Terbit *online* pada laman web jurnal : https://ojisnu.nuponorogo.or.id/index.php/ijitech/

Indonesian Journal Of Information Technology

| ISSN (Print) xxxx-xxxx | ISSN (Online) xxxx-xxxx |

Klik di sini dan tuliskan Kategori Artikel

PENERAPAN METODE *K-NEAREST NEIGHBOR* DAN *DECISION TREE* UNTUK ANALISIS SENTIMEN (Studi Kasus Mario Dandi)

***Yumarlin MZ*[1] *, Jemmy Edwin Bororing*[2] *, Nafisatul Fuadiah*[3]**

[1],[2],[3]Program Studi Informatika, Fakultas Teknik, Universitas Janabadra
Jalan Tentara Rakyat Mataram No. 55 – 57 Yogyakarta 55231
Email : [1]yumarlin@janabadra.ac.id, [2]jemmy@janabadra.ac.id,
[3]nafisafuadiah25@gmail.com

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 19 Oktober 2023
Revisi Akhir: 24 November 2023
Diterbitkan *Online*: 21 Desember 2023

KATA KUNCI

Twitter, Sentiment analysis, *K-Nearest Neighbor*, *Decision Tree*

KORESPONDENSI

Telepon: 08157986629

E-mail: yumarlin@janabadra.ac.id

A B S T R A C T

Media Sosial merupakan salah satu media komunikasi yang saat ini banyak digunakan oleh para pengguna internet, salah satunya adalah Twitter, sebuah situs yang menyediakan layanan mikroblog sehingga penggunanya dapat berbagi ide, pendapat atau sekedar kehidupan melalui postingan pendek yang disebut Tweets. Tweet pengguna Twitter sangatlah beragam, dari tweet tersebut terdapat data yang dapat diolah menjadi Analisis Sentimen sehingga dapat menjadi informasi yang bermanfaat bagi beberapa pihak. Penelitian ini bertujuan untuk membuat sistem analisis sentimen yang dapat menghasilkan data dan informasi berupa sentimen positif, sentimen negatif, sentimen netral. Metode yang digunakan untuk mengklasifikasikan sentimen adalah metode *K-Nerest Neighbor* dan *Decision Tree*. Masukan dari sistem ini berupa tweet dari masyarakat mengenai penganiayaan yang dilakukan oleh anak pegawai pajak mario dandi. Hasil dari sistem ini berupa visualisasi data untuk sentimen positif, sentimen negatif,dan sentimen netral.

Tahapan penelitian dalam melakukan analisis sentimen yakni Studi *Literature*, *Crawling*, *Translate*, *Preprocessing*, *Labeling*, *Split Data*, *Supervised Learning*, *Evaluasion* dengan menggunakan algoritma *K-Nearest Neighbor* dan *Decission Tree*. Hasil perhitungan algoritma *K-Nearest Neighbor* dan *Decission Tree* menggunakan 18.080 data komentar dari Twitter dengan #mariodandi, dengan hasil proses preprocessing sebanyak 2930 data dengan hasil 1.200 komentar negatif, 1.036 data komentar netral dan 687 data positif komentar. Untuk algoritma *K-Nearest Neighbor* menghasilkan tingkat akurasi sebesar 93,50% dan untuk algoritma *Decission Tree* menghasilkan tingkat akurasi sebesar 47,84%. Hasil akurasi algoritma *K-Nearest Neighbor* dalam data set ini tingkat akurasinya lebih baik dari *Decesion tree*.

1. PENDAHULUAN

Perkembangan ilmu dan teknologi saat ini sangat mempengaruhi perilaku anak yang membuat hilangnya nilai -nilai karakter bangsa. Hilangnya nilai -nilai karakter bangsa tidak hanya di pengaruhi oleh perkembangan ilmu dan teknologi saja akan tetapi di pengaruhi oleh lingkungan sekitar dan para generasi muda. Salah satu contoh kekerasan terhadap anak di Indonesia adalah kasus Mario Dandy yang merupakan anak dari Rafael Alun Trisambodo mantan pejabat eselon III di Ditjen Pajak (DJP) Kemenkeu Jakarta selatan. Mario melakukan penganiayaan terhadap David Ozora yang merupakan anak dari Pengurus Pusat Gerakan Pemuda Anshor (PP GP Anshor) Nahdlatul Ulama. Kasus ini mencuat dan menjadi topik hangat dikalangan masyarakat karena terungkapnya kasus bahwa Rafael Alun melakukan tindak pidana pencucian uang yang di taksir mencapai Rp 500 milyar dan juga melakukan penyelewengan kepemilikan

harta yang tidak dilaporkan secara utuh di dalam LHKPN sebesar Rp 56,1 miliar. Dalam kasus Mario Dandi, sentimen analisis dapat membantu dalam mengidentifikasi isu - isu yang kontroversial atau memicu emosi di antara masyarakat.

Dilihat dari permasalahan yang ada maka diperlukan analisis sentiment terhadap komentar masyarakat terhadap kasus penganiayaan / kekerasan yang dilakukan oleh anak pegawai pajak terhadap anak dari Pengurus Pusat Gerakan Pemuda Anshor (PP GP Anshor) Nahdlatul Ulama dengan menggunakan metode *K-nearest neighbor* yang merupakan metode untuk melakukan klasifikasi terhadap objek berdasarkan data jurnal pembelajaran yang jaraknya paling dekat dengan objek tersebut [1]. Metode *K-nearest neighbor* dapat menghasilkan prediksi yang akurat pada kasus-kasus di mana data training sedikit dan data testing hanya sedikit berbeda dari data training. Digunakan juga metode *decision tree* untuk model prediktif yang menggunakan struktur berbentuk diagram alur (*flowchart*) untuk membuat keputusan atau prediksi berdasarkan fitur masukan

2. TINJAUAN PUSTAKA

Sistem informasi adalah sekumpulan subsistem yang saling berhubungan, berkumpul bersama-sama dan membentuk satu kesatuan, saling berintegrasi dan bekerjasama antara bagian satu dengan yang lainnya dengan caracara tertentu untuk melakukan fungsi pengolahan data, menerima masukan (*input*) berupa data-

https://doi.org/10.25077/xxxxx

data, kemudian mengolahnya (*processing*), dan menghasilkan keluaran (*output*) berupa informasi[2].

2.1. Sentimen Analisis

Analisis Sentimen adalah suatu teknik mengekstrak data teks untuk mendapatkan informasi tentang sentimen bernilai positif, netral maupun negatif. Analisis sentimen diberikan oleh pengguna internet pada media sosial untuk memberikan suatu penilaian atau opini pribadi. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan opini terhadap sebuah masalah atau objek yang dikemukakan oleh seseorang, apakah cenderung berpandangan atau beropini negatif atau positif [3].

2.1.1. Kekerasan

Kekerasan diartikan dengan perihal (yang bersifat, berciri) keras, perbuatan seseorang atau sekelompok orang yang menyebabkan cedera atau matinya orang lain, dan paksaan.Kekerasan merupakan suatu ekspresi yang dilakukan oleh individu maupun kelompok baik secara fisik maupun verbal mencerminkan tindakan penyerangan pada kebebasan maupun martabat [4].

2.1.2. Twitter

Twitter memiliki pengguna di seluruh dunia dan sering digunakan oleh individu, organisasi, dan perusahaan untuk berkomunikasi dengan audiens mereka. Banyak orang menggunakan Twitter untuk berbagi berita, opini, gagasan, dan informasi lainnya dengan orang lain, sementara organisasi dan perusahaan menggunakan platform untuk mempromosikan merek mereka dan berinteraksi dengan pelanggan [5]. Twitter juga sering digunakan sebagai alat untuk aktivisme dan perubahan sosial, dengan banyak kampanye dan gerakan yang dimulai di platform. Salah satu fitur terkenal dari Twitter adalah penggunaan tagar atau hashtag, yang memungkinkan pengguna untuk mencari atau memilah tweet berdasarkan topik tertentu.

2.1.3. Metode K-Nearest Neighbor

K-Nearest Neighbor adalah proses untuk mengelompokkan data ke dalam kelas-kelas yang telah ditentukan sebelumnya berdasarkan jarak terdekat / tingkat kemiripan data tersebut dengan dataset / data latih yang ada. Nantinya data akan dikelompokkan ke dalam suatu kelas dengan melihat sejumlah "k" nilai jarak terdekat nya dengan data latih [6]. Metode KNN relatif mudah dipahami, bahkan bagi pemula di bidang *machine learning*. Konsep dasarnya adalah mencari data terdekat dari data baru untuk menentukan label atau nilai target dari data baru Metode KNN juga dapat menghasilkan prediksi yang akurat pada kasus-kasus di mana data training sedikit dan data testing hanya sedikit berbeda dari data training. Dalam penelitian ini, proses *K-Nearest Neighbor* meliputi 2 proses, yaitu : (1) menghitung Bobot Kata (TF-IDF) dan (2) menghitung Tingkat Kemiripan (*Cosine Similarity*).
Berikut ini rumus *Cosine Similarity:*

$$\cos(\Xi_{ij}) = \frac{\sum_k (d_{ik} d_{jk})}{\sqrt{\sum_k {d_{ik}}^2} \sqrt{\sum_k {d_{jk}}^2}} \quad (1)$$

2.1.4. Metode Decision Tree

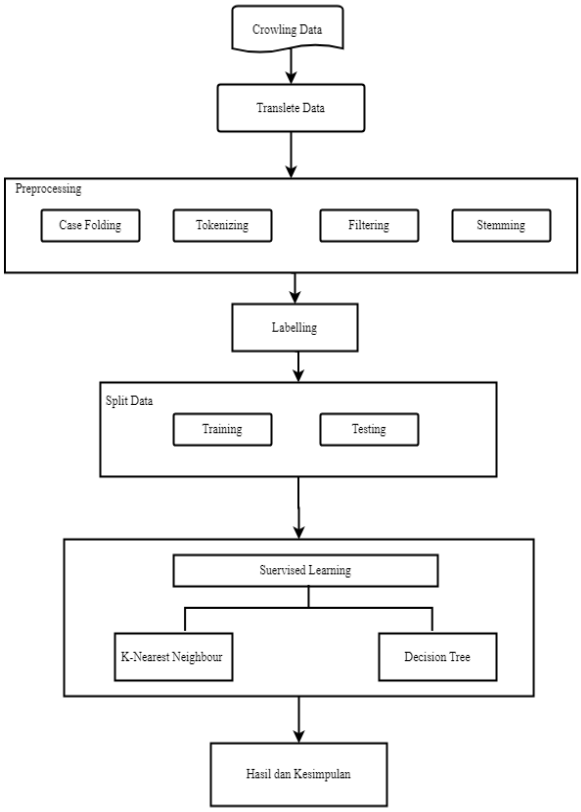
Decision tree merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode decision tree mengubah fakta yang sangat besar menjadi pohon keputusan yang mempresentasikan aturan [7]. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentuk basis data seperti Structure Query Language (SQL) untuk mencari record pada data tertentu. Sebuah decision tree adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Pada decision tree setiap simpul daun menandai label kelas. Simpul yang bukan simpul akhir terdiri dari akar dan simpul internal yang terdiri dari kondisi tes atribut pada sebagian record yang mempunyai karakteristik yang berbeda. Simpul akar dan simpul internal ditandai dengan bentuk oval dan simpul daun ditandai dengan bentuk segi empat.

2.1.5. Rapidminer

RapidMiner merupakan software/perangkat lunak untuk pengolahan data. Dengan menggunakan prinsip dan algoritma data mining, *RapidMiner* mengekstrak pola-pola dari data set yang besar dengan mengkombinasikan metode statistika, kecerdasan buatan dan database [8]. *RapidMiner* memudahkan penggunanya dalam melakukan perhitungan data yang sangat banyak dengan menggunakan operator-operator. Operator ini berfungsi untuk memodifikasi data. Data dihubungkan dengan node-node pada operator kemudian kita hanya tinggal menghubungkannya ke node hasil untuk melihat hasilnya. Hasil yang diperlihatkan RapidMiner pun dapat ditampilkan secara visual dengan grafik. Menjadikan *RapidMiner* adalah salah satu software pilihan untuk melakukan ekstraksi data dengan metode-metode data mining. *RapidMiner* menyediakan sejumlah fitur analitik data, seperti eksplorasi data, transformasi data, analisis statistik, dan pemodelan prediktif.

3. METODOLOGI

Terdapat 7 tahapan metode penelitian yaitu (1) *Crawling data*, (2*) Translate*, (3*) Preprocessing*, (4) *Labelling*, (5) *Split Data*, (6) *Supervised Learning*, (7) Evaluasi, dapat pada gambar 1.



Gambar 1. Metode Penelitian

3.1. Crawling Data

C*rawling* data adalah proses automatis untuk mengumpulkan dan mengindeks data dari berbagai sumber seperti situs web, database, atau dokumen. Proses ini menggunakan *software* atau aplikasi khusus yang disebut "*crawler*" untuk mengakses sumber data dan mengambil informasi yang dibutuhkan. Data yang

 https://doi.org/10.25077/xxxxx

dikumpulkan melalui *crawling* kemudian dapat diproses dan digunakan untuk berbagai tujuan, seperti analisis data, penelitian, atau pengembangan sistem informasi

3.2. Preprocessing

Pada tahap *pre-processing* mencakup operasi untuk pembersihan data (seperti menangani penghapusan noise dan data yang tidak konsisten), transformasi data (data diubah dan dikonsolidasikan kedalam bentuk yang sesuai) dan reduksi data, termasuk seleksi dan ekstraksi fitur. Setelah melalui tahap *pre-processing* diharapkan menjadi data final yang dianggap benar dan berguna untuk algoritma data mining. Pada penelitian lain *pre-processing* yang digunakan untuk mengambil dari sumber data media online yaitu data *cleaning, case folding, tokenizing, normalisasi, stemming dan stopword removal.*

3.2.1. Case Floding

Tahap ini mengubah huruf dalam dokumen menjadi huruf kecil. Selain itu, karakter non huruf akan dihilangkan. Contohnya "*There have been 2 cases*" diubah menjadi "*there have been cases*".

3.2.2. Tokenizing

Tokenizing adalah proses memecah teks atau dokumen menjadi unit-unit yang lebih kecil yang disebut token. Token adalah unit dasar dalam pemrosesan teks yang dapat dianalisis lebih lanjut, seperti kata, frasa, atau simbol. Sebagai contoh kalimat "*im also annoyed with the rubbicon*" ditokenize menjadi "*im", "also", "annoyed", "with", "the", "rubbicon*".

3.2.3. Filtering

Tahpan filtering untuk proses pemeriksaan kumpulan data untuk mengecualikan, mengatur ulang, atau membagi data menurut criteria tertentu. Misalnya, kalimat "*the fathers account was blocked*" setelah di filtering menjadi "*fathers account blocked*". *Stemming* adalah tahap transformasi suatu kata menjadi kata dasar (*rootword)* dengan menggunakan aturan-aturan tertentu. Contohnya pada kalimat "*fathers account blocked*" distemming menjadi "*father account block*".

3.3. Labelling

Di dalam tahapan *Labeling* penulis menggunakan kamus lexicon SentiWordnet untuk menentukan setiap kelas pada setiap komentar. Dari 2.930 data yang melalui tahap ini didapatkan hasil dari komentar negative sebanyak 1.200 data, komentar netral sebanyak 1.036 data dan komentar positif sebanyak 687 data.

3.4. Split data

Beberapa referensi atau *rule of thumb* yang sering digunakan dalam membagi data *train* dan *test* yaitu berdasarkan profesor Andrew Ng yang mengatakan bahwa *rule of thumb* dalam pembagian data split yaitu 70:30. Selain itu terdapat *rule of thumb* lain dalam pembagian data yaitu 80:20 atau dikenal dengan Pareto principle yang sering dipakai dalam matematika, ekonomi dan computer.

3.5. Supervised Learning

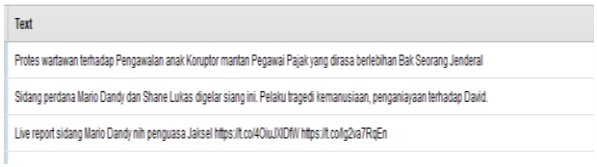
Supervised learning adalah salah satu jenis pembelajaran mesin (*machine learning*) di mana algoritma belajar menggunakan data yang sudah diketahui label atau targetnya. Dalam *supervised learning*, ada dua komponen utama, yaitu input (fitur) dan output (label). Tujuannya adalah untuk mengembangkan model atau algoritma yang dapat mempelajari pola dari data yang ada dan kemudian dapat memprediksi label atau output yang tepat untuk data baru. Penulis menggunakan 2 metode algoritma yang akan digunakan pada saat testing data, yaitu metode *K-Nearst Neigbor dan Decision Tree.*

Hasil dan Kesimpulan
Proses ini bertujuan untuk mengetahui tingkat akurasi pada metode klasifikasi yang digunakan dalam penelitian ini.

4. HASIL DAN PEMBAHASAN

4.1 Crawling

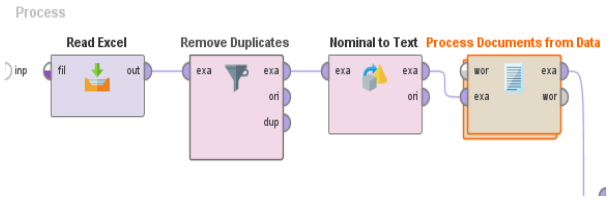
Crawling data mengambil data mentah dari twitter dengan #mariodandy. Data komentar yang didapat dalam proses *Crawling* dengan #mariodandy mendapatkan data sebanyak 18.080 data yang dilakukan pada tanggal 3 Juni 2023 dan 8 Juni 2023. Berikut ini merupakan beberapa contoh data ulasan yang digunakan dan terdapat dalam gambar 2.



Gambar 2 Hasil crawling data

4.2 Preprocessing

Data yang sudah di *translate* kebahasa Inggris akan diolah pada *tools Rapidminer* dengan beberapa tahap proses yang akan dilakukan dapat dilihat pada gambar 3.

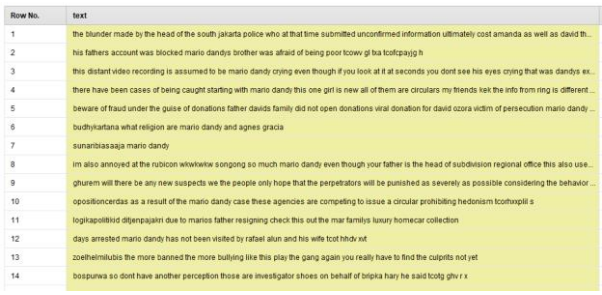


Gambar 3 *Preprocessing*

Preprocessing dilakukan agar data menjadi lebih bersih dan siap untuk dianalisis. Sebelum melakukan proses preprocessing, file excel terlebih dahulu diubah dalam bentuk file CSV. *Preprocessing* data dapat dilihat sebagai berikut:

4.1.1. Case Folding

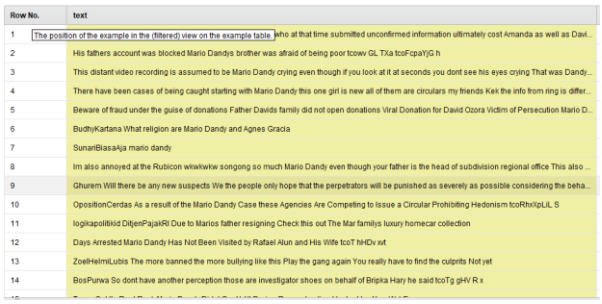
Hasil dari proses cleansing dapat dilihat pada gambar 4.



Gambar 4. Hasil case folding

4.1.2. Tokenizing

Proses *tokenizing* merupakan proses untuk menghilangkan karakter-karakter tertentu seperti angka dan tanda baca. Contoh data sesudah melalui tahapan *tokenizing* dapat dilihat pada gambar 5.

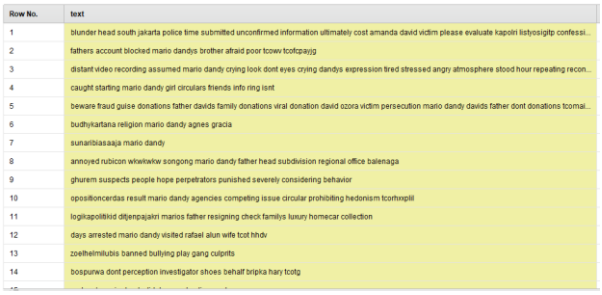


Row No.	text
1	The position of the example in the (filtered) view on the example table. who at that time submitted unconfirmed information ultimately cost Amanda as well as Davi...
2	His fathers account was blocked Mario Dandys brother was afraid of being poor tcoww GL TXa tcoFcpaYjG h
3	This distant video recording is assumed to be Mario Dandy crying even though if you look at it at seconds you dont see his eyes crying That was Dandy...
4	There have been cases of being caught starting with Mario Dandy this one girl is new all of them are circulars my friends Kek the info from ring is differ...
5	Beware of fraud under the guise of donations Father Davids family did not open donations Viral Donation for David Ozora Victim of Persecution Mario D...
6	BudhyKartana What religion are Mario Dandy and Agnes Gracia
7	SunariBiasaAja mario dandy
8	Im also annoyed at the Rubicon wkwkwkw songong so much Mario Dandy even though your father is the head of subdivision regional office This also ...
9	Ghurem Will there be any new suspects We the people only hope that the perpetrators will be punished as severely as possible considering the beha...
10	OpositionCerdas As a result of the Mario Dandy Case these Agencies Are Competing to Issue a Circular Prohibiting Hedonism tcoRhrXpLiL S
11	logikapolitikid DitjenPajakRI Due to Marios father resigning Check this out The Mar familys luxury homecar collection
12	Days Arrested Mario Dandy Has Not Been Visited by Rafael Alun and His Wife tcoT hHDv xvt
13	ZoelHelmiLubis The more banned the more bullying like this Play the gang again You really have to find the culprits Not yet
14	BosPurwa So dont have another perception those are investigator shoes on behalf of Bripka Hary he said tcoTg gHV R x

Gambar 5. Hasil tokenizing

4.1.3. *Filtering*

Proses *filtering* merupakan proses mengambil kata-kata penting setelah melakukan proses *tokenizing*. Contoh data sesudah melalui tahapan *filtering* dapat dilihat pada gambar 6.

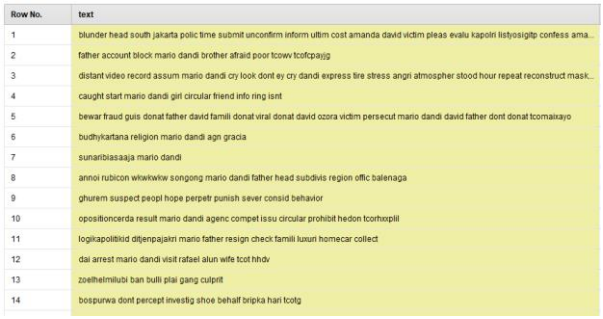


Row No.	text
1	blunder head south jakarta police time submitted unconfirmed information ultimately cost amanda david victim please evaluate kapolri listyosigitp confessi...
2	fathers account blocked mario dandys brother afraid poor tcoww tcofcpayjg
3	distant video recording assumed mario dandy crying look dont eyes crying dandys expression tired stressed angry atmosphere stood hour repeating recon...
4	caught starting mario dandy girl circulars friends info ring isnt
5	beware fraud guise donations father davids family donations viral donation david ozora victim persecution mario dandy davids father dont donations tcomai...
6	budhykartana religion mario dandy agnes gracia
7	sunaribiasaaja mario dandy
8	annoyed rubicon wkwkwkw songong mario dandy father head subdivision regional office balenaga
9	ghurem suspects people hope perpetrators punished severely considering behavior
10	oppositioncerdas result mario dandy agencies competing issue circular prohibiting hedonism tcorhrxplil
11	logikapolitikid ditjenpajakri marios father resigning check familys luxury homecar collection
12	days arrested mario dandy visited rafael alun wife tcot hhdv
13	zoelhelmilubis banned bullying play gang culprits
14	bospurwa dont perception investigator shoes behalf bripka hary tcotg

Gambar 6. Hasil *filtering*

4.1.4. *Stemming*

Proses *Stemming* merupakan proses menjadikan suatu kata yang berimbuhan menjadi kata dasarnya (*rootword*) dengan menggunakan suatu algoritma. Contoh data sesudah melalui tahapan stemming dapat dilihat pada gambar 7.



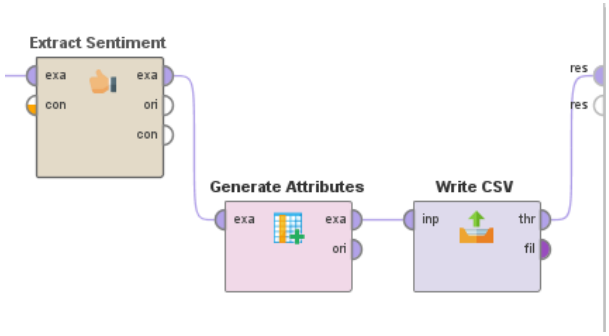
Row No.	text
1	blunder head south jakarta polic time submit unconfirm inform ultim cost amanda david victim pleas evalu kapolri listyosigitp confess ama...
2	father account block mario dandi brother afraid poor tcoww tcofcpayjg
3	distant video record assum mario dandi cry look dont ey cry dandi express tire stress angri atmospher stood hour repeat reconstruct mask...
4	caught start mario dandi girl circular friend info ring isnt
5	bewar fraud guis donat father david famili donat viral donat david ozora victim persecut mario dandi david father dont donat tcomaixayo
6	budhykartana religion mario dandi agn gracia
7	sunaribiasaaja mario dandi
8	annoi rubicon wkwkwkw songong mario dandi father head subdivis region offic balenaga
9	ghurem suspect peopl hope perpetr punish sever consid behavior
10	oppositioncerda result mario dandi agenc compet issu circular prohibit hedon tcorhrxplil
11	logikapolitikid ditjenpajakri mario father resign check famili luxuri homecar collect
12	dai arrest mario dandi visit rafael alun wife tcot hhdv
13	zoelhelmilubi ban bulli play gang culprit
14	bospurwa dont percept investig shoe behalf bripka hari tcotg

Gambar 7. Hasil stemming

Data yang tersisa setelah melalui proses *preprocessing* ini sebanyak 2.930 data.

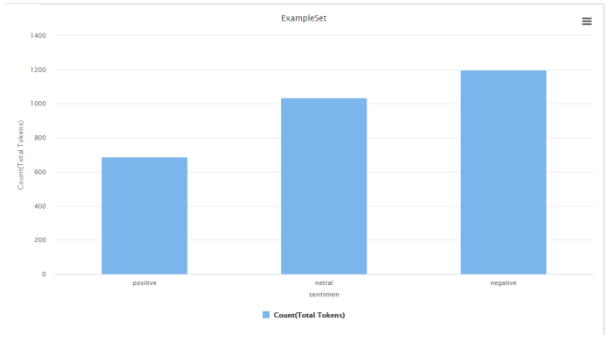
4.3 Labeling Data

Setelah proses *preprocessing* data kemudianambil data *training* yang telah dibagi diklasifikasikan kedalam sentimen (positif, netral dan negatif), berikut langkah *labeling* dan dapat dilihat pada gambar 8.



Gambar 8. Proses *labeling* data

Dari proses tersebut mendapatkan hasil *analisis sentimen* berupa sentimen positif sebanyak 687 data, sentimen negatif sebanyak 1200 data dan hasil sentimen netral sebanyak 1036 data. Grafik hasil *labeling* dapat dilihat pada gambar 9.



Gambar 9. Grafik *labeling*

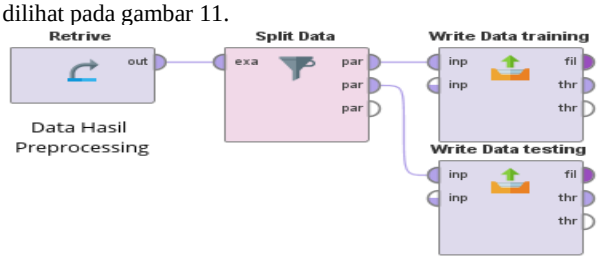
4.4 Wordcloud

WordCloud merupakan gambar yang terdiri dari kumpulan kata, dimana besarnya kata mempresentasikan kemunculan kata. Semakin besar suatu kata, maka semakin sering kata tersebut muncul dalam sebuah dokumen. *WordCloud* dapat dilihat pada gambar 10.

Gambar 10. *Wordcloud*

4.5 Split Data

Setelah proses split data kemudian dataset dibagi menjadi data training dan data testing. Berikut langkah split data dan dapat dilihat pada gambar 11.



Gambar 11. Proses split data

https://doi.org/10.25077/xxxxx

4.6 K-Nearest Neighbour

Setelah proses *labeling* kemudian data di testing menggunakan algoritma KNN. Diketahui terdapat 6 dokumen yang ada pada tabel 1 dimana dari D1 sampai D5 sebagai data *training* dan Q sebagai data *testing*.

Tabel 1. dokumen TF-ID

D1	D2	D3	D4	D5	D6 / Q
Brand	father	Mario	dettolid	cry	reveal
Shoes	account	Dandi	look	traumat	reconstruct
Mario	block	Look	reconstruct	with	persecute
Dandi	mario	Cool	mario	scream	david
reconstruct	dandi	david	dandi	stop	ozora
highlight	brother	abuse	agnes	help	mario
Police	afraid	reconstruct	stage	david	dandi
investig	poor	wear		mario	smoke
		shoes		dandi	david
		worth		didnt	torture
		million		reconstruct	reconstruct
					model

Untuk menghitung tingkat kemiripan *cosine similarity dapat menggunakan rumus berikut:*

$$\cos(\Xi_{ij}) = \frac{\sum_k (d_{ik} d_{jk})}{\sqrt{\sum_k {d_{ik}}^2} \sqrt{\sum_k {d_{jk}}^2}} \qquad (2)$$

Mencari nilai Dx Q yaitu dengan melakukan perkalian antara TF-IDF masing-masing D terhadap TF-IDF*Q, kemudian dicari totalnya:

Contoh:

DxQ (“dandi” D1*Q) = TF-IDF(D1”dandi”)*TF-IDF(Q”dandi”)

=0,93305321*1,866106421

= 1,741176587

Hasilnya dapat dilihat pada tabel 2.

Tabel 2. TF-IDF dengan Q

TF-IDF dengan Q					
TERM	D1	D2	D3	D4	D5
abus	0	0	0	0	0
cry	0	0	0	0	0
dandi	1.741176587	1.741176587	1.741176587	1.741176587	1.741176587
david	0	0	2.766381299	0	2.766381299
...	...	...	...	...	...
Jumlah	5.482353174	3.482353174	8.248734473	5.482353174	8.248734473

Selanjutnya mencari kuadrat dari TF-IDF, yang kemudian ditotalkan lalu hasil tersebut diakar kuadratkan.

Contoh:

Kuadrat TF-IDF (“abuse” D3)

= 1,77815125^2

= 3,161821869

Hasil dapat dilihat pada tabel 3.

Tabel 3. TF-IDF Kuadrat

TF-IDF ^ 2						
TERM	D1	D2	D3	D4	D5	D6/Q
abus	0	0	3.161821869	0	0	0
account	0	3.161821869	0	0	0	0
afraid	0	3.161821869	0	0	0	0
agn	0	0	0	3.161821869	0	0
...	...	...	...	...	...	...
Jumlah	17.57035126	20.7121078	25.27718565	12.22664219	26.25712032	49.7616759
Akar	4.191700283	4.55105568	5.027642156	3.496661578	5.124170208	7.054195624

Setalah dilakukan pembobotan pada TF-IDF, kemudian dikalikan hasil antara data *training* dan data *testing* sehingga menghasilkan nilai yang nantinya dipakai dalam tahap selanjutnya. Dapat dilihat pada tabel 4.

Tabel 4. Hasil Perhitungan*Cosine Similarity*

D1	D2	D3	D4	D5
0,18540835	0,101033026	0,23258166	0,222262353	0,228200335

Dari hasil tersebut dilakukan pengurutan nilai *similarity* Q dari yang tertinggi ke nilai yang terendah yaitu pada tabel 5.

Tabel 5. Hasil Urutkan Nilai *Similarity* Q

	Cosine Similarity	Sentimen
D3	0,23258166	positive
D5	0,228200335	negative
D4	0,222262353	netral
D1	0,18540835	positive
D2	0,101033026	negative

Data kemudian di *testing* menggunakan algoritma *K-Nearest Neighbor*, hasilnya dapat dilihat pada tabel 6.

Tabel 6. Nilai analisis sentimen metode *K-Nerest Neighbor*

	True Positif	True Negatif	True Netral	Class Precision
Pred. Positif	121	8	0	93,30%
Pred. Negatif	14	199	13	88,05%
Pred. Netral	3	0	227	98,70%
Class Recall	87,6%	96,14%	94,58%	

Selanjutnya dilakukan perhitungan akurasi dari hasil tersebut:

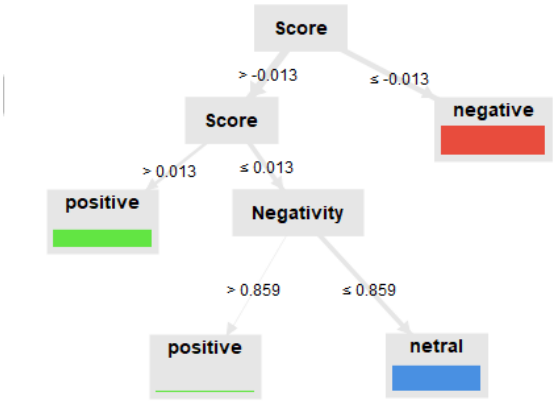
$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} x100 \qquad (3)$$

$$Accuracy = \frac{121 + 199 + 277}{121 + 14 + 3 + 8 + 199 + 13 + 227}$$

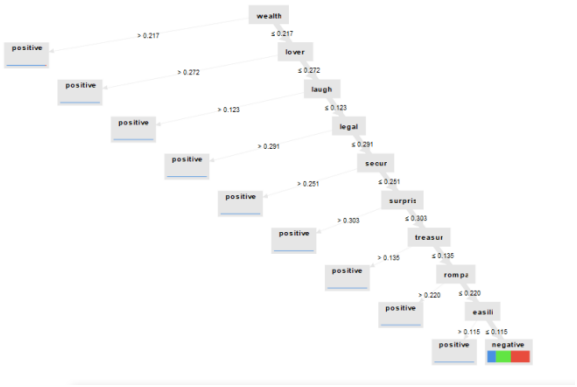
$$= \frac{574}{585} x100 = 93{,}50\%$$

4.7 Decision Tree

Data kemudian di *testing* menggunakan algoritma *Decision Tree.* Hasilnya dapat dilihat pada gambar 11, 12 dan tabel 7.



Gambar 11. Decision Tree 1



Gambar 12. Decision Tree 2

Tabel 7. Nilai analisis sentimen metode Decision Tree

	True Positif	True Negatif	True Netral	Class Precision
Pred. Positif	212	0	2	99,07%
Pred. Negatif	0	0	0	0,00%
Pred. Netral	652	1132	1428	44,46%
Class Recall	24,54%	0,00%	99,86%	

Selanjutnya dilakukan perhitungan akurasi dari hasil tersebut:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} x100 \quad (4)$$

$$Accuracy = \frac{212 + 1428}{121 + 652 + 1132 + 1428} == 47{,}84\%$$

Hasil dan Kesimpulan dari data set yang di peroleh dengan hasil klasifikasi sentimen positif sebanyak 687, negatif sebanyak 1200, dan netral sebanyak 1036. Tingkat akurasi K-Nerest Neighbour sebesar 93,50%, dan tingkat akurasi pada Decision Tree sebesar 47,84%.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil pengujian dan analisis yang diperoleh, kesimpulan yang didapatkan dari penelitian ini adalah:
Data komentar yang didapat dalam proses *crawling* mendapatkan data sebanyak 18.507, kemudian data melalui proses *preprocessing* terdapat beberapa tahapan yaitu *cleaning, case folding, tokenizing, filtering dan stemming,* dimana dari proses tersebut didapatkan data ulasan yang sebanyak 18.507 menjadi 2.930 data. Hasil dari analisis sentimen terhadap komentar pada twitter dengan kasus penganiayaan oleh anak pegawai pajak didapatkan sentimen positif sebanyak 687 data, sentimen negatif sebanyak 1200 data dan hasil sentimen netral sebanyak 1036 data. Pada proses yang dilakukan dengan akurasi dengan menggunakan algoritma *K-Nearest Neighbor* adalah 93,50% dan algoritma Decision Tree 47,84%. Dari hasil akurasi dalam dataset dalam penelitian ini menunjukan bahwa algotitma *K-Nearest Neighbor* lebih baik dibandingkan algoritma *Decision Tree.*
Beberapa hal yang sarankan yaitu penambahan waktu menambang data pada saat proses crawling data sehingga dalam meningkatkan ketepatan pengkategorian data, dan untuk menangani klasifikasi dapat menggunakan metode lain seperti metode SVM.

DAFTAR PUSTAKA

[1] Akhmad, Romadhoni, B., Karim, K., Tajibu, M. J., and Syukur, M, "The Impact of Fuel Oil Price Fluctuations on Indonesia's Macro Economic Condition. International," *Journal of Energy Economics and Policy.*, vol. 9 no. 2, pp. 277–282, 2019.A.

[2] Yanuardi, & Permana, A. A., "Rancang Bangun Sistem Informasi Keuangan Pada Pt. Secret Discoveries Travel and Leisure Berbasis Web," *Jurnal Teknik Informatika*, pp. 1–7, 2018.

[3] Dewi Susanti, and Fransiska Ristiana, "Universitas Pendidikan Ganesha, Indonesia," MIMBAR PGSD Undiksha. vol. 6, no. 3, 2018.

[4] KBBI. (2016). *Kamus Besar Bahasa Indonesia (KBBI)*. [Online] Available at: http://kbbi.web.id/pusat, [Diakses 15 Mei 2023].

[5] Tankovska, H. (2021, Februari 9). Global social networks ranked by number of users 2021. Retrieved from Statista: https://www.statista.com/statistics/272014/global-socialnetworks-ranked-by-number-of-users/

[6] Deviyanto and M. D. R. Wahyudi, "Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor," *JISKA (Jurnal Inform. Sunan Kalijaga).*, vol. 3, no. 1, pp. 1, 2018.

[7] Muzakir, A., and Wulandari, R. A., "Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree," *Scientific Journal of Informatics*, vol. 3 no. 1, pp. 19–26, 2016.

[8] Rahmat Brillian C.T.I, Agum Agidtama Gafar, dkk., "Implementasi K-Means Clustering Pada RapidMiner Untuk Analisis Daerah Rawan Kecelakan," *Jurnal Sains dan Informatika, Paper yang dipresentasikan pada Seminar Nasional Riset Kuantitatif Terapan. Kendari : Universitas Halu Oleo*, 2017.

 https://doi.org/10.25077/xxxxx