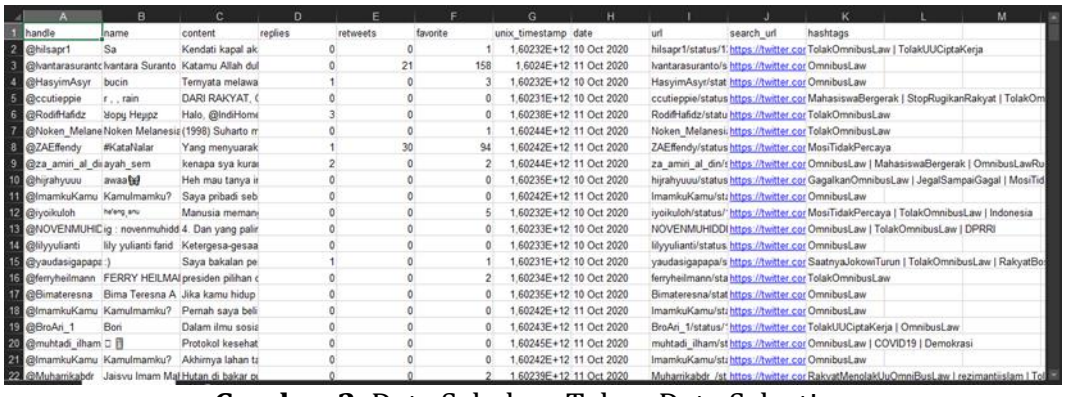
e) Evaluation
Tahap evaluasi merupakan tahap merepresentasikan pengetahuan dari proses sebelumnya. Selain itu dapat juga dijadikan sebagai bahan untuk pertimbangan pada penelitian selanjutnya.

3. HASIL DAN PEMBAHASAN

Hasil penelitian yang dilakukan adalah melakukan analisis sentimen dengan mengklasifikasikan ulasan dari twitter terkait dengan #omnibuslaw dengan menggunakan seleksi fitur Chi-Square dan algoritma K-Nearest Neighbor.

3.1. Data Selection

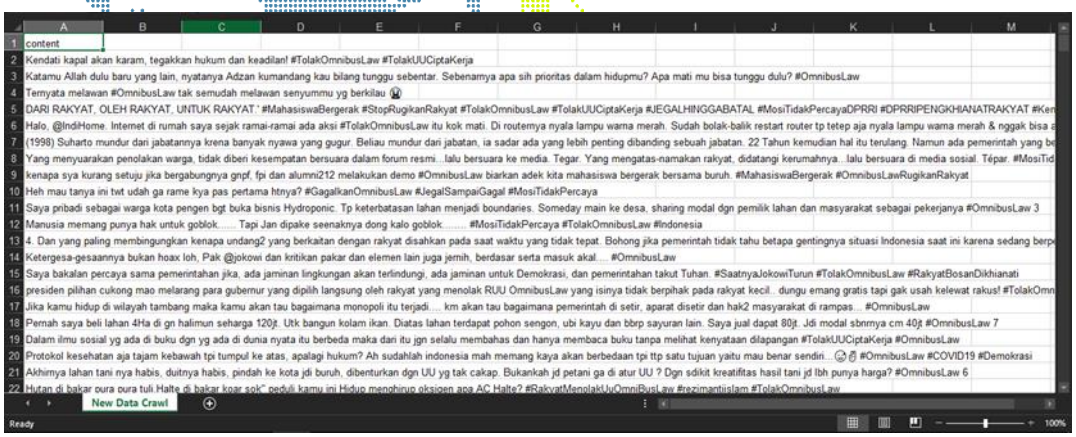
Pada tahapan ini data yang telah diperoleh melalui scrapping dari data twitter akan dilakukan proses data selection, yaitu akan dipilih atribut-atribut yang akan digunakan dalam proses pembersihan data, dan klasifikasi nantinya. Sebelumnya data yang telah diperoleh memiliki banyak atribut diantaranya handle, name, content, replies, retweets, favorite, unix_timestamp, url, search url, dan hastag. Dalam data selection ini, dipilih hanya content yang akan dijadikan sebagai data pada tahapan berikutnya.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	handle	name	content	replies	retweets	favorite	unix_timestamp	date	url	search_url	hashtags		
2	@hilsapr1	Sa	Kendati kapal ak	0	0	1	1,60232E+12	10 Oct 2020	hilsapr1/status/1	https://twitter.cor	TolakOmnibusLaw | TolakUUCiptaKerja		
3	@Ivantarasurantc	Ivantara Surante	Katamu Allah dul	0	21	158	1,6024E+12	11 Oct 2020	Ivantarasuranto/s	https://twitter.cor	OmnibusLaw		
4	@HasyimAsyr	bucin	Ternyata melawa	1	0	3	1,60232E+12	10 Oct 2020	HasyimAsyr/stat	https://twitter.cor	OmnibusLaw		
5	@ccutieppie	r , , rain	DARI RAKYAT, (	0	0	0	1,60231E+12	10 Oct 2020	ccutieppie/status	https://twitter.cor	MahasiswaBergerak | StopRugikanRakyat | TolakOm		
6	@RodifHafidz	dopy Heppz	Halo, @IndiHome	3	0	0	1,60238E+12	11 Oct 2020	RodifHafidz/statu	https://twitter.cor	TolakOmnibusLaw		
7	@Noken_Melane	Noken Melanesia	(1998) Suharto m	0	0	1	1,60244E+12	11 Oct 2020	Noken_Melanesi	https://twitter.cor	TolakOmnibusLaw		
8	@ZAEffendy	#KataNalar	Yang menyuarak	1	30	94	1,60242E+12	11 Oct 2020	ZAEffendy/status	https://twitter.cor	MosiTidakPercaya		
9	@za_amiri_al_di	ayah_sem	kenapa sya kura	2	0	2	1,60244E+12	11 Oct 2020	za_amiri_al_din/s	https://twitter.cor	OmnibusLaw | MahasiswaBergerak | OmnibusLawRu		
10	@hijrahyuuu	awaa	Heh mau tanya i	0	0	0	1,60235E+12	10 Oct 2020	hijrahyuuu/status	https://twitter.cor	GagalkanOmnibusLaw | JegalSampaiGagal | MosiTid		
11	@ImamkuKamu	Kamulmamku?	Saya pribadi seb	0	0	0	1,60242E+12	11 Oct 2020	ImamkuKamu/st	https://twitter.cor	OmnibusLaw		
12	@iyoikuloh		Manusia meman	0	0	5	1,60232E+12	10 Oct 2020	iyoikuloh/status/	https://twitter.cor	MosiTidakPercaya | TolakOmnibusLaw | Indonesia		
13	@NOVENMUHIDI	ig : novenmuhidd	4. Dan yang palir	0	0	0	1,60233E+12	10 Oct 2020	NOVENMUHIDDI	https://twitter.cor	OmnibusLaw | TolakOmnibusLaw | DPRRI		
14	@lilyyulianti	lily yulianti farid	Ketergesa-gesaa	0	0	0	1,60233E+12	10 Oct 2020	lilyyulianti/status	https://twitter.cor	OmnibusLaw		
15	@yaudasigapapa	:)	Saya bakalan pe	1	0	1	1,60231E+12	10 Oct 2020	yaudasigapapa/s	https://twitter.cor	SaatnyaJokowiTurun | TolakOmnibusLaw | RakyatBo		
16	@ferryheilmann	FERRY HEILMA	presiden pilihan	0	0	2	1,60234E+12	10 Oct 2020	ferryheilmann/sta	https://twitter.cor	TolakOmnibusLaw		
17	@Bimateresna	Bima Teresna A	Jika kamu hidup	0	0	0	1,60235E+12	10 Oct 2020	Bimateresna/stat	https://twitter.cor	OmnibusLaw		
18	@ImamkuKamu	Kamulmamku?	Pernah saya beli	0	0	0	1,60242E+12	11 Oct 2020	ImamkuKamu/st	https://twitter.cor	OmnibusLaw		
19	@BroAri_1	Bori	Dalam ilmu sosi	0	0	0	1,60243E+12	11 Oct 2020	BroAri_1/status/	https://twitter.cor	TolakUUCiptaKerja | OmnibusLaw		
20	@muhtadi_ilham		Protokol kesehat	0	0	0	1,60245E+12	11 Oct 2020	muhtadi_ilham/st	https://twitter.cor	OmnibusLaw | COVID19 | Demokrasi		
21	@ImamkuKamu	Kamulmamku?	Akhirnya lahan t	0	0	0	1,60242E+12	11 Oct 2020	ImamkuKamu/st	https://twitter.cor	OmnibusLaw		
22	@Muharnikabdr	Jaisvu Imam Mal	Hutan di bakar p	0	0	2	1,60239E+12	11 Oct 2020	Muharnikabdr_/st	https://twitter.cor	RakyatMenolakUuOmniBusLaw | rezimantisIam | Tol		

Gambar 2. Data Sebelum Tahap Data Selection

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 897-910
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti



Gambar 3. Data Setelah Tahap Data Selection

Setelah dilakukan tahapan *data selection* terdapat 1 atribut berupa content dan 500 baris sebagai data yang akan digunakan dalam klasifikasi kelas sebuah sentimen. Setelah didapatkan data sebanyak 500, maka dilakukan sebuah labeling pada data yang dilakukan oleh pakar. Digunakan label sebanyak 2 yaitu positif dan negatif.

Tabel 1. Perbandingan Jumlah Ulasan pada Kelas/Label

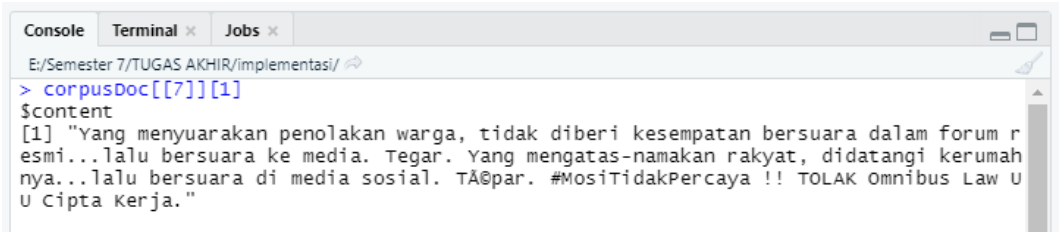
Kelas Sentimen	Jumlah Ulasan
Positif	106
Negatif	394
Total	**500**

3.2. Data Preprocessing

Data preprocessing digunakan agar gangguan terhadap data bisa berkurang. Berikut ini merupakan tahapan-tahapan dalam proses data preprocessing dengan software RStudio.

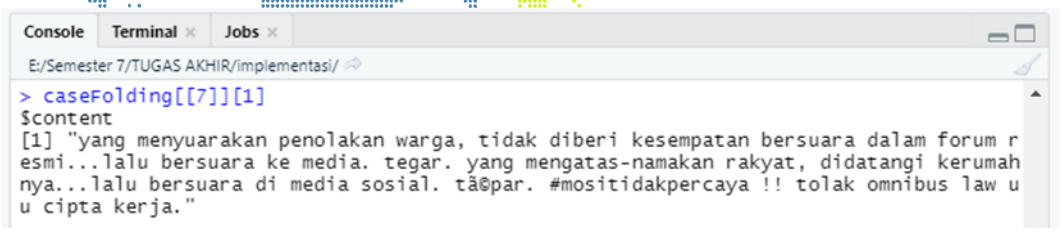
3.2.1. *Case Folding*

Tahapan *case folding* merupakan tahapan dengan mengubah seluruh huruf yang ada dalam data ulasan berupa teks dirubah menjadi huruf kecil (*lowercase*).



Gambar 4. Data Sebelum Tahap *Case Folding*

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 897-910
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

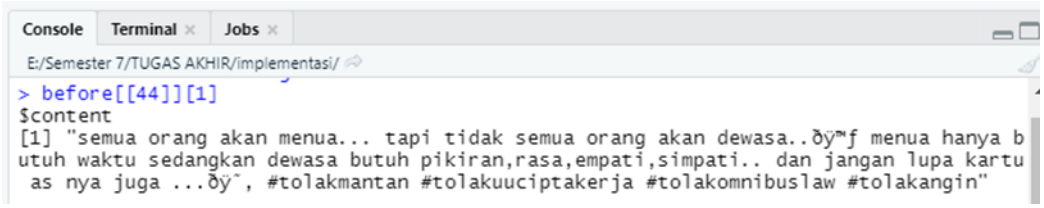


Gambar 5. Data Setelah Tahap *Case Folding*

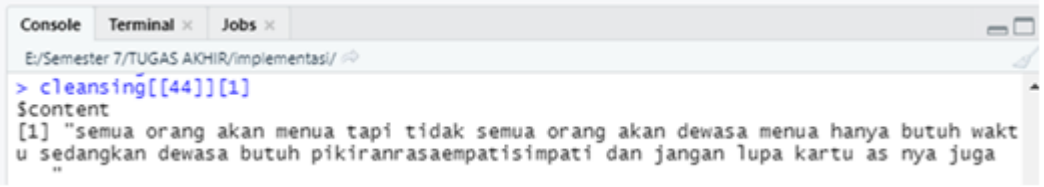
3.2.2. Cleansing

Karakter yang kurang berpengaruh akan dihilangkan atau dihapus pada tahapan cleansing ini. Karakter yang kurang berpengaruh antara lain berupa punctuation atau tanda baca, simbol-simbol dalam emoticon, URL atau link yang ada dalam ulasan, hastag, karakter dan angka

.



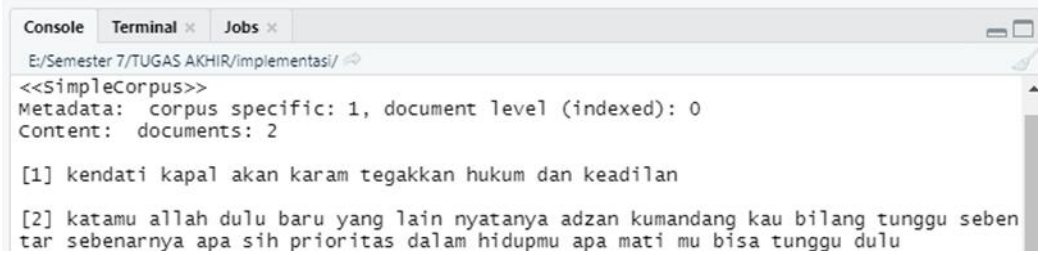
Gambar 6. Data Sebelum Tahap *Cleansing*



Gambar 7. Data Setelah Tahap *Cleansing*

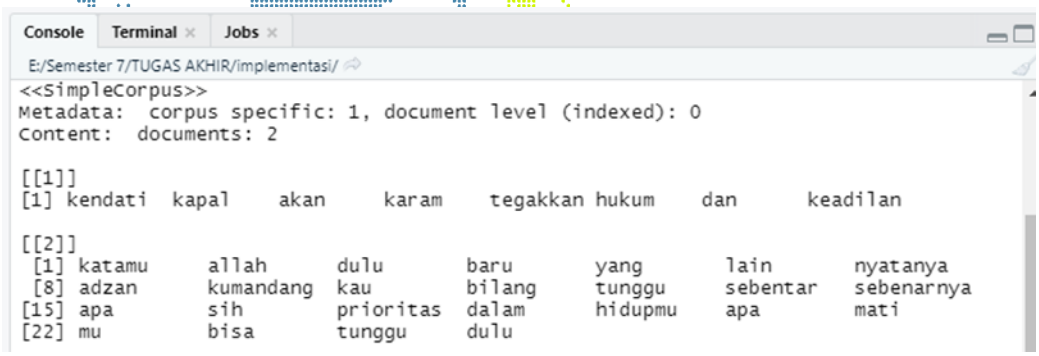
3.2.3. Tokenizing

Proses *tokenizing* merupakan tahapan dengan memisahkan setiap kata dalam sebuah kalimat atau ulasan yang ada berupa dokumen teks yang dihubungkan dengan karakter spasi.



Gambar 8. Data Sebelum Tahap *Tokenizing*

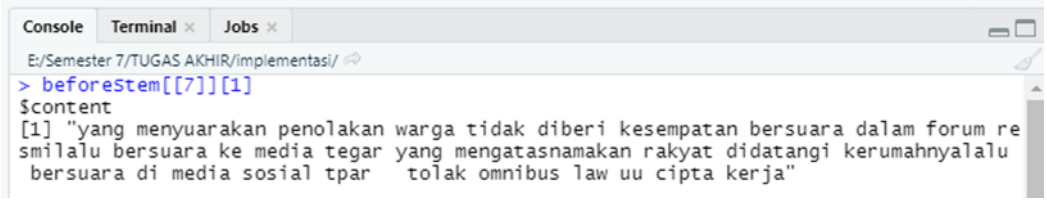
Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 897-910
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti



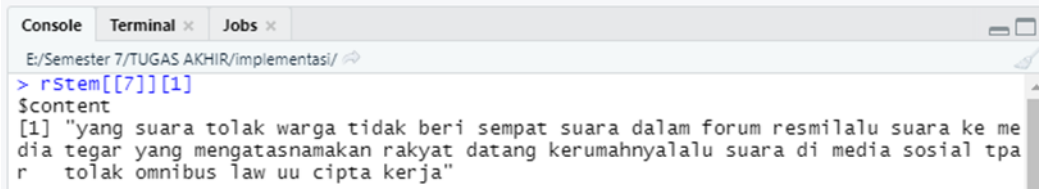
Gambar 9. Data Setelah Tahap *Tokenizing*

3.2.4. Stemming

Tahapan *stemming* merupakan tahapan menghapus huruf yang menjadi kata imbuhan dalam sebuah kata yang ada dalam ulasan yang telah dilakukan proses dalam data preprocessing sebelumnya. Imbuhan yang akan dihapus yaitu baik awalan, akhiran, maupun awalan dan akhiran sehingga menjadi kata dasar.



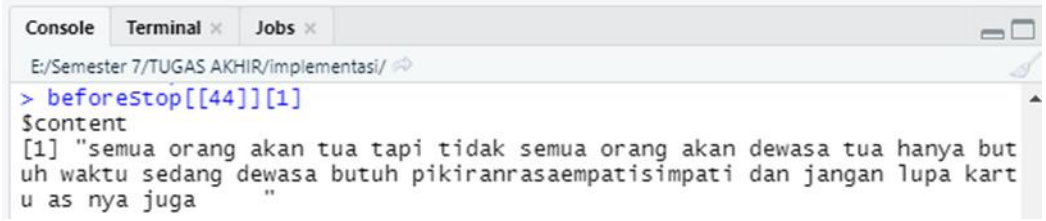
Gambar 10. Data Sebelum Tahap *Stemming*



Gambar 11. Data Setelah Tahap *Stemming*

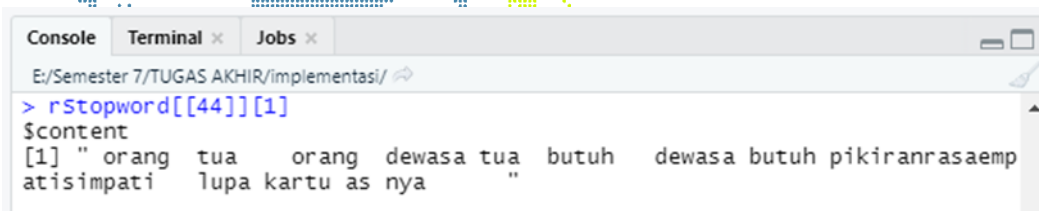
3.2.5. Stopword Removal

Tahapan stopword removal merupakan tahapan menghapus beberapa kata yang dirasa kurang berkepentingan dalam proses selanjutnya. Kata yang nantinya dihapus berupa kata penghubung "di", "dan", "yang", dan lain sebagainya.



Gambar 12. Data Sebelum Tahap *Stopword Removal*

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 897-910
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti



Gambar 13. Data Setelah Tahap *Stopword Removal*

3.3. *Transformation*

Sebelum nantinya dilakukan transformation dengan menggunakan TF-IDF (Tern Frequency Inverse Document Frequency) akan dilakukan terlebih dahulu tahapan feature selection dengan menggunakan Chi-Square.

3.2.1. *Seleksi Fitur Chi Square*

Langkah-langkah dalam penerapan feature selection Chi-Square diantaranya adalah:

a) Menghitung nilai chi-square berdasarkan persamaan.
b) Mengurutkan nilai chi-square.
c) Menentukan penggunaan fitur.

Tabel 2. Hasil dan Pengurutan 10 Data Teratas Nilai Chi Square

Fitur	Kategori	Nilai Chi
utk	Positif	20.425331
lahan	Positif	19.837317
jual	Positif	18.216837
desa	Positif	17.081110
rakyat	Positif	13.972787
nggk	Positif	12.144558
ubi	Positif	12.144558
rumah	Positif	11.733190
bisnis	Positif	11.363560
baca	Positif	11.357405

Dari nilai Chi Square yang telah didapatkan serta diurutkan, akan dilakukan skema penggunaan fitur sebanyak 4 kali yaitu sebanyak 25%, 50%, 75%, dan 100% fitur akan digunakan dengan dikombinasikan terhadap klasifikasi KNN dalam data mining.

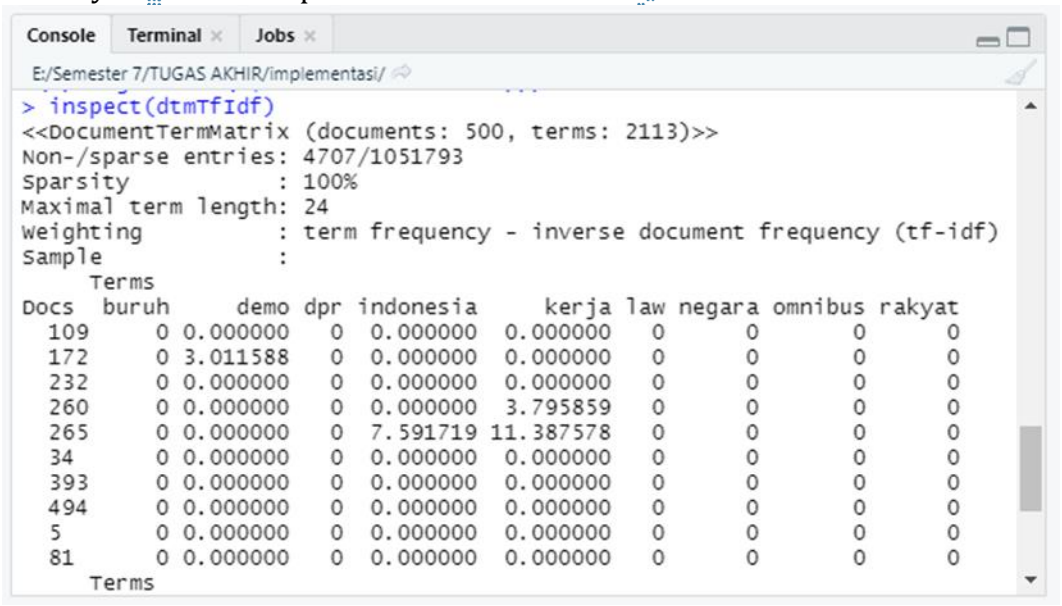
Tabel 3. Banyaknya Penggunaan Fitur dalam Beberapa Skema

Persentase Skema	Banyaknya Fitur
25%	1.099 Fitur
50%	2.197 Fitur
75%	3.295 Fitur
100%	4.394 Fitur

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 897-910
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

3.2.1. *TF-IDF*

Pembobotan dengan menggunakan TF-IDF dilakukan dengan menghitung banyaknya term dari masing-masing dokumen dengan frekuensi terbanyak akan ditampilkan terlebih dahulu.



Gambar 14. Hasil TF-IDF

3.4. Data Mining

Pada tahapan data mining ini dilakukan tahapan klasifikasi terhadap sentimen dari twitter dengan menggunakan algoritma K-Nearest Neighbor (KNN). Dari data hasil transformation tadi selanjutnya akan dilakukan sebuah permodelan dengan algoritma K-Nearest Neighbor dengan menggunakan K-Fold Cross Validation dengan nilai K adalah 10. Dengan menggunakan K-Fold Cross Validation nantinya data akan dibagi sebanyak K bagian yang sama dan akurasi ditentukan oleh data uji pada setiap bagian dan diambil rata-rata akurasinya. Data mining dilakukan sebanyak 4 kali sesuai dengan persentase fitur yang digunakan yaitu 25%, 50%, 75%, dan 100% (tanpa seleksi fitur). Klasifikasi yang digunakan adalah klasifikasi dengan K-Nearest Neighbor dengan dengan nilai k yang digunakan adalah 1, 3, 5, 7, 9, dan 11.

Tabel 4. Hasil Klasifikasi dengan Berbagai Persentase Penggunaan Fitur

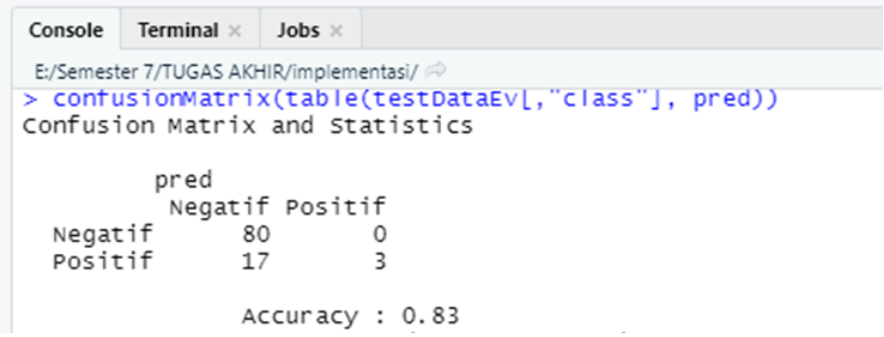
Persentase Penggunaan Fitur	K (KNN)	*Accuracy*
25%	1	0.8019944
	3	0.8120368
	5	0.8140776
	7	0.8121168
	9	0.8121168
	11	0.8121168
50%	1	0.8038880

Persentase Penggunaan Fitur	K (KNN)	*Accuracy*
	3	0.8119288
	5	0.7999664
	7	0.7999664
	9	0.8019272
	11	0.8019272
75%	1	0.7716575
	3	0.7820216
	5	0.7739416
	7	0.7359792
	9	0.7060576
	11	0.6941361
100% (tanpa seleksi fitur)	1	0.4524810
	3	0.3963850
	5	0.4180208
	7	0.2658631
	9	0.2239336
	11	0.2119712

Berdasarkan tabel 4.2, telah dilakukan 4 skema data mining sesuai dengan persentase penggunaan fitur serta penggunaan nilai k yaitur k = 1, 3, 5, 7, 9, dan 11. Dari hasil yang telah didapatkan dan ditampilkan melalui tabel 4.2, diketahui bahwa penggunaan fitur sebesar 25% dengan k = 5 merupakan hasil dengan akurasi cukup baik dan lebih baik dari skema lainnya yaitu sebesar 0.8140776 atau 81.4%.

3.5. *Evaluation*

Evaluasi digunakan untuk mengetahui apakah model yang telah dibuat dapat digunakan untuk memprediksi data terkait. Dalam melakukan evaluasi kali ini model yang sebelumnya telah dibuat yaitu dengan penggunaan fitur sebesar 25% dan dengan k =5. Data yang digunakan menggunakan data acak dari 500 data yang ada. Digunakan 100 data sebagai penguji dari model yang telah ada sebelumnya.



Gambar 15. Evaluasi Klasifikasi (*Convussion Matrix*)

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 897-910
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Berdasarkan gambar 15 diperoleh informasi terkait hasil pengujian klasifikasi K-Nearest Neighbor terhadap data uji diperoleh akurasi sebesar 83%. Sebagai informasi tambahan dan memperjelas hasil tersebut berikut ditampilkan perhitungan hasil kinerja dengan mencari nilai akurasi, precission, recall, dan F-Measure.

a) Akurasi

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Accuracy = \frac{3+80}{3+17+80+0}. = \frac{83}{100} = 83\%$$

b) Precission

$$Precission = \frac{TP}{TP+FP}$$

$$Precission = \frac{3}{3+0} = \frac{3}{3} = 1$$

c) Recall

$$Recall = \frac{TP}{TP+FN}$$

$$Recall = \frac{3}{3+17} = \frac{3}{3+17} = \frac{3}{20} = 0.15$$

d) F-*Measure*

$$F1\ score = \frac{2*precission*recall}{precission+recall}$$

$$F1\ score = \frac{2 * 1 * 0.15}{1+0.15} = \frac{0.3}{1.15} = 0{,}2608$$

4. SIMPULAN

Berdasarkan penelitian yang telah dilakukan disimpulkan beberapa hal sebagai berikut :

a) Klasifikasi pada sentimen twitter telah berhasil dilakukan dengan menggunakan seleksi fitur Chi Square dan diambil fitur sebanyak 25% dari data yang ada, dilakukan pembobotan dengan TF-IDF dan menggunakan K-Nearest Neighbor dengan nilai k = 5 serta K-Cross Validation dengan nilai K = 10. Dengan permodelan yang dilakukan diperoleh akurasi sebesar 81,4%. Setelah dilakukan pengujian terhadap model melalui 100 data acak yang diambil diperoleh 80 data kelas negatif diprediksi benar serta 3 data kelas positif diprediksi benar atau dalam kata lain 83 dari 100 data diprediksi benar.

b) Evaluasi dilakukan dengan melakukan sebuah pengujian model klarifikasi dengan KNN dan K Fold Cross Validation serta fitur seleksi Chi Square terhadap data acak yang diambil sebanyak 100 data. Dari pengujian tersebut diperoleh akurasi sebesar 83%, precission sebesar 100%, recall sebesar 15% dan nilai F-Measure sebesar 26,08%.

c) Dari 500 data yang diambil, kata “rakyat” merupakan kata yang paling mendominasi dari seluruh kata yang ada. Dari 500 data yang diambil

sebanyak 78,8% merupakan sentimen negatif dan 21,2% merupakan sentimen positif.

Sebagai pertimbangan pada penelitian selanjutnya dapat dilakukan dengan :

a) Menggunakan data yang lebih seimbang antara suatu label dengan label lainnya. Keseimbangan data antar label dapat mempengaruhi hasil dari klasifikasi.
b) Data yang digunakan dapat diperluas dengan menggunakan lebih dari 500 data. Jumlah data juga dapat mempengaruhi hasil dari klasifikasi.
c) Klasifikasi dengan K-Nearest Neighbor dapat dilakukan dengan menambah variasi nilai k pada K-Nearest Neighbor sebagai pertimbangan klasifikasi.

DAFTAR PUSTAKA

[1] Mahardhika, Y.S, and Zuliarso, E. "Analisis Sentimen Terhadap Pemerintahan Joko Widodo Pada Media Sosial Twitter Menggunakan Algoritma Naives Bayes." Prosiding SINTAK 2018 (2015):409–13, 2018.
[2] Ankit, and Saleena, N. "An Ensemble Classification System for Twitter Sentiment Analysis." Procedia Computer Science 132(Iccids):937–46. doi: 10.1016/j.procs.2018.05.109, 2018.
[3] Ain, Q.T., Ali,M., Riaz,A., Noureen,A., Kamran,M., Hayat,B., and Rehman, A. "Radiotherapy Is the Gold Standard in Treating Bone Malignancy . Effective in 50-90 % Expectancy Months)." 8(6), 2017.
[4] Nasution, Muhammad Rangga Aziz, and Mardhiya Hayaty. "Perbandingan Akurasi Dan Waktu Proses Algoritma K-NN Dan SVM Dalam Analisis Sentimen Twitter." Jurnal Informatika 6(2):226–35. doi: 10.31311/ji.v6i2.5129, 2019.
[5] Luthfiana, Lulu, Julio Christian Young, and Andre Rusli. "Implementasi Algoritma Support Vector Machine Dan Chi Square Untuk Analisis Sentimen User Feedback Aplikasi." XII(2):125–28, 2020.
[6] Han, Jiawei, Micheline Kamber, and Jian Pei. "Data Mining: Concepts and Techniques.", 2012.
[7] Kamila, Insanul, Ulya Khairunnisa, and Mustakim. "Perbandingan Algoritma K-Means Dan K-Medoids Untuk Pengelompokan Data Transaksi Bongkar Muat Di Provinsi Riau." Jurnal Ilmiah Rekayasa Dan Manajemen Sistem Informasi 5(1):119. doi: 10.24014/rmsi.v5i1.7381, 2019.
[8] R. Feldman dan J. Sanger, "*The Text Mining Handbook, New York: Cambridge University Press*", 2007.
[9] Rivki, M., and Bachtiar, A.M. "Implementasi Algoritma K-Nearest Neighbor Dalam Pengklasifikasian Follower Twitter Yang Menggunakan Bahasa Indonesia." Jurnal Sistem Informasi 13(1):31. doi: 10.21609/jsi.v13i1.500, 2017.
[10] Amrullah, Ahmad Zuli, Andi Sofyan Anas, and Muh Adrian Juniarta Hidayat. 2020. "Analisis Sentimen Movie Review Menggunakan Naive

Bayes Classifier Dengan Seleksi Fitur Chi Square." *Jurnal* 2(1):40–44. doi: 10.30812/bite.v2i1.804, 2020.

[11] Ipmawati, Joang, Kusrini, and Emha Taufiq Luthfi. "Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen." Indonesian Journal on Networking and Security 6(1):28–36, 2017.

[12] Deviyanto, A., and Wahyudi, M.D.R. "Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor." JISKA (Jurnal Informatika Sunan Kalijaga) 3(1):1. doi: 10.14421/jiska.2018.31-01., 2018.

[13] Ibrahim, N., Bacheramsyah, T.F., Hidayat,B., and Darana, S. "Pengklasifikasian Grade Telur Ayam Negeri Menggunakan Klasifikasi K-Nearest Neighbor Berbasis Android." ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika 6(2):288. doi: 10.26760/elkomika.v6i2.288, 2018

[14] Han, Jiawei, Micheline Kamber, and Jian Pei. "Data Mining: Concepts and Techniques.", 2012.

[15] Mardi, Y. "Data Mining : Klasifikasi Menggunakan Algoritma C4.5." Jurnal Edik Informatika 2(2):213–19. 2017.

[16] Gullo, F. "From Patterns in Data to Knowledge Discovery: What Data Mining Can Do." Physics Procedia 62:18–22. doi: 10.1016/j.phpro.2015.02.005. 2015.