JURNAL INFORMATIKA, Vol.10 No.1 April 2023, Halaman 45-53
ISSN: 2355-6579 | E-ISSN**:** 2528-2247
DOI: https://doi.org/10.31294/inf.v10i1.14546

Sentiment Analysis Terhadap Perspektif Warganet Atas Tragedi Kanjuruhan Malang di Twitter Menggunakan Naïve Bayes Classifier

Minardi[1], Ranita Lasepa[2], Santoso Riyadi[3], Syahrur Ramadhan[4], Dedi Dwi Saputra[5]

[1,2,3,4,5] Fakultas Teknologi dan Informasi, Universitas Nusa Mandiri
Jalan Kramat Raya No.18, Jakarta, Indonesia

e-mail: [1]11220070@nusamandiri.ac.id, [2]11220230@nusamandiri.ac.id,
[3]11220099@nusamandiri.ac.id, [4]11220121@nusamandiri.ac.id, [5]dedi.eis@nusamandiri.ac.id

Informasi Artikel Diterima: 05-12-2022 Direvisi: 20-12-2022 Disetujui: 05-01-2023

Abstrak

Situs media sosial *Twitter* adalah tempat di mana pengguna Internet di seluruh dunia dapat bertukar perspektif tentang diskusi terkini. Salah satunya sepak bola, olahraga ini merupakan hobi yang digandrungi oleh seluruh penjuru dunia, termasuk warga Malang, dengan kecintaan mereka terhadap olahraga tersebut mereka menamakan dirinya Aremania yaitu suporter tim Arema Malang, namun terjadi peristiwa kelam. Kanjuruhan di Stadion Malang pada 01/10/2022, memunculkan pandangan berbeda dari semua akun pengguna Twitter, yang menyebabkan peningkatan tweet dan menjadi trending topik saat itu. Untuk mengembangkan perspektif yang berbeda berdasarkan apa yang membawa keuntungan dan kerugian di komunitas, diterapkan prosedur untuk mengklasifikasikan perspektif positif atau negatif pengguna Twitter melalui analisis sentimen dengan pengklasifikasi Naïve Bayes. Analisis sentimen dilakukan dengan mengindeks *tweet* pengguna *Twitter* dengan tagar UsutTuntasTragediKanjuruhan, mengambil (*crawling*) data 1.500 tweet yang ada sebagai kumpulan data (dataset), setelah itu data untuk diproses diidentifikasi *(labeling)* untuk langkah selanjutnya yaitu tahap *Preprocessing* data yang terdiri dari *Annotation Removal*, *Remove Hashtag*, *Transformation Remove Url*, *Regexp*, *Indonesian Steaming*, *Indonesian Stopword Removal* dipadukan dengan operator *Smote Upsampling*. Pembuatan Confusion Matrix yang menunjukkan hasil akhir analisis berjalan dengan baik yaitu nilai *accuracy* 77,67%, nilai *precision* sebesar 77,19%, nilai *recall* sebesar 78,50%, dan nilai *AUC* 0.820 (*good classification*).

Kata Kunci: Sentiment Analysis; Tragedi Kanjuruhan Malang; *Naïve Bayes Classifier*

Abstract

The social media site Twitter is a place where Internet users around the world can exchange perspectives on current discussions. One of them is football; this sport is a hobby that is loved by all corners of the world, including the people of Malang. With their love for this sport, they call themselves Aremania, namely Arema Malang team supporters, but a dark incident occurred. The Kanjuruhan at Malang Stadium on January 10, 2022, raised different views from all Twitter user accounts, which led to an increase in tweets and became a trending topic at that time. To develop different perspectives based on what brings advantages and disadvantages to the community, a procedure was applied to classify Twitter users' positive or negative perspectives through sentiment analysis with the Nave Bayes classifier. Sentiment analysis was carried out by indexing Twitter user tweets with the hashtag "UsutTuntasTragediKanjuruhan," crawling data from 1,500 existing tweets as a dataset, after which the data to be processed is identified. (labeling) for the next step, namely stage Data preprocessing includes annotation removal, hashtag removal, URL removal, regexp, Indonesian steaming, and Indonesian stopword removal, as well as operators' smote upsampling. Making a confusion matrix that shows the final result of the analysis is going well, namely the value accuracy of 77.67%, the value precision of 77.19%, the value recall of 78.50%, and the value AUC of 0.820 (good classification).

Keywords*: Sentiment Analysis; Kanjuruhan Malang Tragedy; Naïve Bayes Classifier*

1. Pendahuluan

Suporter sepak bola adalah nyawa atau penghidup sebuah klub. Tidak hanya sebagai pendukung klub, suporter sepak bola adalah sebuah identitas dari kota klub sepak bola tersebut. Sejarah *Hooliganisme* sepak bola di Indonesia sudah ada sejak tahun 1990-an, *hooliganisme* didefinisikan sebagai paham budaya yang terkait dengan perilaku nakal dan destruktif yang biasa dilihat oleh kelompok penggemar sepak bola seperti perkelahian, vandalisme, atau perilaku mengancam (Hendika & Nuraeni, 2020)**.**

Pada tanggal 1 Oktober 2022 telah terjadi pertandingan sepak bola yang mempertemukan dua tim yang dianggap bersaing keras, Arema dan Persebaya, di stadion Kanjuruhan, Malang, Jawa Timur. Pertandingan ini menjadi salah satu *trending* topik di Twitter yang memunculkan berbagai *tweet perspektif* warganet yang memicu timbulnya pro-kontra antara masyarakat, karena terjadi insiden yang sangat di sesalkan, yang merenggut korban baik dari kalangan supporter maupun bukan supporter di pertandingan tersebut.

Analisis sentimen masih merupakan bagian dari penelitian penambangan opini**.** Artinya**,** proses memahami, mengekstrak**,** dan memproses data teks secara otomatis untuk mendapatkan informasi sentimen didalam kalimat opini (Buntoro, 2019). Pada Penelitian ini penulis melakukan Analisis Sentimen *tweet* warganet media sosial twitter terhadap tragedi yang terjadi dari insiden pertandingan sepak bola di Kanjuruhan Malang Jawa Timur. Salah satu yang dilakukan adalah melakukan klasifikasi dari *tweet* yang ber *hashtag* UsutTuntasTragediKanjuruhan di Twitter. Penilitan ini menggunakan metode klasifikasi algoritma *Naïve Bayes* untuk mengklasifikasikan *perspektif* positif atau negatif dari *tweet* tersebut.

2. Metode Penelitian

Sentiment Analysis akan diterapkan dalam percobaan ini untuk mengekstrak pengetahuan tentang bagaimana ulasan positif atau negatif yang terlihat seperti dalam format teks. Dengan menggunakan *sentiment analysis*, dataset tersebut akan dipelajari dengan melakukan *training* dan *testing* (*supervised machine learning*) untuk mengklasifikasikan ulasan mana yang mewakili sentimen positif atau negatif (Hakim, 2021). Polaritas dataset besar itu akan diturunkan dengan menggunakan teknik tingkat fitur untuk kalimat yang lebih kompleks. Perbandingan antara teks *preprocessing* yang dilakukan dan diimplementasikan di masing-masing dataset akan dijadikan bahan evaluasi pada penelitian ini.

Metode yang digunakan dalam penilitan ini adalah metode *Naive Bayes***.** *Naive Bayes* adalah metode klasifikasi berbasis probabilitas sederhana yang dimaksudkan untuk digunakan dengan asumsi bahwa tidak ada saling ketergantungan antara satu kelas dengan kelas lainnya. (Nurmalasari et al., 2021).

Klasifikasi adalah suatu *fungsionalitas data mining* yang menghasilkan model untuk memprediksi kelas atau kategori dari objek - objek didalam basis data (Susana & Suarna, 2022). Pengklasifikasi probabilistik akan mengembalikan kelas yang memiliki probabilitas posterior maksimum di setiap dokumen yang diberikan.

Probabilitas bayes digunakan untuk menyelesaikan permasalahan ketidakpastian data berdasarkan formula bayes yang dinyatakan (Imamah & Siddiqi, 2019).

$$P(H|X) = \frac{P(H|X).\ P(H)}{P(X)} \quad \text{..... (1)}$$

Keterangan:

X	: Data yang belum di ketahui kelasnya.
H	: Hipotesis data X suatu kelas
P(H|X)	: Probabilitas hipotesis H pada kondisi X (posteriori probability)
P(H)	: Probabilitas hipotesis H (prior probability)
P(X|H)	: Probabilitas X pada kondisi hipotesis H
P(X)	: Probabilitas X

Dengan demikian komputasi kelas kemungkinan diberikan dalam dokumen dengan memilih kelas yang mencapai produk tertinggi dari dua probabilitas, atau disebut probabilitas sebelumnya dari kelas dan kemungkinan probabilitas dokumen terjadi dalam metode ini.

Penelitian lain yang bisa menggambarkan penggunaan metode *Naïve bayes* pada penelitian yang dilakukan oleh (Astiningrum Mungki et al., 2020) dengan judul “Analisis Sentimen tentang opini terhadap perfoma timnas sepak bola Indonesia pada Twitter”. Dimana dalam penelitian ini algoritma *Naïve Bayes* dapat digunakan untuk mengklasifikasikan *tweet* positif atau negatif tentang perfoma timnas sepak bola Indonesia. Dari tiga pengujian didapatkan hasil nilai algoritma *Naive Bayes* pada komposisi data *training* 371 dan data *testing* 159 sebesar 78%, komposisi data *training* 424 dan data *testing* 106 sebesar 84% dan komposisi data *training* 477 dan data *testing* 53 sebesar 87%. Nilai akurasi

terendah adalah 78% dan tertinggi adalah 87%. Setiap terjadi penambahan komposisi data *training*, maka nilai dari *accuracy*, *precision* dan *recall* juga mengalami peningkatan. Hal ini dikarenakan algoritma *Naive Bayes* merupakan algoritma yang sangat bergantung pada data *training*, kemungkinan akurasi dapat ditingkatkan lagi dengan menambahkan data *training* yang lebih banyak lagi. Seperti yang ditunjukkan oleh hasil penelitian ini, akurasi meningkat setiap kali ada data pelatihan tambahan**.**

Pada Judul "*Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning*" (Patel & Passi, 2020). Metode *Naive Bayes* digunakan untuk mengklasifikasikan sentimen positif atau negatif tentang piala dunia 2014 di Brazil. Tahap-tahap *pre-processing* yang dilakukan adalah *word tokenization, word stemming and lemmaztizing, removing URLs, Rename and removing of "RT@username" from tweets, filtering #Hashtag, removing repeated characters, removing special characters.* Dari hasil penelitian *Naïve Bayes* memberikan akurasi terbaik sebesar 88.17%.

Dalam studi literatur penulis mengenai penggunaan metode *Naïve Bayes* ini salah satu jurnal yang menarik untuk di pelajari adalah karya tulis (Hartati et al., n.d.) dengan judul "Optimasi Analisis Sentimen Pada *Twitter* Olshop Tokopedia Menggunakan Textmining Dengan Algoritma *Naïve Bayes & Adaboost.* Dari hasil penelitian ini menunjukkan bahwa Algoritma algoritma *Naïve Bayes & Adaboost* jika dioptimasi dengan menggunakan Synthetic Minority Over-sampling Technique (SMOTE) menghasilkan milai accuracy: 94.95%,precision: 90.86% , recall: 100.00% dan AUC: 0.950. pada penelitian ini juga menggunakan algoritma *Naïve Bayes* dengan optimasi SMOTE menghasilkan accuracy: 95.16%, precision: 91.29%, recall: 100.00% dan AUC: 0.780. sedangkan algoritma *Naïve Bayes* tanpa optimasi SMOTE menghasilkan accuracy: 86.80%, precision: 23.29%, recall: 49.33%, dan AUC: 0.511. jadi berdasarkan hasil penelitian ini, dapat disimpulkan bahwa Algoritma *Naïve Bayes & Adaboost* yang dioptimasi dengan fitur Synthetic Minority Over-sampling Technique (SMOTE) adalah klasifikasi yang lebih baik digunakan dibandingkan dengan Algoritma *Naïve Bayes* dengan optimasi SMOTE maupun Algoritma *Naïve Bayes* tanpa optimasi SMOTE.

Sementara itu dalam karya tulis "Klasifikasi dan Analisis Sentimen Pada Data Twitter menggunakan algoritma *Naïve Bayes* (Studi Kasus: Timnas Indonesia senior, U-23, dan U-19)" (Prajamukti & Mega Santoni, 2021). Metode *Naïve Bayes* digunakan untuk mengklasifikasikan *tweet* positif atau negatif yang masyarakat berikan tentang timnas sepak bola Indonesia. Alur penelitian dilakukan secara bertahap yaitu identifikasi masalah, studi literatur, akuisisi data dari Twitter, pra proses data, pembobotan *Term TF-IDF*, klasifikasi *Naïve Bayes,* hasil, dan evaluasi. Algoritma *Naïve Bayes* pada klasifikasi tweet tentang sentimen terhadap Timnas Indonesia berjalan dengan baik dengan nilai *accuracy* 83%.

Untuk metode *Naïve Bayes* ada beberapa karya tulis yang mampu menambah khasanah pengetahuan penulis dalam menggunakan metode *Naïve Bayes.* Penelitian tersebut adalah "Text Mining untuk Sentimen Analisis dengan Metode *Naïve Bayes*, SMOTE, N-Gram dan AdaBoost pada Twitter CommuterLine" (Pratama Putra et al., 2022). Dimana pada tulisan ini penulis menambahkan *feature selection SMOTE, N-Gram,* dan *Adaboost.* Berdasarkan hasil penelitian yang telah dilakukan mengenai permasalahan peng-kategorian tweet dengan sentiment "*Positive*" dan "*Negative*" terhadap pelayanan PT KAI Commuter, di sosial media Twitter dengan menggunakan *mention* @CommuterLine dapat ditarik kesimpulan yaitu, pendekatan dengan menggunakan metode *text mining* dan pemodelan *algoritma Naïve Bayes* yang ditambahkan dengan *feature selection SMOTE*, *N-Gam*, dan *Adaboost* terbukti efektif dalam hal klasifikasi pengkategorian narasi tweet "*Positive*" dan "*Negative*", hal ini didukung dengan dihasilkannya kategori tertinggi pada hasil penelitian *algoritma Naïve Bayes* yang ditambahkan dengan *feature selection SMOTE*, *N-Gram*, dan *Adaboost* didapati tingkat *Acuraccy* 90.18%, *Precission* 96.61%, *Recall* 83.43%, *AUC* 0.988%.

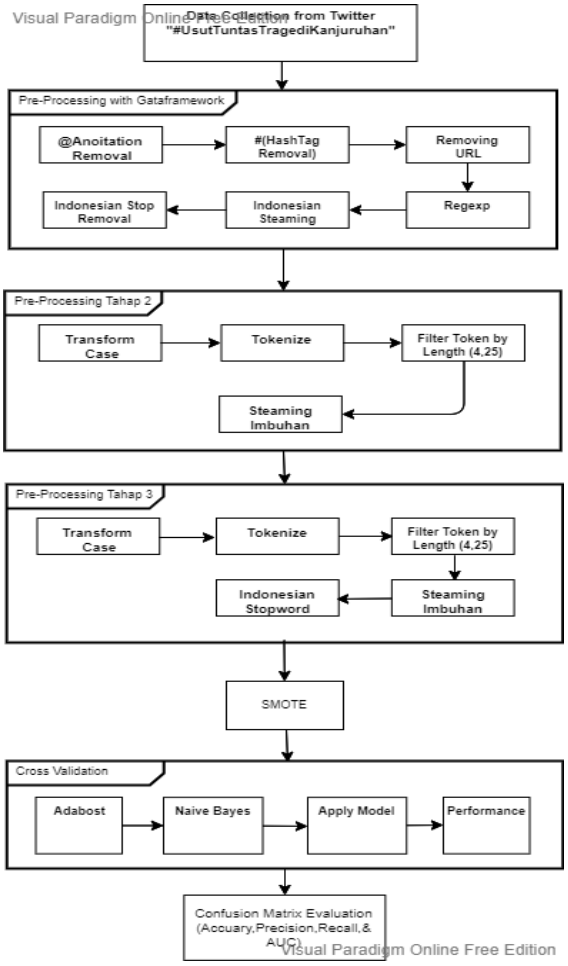
Penelitian-penelitian diatas (yang meng-gunakan metode *Naïve Bayes Classifier)* menunjukan bahwa metode *Naïve Bayes* dapat mengklasifikasikan sentimen positif dan negatif dari sebuah cuitan *twitter.* Hal tersebut didukung dengan dihasilkannya nilai *accuracy* yang cukup besar yaitu di atas 80%.

Pada penlitian ini metode penelitian dipresentasikan dalam sebuah model kerangka kerja penelitian, pada gambar 1 dengan penjelasan kerangka kerja analisis sentimen adalah sebagai berikut:

1. Data *Collection from Twitter*

Pada tahapan ini dilakukan pengumpulan atau *colleecting* data melalui aplikasi *RapidMiner* dengan menggunakan koneksi ke akun Twitter dengan mengkoneksikan Twitter *API* melalui akun Twitter *Developer*. Twitter API merupakan sekumpulan URL yang mengambil parameter. URL ini mengizinkan pengguna maengakses fitur-fitur Twitter, seperti memposting tweet atau mencari tweet yang ber-

isi suatu kata dan lain-lain M.Cindy(Buslim et al., 2018). Data Twitter dapat diakses melalui *API REST* (*Representational State Transfer*) Twitter yang telah disediakan oleh pihak Twitter dengan terlebih dahulu mengajukan permintaan kepada pihak Twitter untuk memperoleh akses data dari Twitter dengan melakukan pendaftaran sebagai akun *developer*.



Gambar 1. Kerangka Kerja Analisis Sentimen

2. *Pre-processing* Tahap 1

Data yang didapat pada proses *crawling* biasanya tidak langsung siap digunakan untuk proses pengujian, maka dari itu perlu dilakukan *pre-processing*. Tahap dilakukan untuk mengubah data yang diperoleh dari sumber data menjadi dataset agar siap diproses lebih lanjut ke tahap pengujian data. *Pre-processing* bertujuan untuk menghilangkan noise serta melakukan beberapa tahapan untuk mengubah data yang belum terstruktur menjadi tersetruktur atau mengubah teks menjadi term index yang mewakili sebuah dokumen sehingga siap diproses lebih lanjut.

3. *Pre-processing* Tahap 2

Pada tahap ini akan dilakukan operasi beberapa fungsi, metode dan operator dalam pengolahan data yaitu *read* excel, yaitu membaca data yang telah dilakukan pada *preprocessing* tahap 1, Tahapan dimulai dengan mengimport data *excel* dari *pre-processing* 1, kemudian parameter data yang dipakai adalah no, class, dan indonesian stop word. Selanjutnya memberikan attribute data mengedit role attribute, *tokenize, transform cases, filter token by length (4,25),* dan *steaming* Imbuhan.

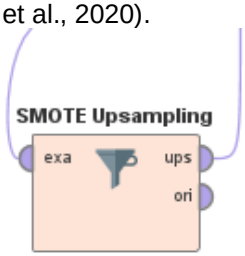
4. *Pre-processing* Tahap 3

Pada tahapan ini melanjutkan dari tahapan *pre-processing* tahap 2. Pengujian dilakukan dengan menambahkan salah satu fungsi operator yaitu filter stopwords. Tehnik menggunakan operator *filter stopword (Dictionary)* berfungsi untuk membuang kata yang memiliki informasi yang rendah dalam sebuah teks.

5. *Smote*

Setelah dilakukan *pre-processing* tahap berikutnya dilakukan *Synthetic Minority Oversampling Technique (*SMOTE) agar tidak terjadi *imbalance* data atau kondisi dimana antar kelas memiliki selisih yang signifikan, Smote perlu dilakukan untuk menghindari *imbalance* karena pemodelan dengan algoritma yang tidak memperhatikan Ketidakseimbangan data mendominasi oleh kelas mayor dan tidak memperhatikan kelas minornya.

Untuk mengatasi masalah ketidakseimbangan kelas**,** salah satu metode yang digunakan adalah tehnik *Synthetic Minority Oversampling Technique (*SMOTE yang mengubah distribusi data antara kelas mayoritas dan kelas minoritas dalam dataset untuk menyeimbangkan jumlah data di setiap kelas. (Sutoyo et al., 2020).



Gambar 2. Operator Smote

6. *Cross Validation*

Cross Validation (k-fold) adalah metode untuk mengevaluasi model atau algoritma yang bertujuan memisahkan data menjadi data *training* dan data validasi (Augustia et al., 2021).

Tahapan pemodelan data menggunakan algoritma *Naïve Bayes,* yang menghasilkan model *cross validation* ini mempunyai beberapa operator yaitu algoritma untuk training sehingga menghasilkan model dan *testing perfomance* model.

7. *Confusion Matrix*

Pada penilitan ini, hasil dari tahap pengujian akan di evaluasi menggunakan table *Confusion Matrix* yaitu, *Accuracy, Precision, Recall,* dan AUC.

Tabel 1. *Confusion Matrix*

Correct Clasification	*Classified as*	
	+	-
+	*True Positif*	*False Negatif*
-	*False Positif*	*True Negatif*

Nilai *accuracy* adalah presentase jumlah record data yang diklasifikasikan secara benar oleh sebuah algoritma dapat membuat klasifikasi setelah dilakukan pengujian pada hasil klasifikasi tersebut, Han & Kamber(dalam Saputra D, et al., 2018).

Nilai *Precision* atau dikenal juga dengan nama *confidence* merupakan proporsi jumlah kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. Sedangkan nilai dari *recall* atau *sensitivity* merupakan proporsi jumlah kasus positif yang sebenarnya yang diprediksi positif secara benar, Powers(dalam Saputra D, et al., 2018).

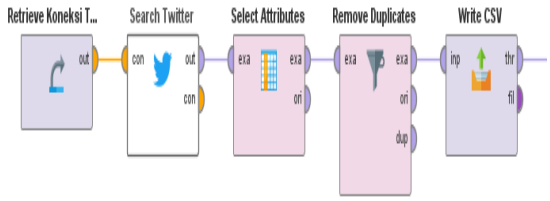
Nilai AUC menggambarkan hasil pengukuran kesesuaian model secara keseluruhan yang digunakan. Nilai AUC yang meningkat memiliki arti bahwa variable yang di teliti semakin baik dalam memprediksi kejadian, Maskoen & Purnama(dalam Augustia et al., 2021).

3. Hasil dan Pembahasan

Berdasarkan tahapan dan metode penelitian yang sudah dilakukan, maka selanjutnya akan menjelaskan hasil dan pembahasan terhadap penelitian yang dikerjakan adalah sebagai berikut:

3.1 Crawling Data Twitter

Pengumpulan data *twitter* dilakukan dengan *crawling tool Rapidminer* menggunakan operator *search twitter* dengan parameter *query "#UsutTuntasTragediKanjuruhan"*, dengan jumlah *record* 1.500. Proses *crawling* data disajikan pada gambar 3.



Gambar 3. Proses *Crawling* Data

Data yang dihasilkan dari proses *crawling* disimpan ke dalam bentuk *file excel*. Setelah itu data pada *file excel* tersebut dihilangkan data duplikatnya menggunakan *tool microsoft excel* yaitu *remove duplicate.* Penghapusan duplikat menghasilkan 1.114 data yang akan dilakukan *pre-processing.*

3.2 Labeling

Pada tahapan ini dilakukan labeling terhadap 1.114 dataset yang dilakukan oleh publik atau masyarakat dan menghasilkan data dengan klasifikasi "*Positive*" dan "*Negative*".

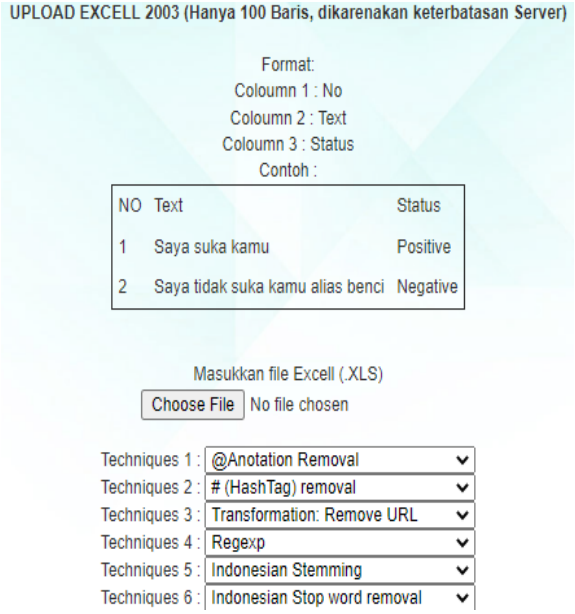
No	Text	Class
1	Mention aja hashtag #UsutTuntasTragediKanjuruhan itu ke mereka para pemain, pelatih atau siapapun itu lah. #Justiceforkanjuruhan;"1580207894024245249"	Positif
2	Sejak awal, PSI mendukung pengusutan tuntas hilangnya ratusan nyawa dalam Tragedi Kanjuruhan dan pihak-pihak yang bertanggung jawab harus diberi sanksi. #UsutTuntasTragediKanjuruhan;"1580149628821213184"	Negatif

Gambar 4. Hasil Labeling

Dari dataset yang sudah dilakukan penghapusan duplikat dipilih 600 data *twitter* untuk klasifikasi *class* sentimen Positif dan 514 data *twitter* untuk klasifikasi *class* sentimen Negatif.

3.3 *Pre-processing* Tahap 1

Dari hasil labeling terhadap dataset yang telah dikumpulkan tersebut kemudian dilakukan proses *preprocessing* tahap I dengan membagi file menjadi sebanyak 10 file, masing-masing file berisi 50 dataset dikarenakan keterbatasan server pada aplikasi *gata framework*. Pada tahapan ini *preprocessing* dan *cleansing* dengan menggunakan beberapa teknik pada *website gataframework* untuk membuat *design preprocessing* dan *cleansing* pada dataset *local*. *@Annotation Removal, Transformation Remove URL, Regexp, Indonesian Stemming, Indonesian Stop Word Removal*.

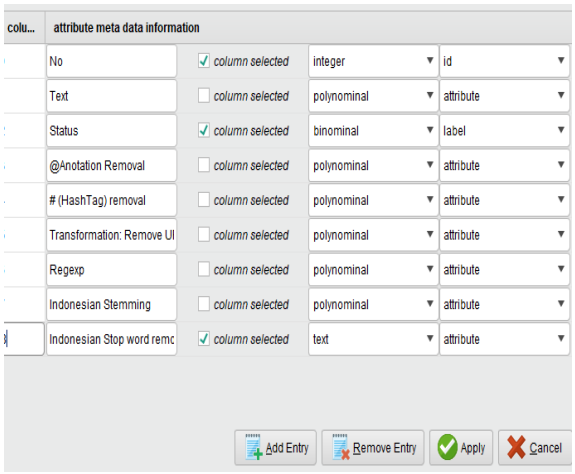


Gambar 5. *Pre-processing* Tahap 1 *with Gataframework*

Beberapa tahapan yang dilakukan dalam *pre-processing* tahap 1 yaitu:
a).*@Anotation Removal*, bertujuan untuk menghapus tanda *anotation* yang terdapat pada text *tweet*. *Anotation* merupakan salah satu noise yang tidak memiliki arti.
b).*#(Hashtag) removal,* dilakukan untuk menghapus kata kunci *hashtag.*
c).*Removing URL,* yaitu proses menghilangkan URL yang biasanya terdapat pada *text tweet.*
d). Regexp, yaitu proses menghilangkan simbol-simbol yang terdapat pada *text tweet.*
e).Indonesian steaming, adalah tahapan untuk memperkecil jumlah indeks yang berbeda dari satu sehingga sebuah kata yang memiliki *suffix* maupun *prefix* akan kembali ke bentuk dasarnya.
f). Indonesia stop word removal, yaitu proses menghilangkan kata penghubung dalam bahasa indonesia misalnya adalah “di” dan “yang”. Proses stop word dilakukan dengan cara mengumpulkan kata yang paling sering muncul di corpus.

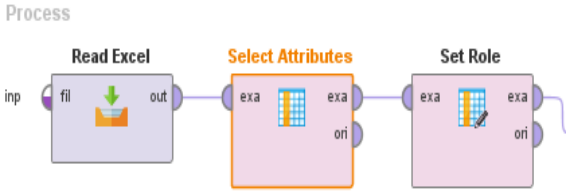
3.4 *Pre-processing* Tahap 2

Pada tahapan ini dilakukan pengujian dilakukan menggunakan *tool Rapidminer.* Tahapan *pre-processing* tahap 2 dilakukan beberapa proses uji coba pada dataset sehingga menghasilkan data yang lebih akurat untuk pengimplementasian *machine learing*.



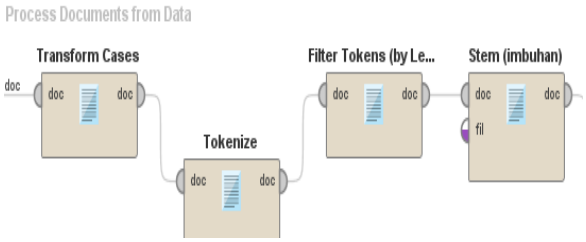
Gambar 6. *Import Data*

Pada tahapan ini dilakukan dengan cara melakukan *import* data yang sudah di *pre-processing* pada tahap 1 ke aplikasi *Rapid Miner* dengan menggunakan operator *read* excel



Gambar 7. Operator *Attribute* dan *Set Role*

Tahap ini dilakukan untuk membuang atribut yang tidak diperlukan. Kemudian melakukan perubahan pada parameter role menjadi status. Hal ini akan memungkinkan kita memilih sub kumpulan kolom untuk disimpan dalam data.



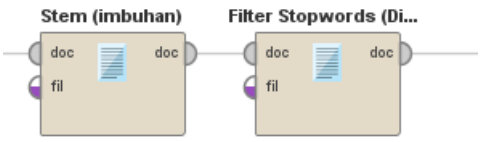
Gambar 8. *Pre-processing Tahap* 2

Tahap selanjutnya yang meliputi tahapan *pre-processing* tahap 2 sebagai berikut:
a).*Tokenize*, proses *tokenize* adalah proses pemotongan string input berdasarkan setiap kata komposisinya. Tahapan ini dilakukan untuk memisahkan kata per kata dari suatu teks kalimat.
b). *Transform Cases*, tahapan ini berfungsi untuk membuat data tweet menjadi huruf kecil

semua, misalnya dari huruf kapital menjadi huruf kecil biasa.
c). *Filter Token by Length (4,25),* tahapan ini bertujuan agar mendapat kata-kata yang panjangnya antara 4 sampai 25 karakter.
d). *Steaming Imbuhan,* tahapan ini bertujuan membandingkan data distributor yang dihasilkan dengan menghilangkan imbuhan data sepert "di", "an", "nya" dan lain-lain.

3.5 *Pre-processing* Tahap 3

Pada tahapan ini dilakukan pembobotan nilai pada dataset distribusi yang dihasilkan, pembobotan ini dilakukan untuk mengetahui kata dalam text yang tidak relevan atau kolerasi dengan *object* yang diteliti.



Gambar 9. *Pre-processing* Tahap 3

Pada *pre-processing* tahap 3 dilakukan penambahan operator filter stopword. Data text stopword di import dari data hasil nilai pembobotan masing-masing kata.

Tabel 2. Hasil Stopword

Text	Nilai Bobot
borgol	0.1862
distrik	0.1862
republik	0.1862
pegang	0.1862
…..	…..
penyidik	1.0978

Tabel 2 menjelaskan hasil dari pemberian nilai bobot masing-masing kata yang didapat dari *simple distribrution* data. Semakin kecil nilai pembobotannya, maka kata tersebut tidak ada kolerasinya dengan object penelitian.

3.6 Evaluasi

Dari hasil klasifikasi keseluruhan yang sudah didapat, maka pada proses ini dilakukan untuk menguji hasil klasifikasi dengan menggunakan metode *confusion matrix* dengan sejumlah data yang sudah diuji. Pada tahap ini penulis mencari nilai *accuracy, precision, recall, dan auc.*

Tabel 3. Hasil *Accuracy* Algoritma NB

accuracy: 77.67% +/- 6.37% (micro average: 77.67%)			
	true Positif	true Negatif	class precision
pred. Positif	461	129	78.14%
pred. Negatif	139	471	77.21%
class recall	76.83%	78.50%	

Pada tabel 3 menunjukkan bahwa nilai *accuracy* sebesar 77,67% dengan toleransi kesalahan sebesar 6,37%, dengan nilai *true negatif* 471 *records* dan *true positif* 461 *records*.

Tabel 4. Hasil *Precision* Algoritma NB

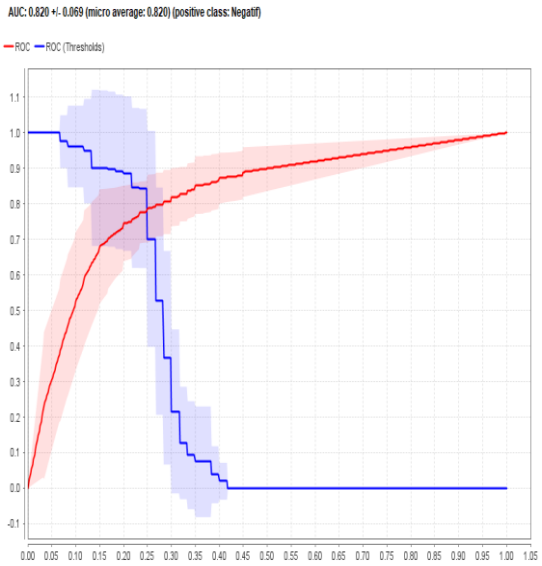
precision: 77.19% +/- 5.93% (micro average: 77.21%) (positive class: Negatif)			
	true Positif	true Negatif	class precision
pred. Positif	461	129	78.14%
pred. Negatif	139	471	77.21%
class recall	76.83%	78.50%	

Pada tabel 4 menunjukkan bahwa nilai *precision* sebesar 77,19% dengan toleransi kesalahan sebesar 6,93%, dengan nilai *true negatif* 471 *records* dan *true positif* 461 *records*.

Tabel 5. Hasil *Recall* Algoritma NB

recall: 78.50% +/- 8.73% (micro average: 78.50%) (positive class: Negatif)			
	true Positif	true Negatif	class precision
pred. Positif	461	129	78.14%
pred. Negatif	139	471	77.21%
class recall	76.83%	78.50%	

Pada tabel 5 menunjukkan bahwa nilai *accuracy* sebesar 78,50% dengan toleransi kesalahan sebesar 8,73%, dengan nilai *true negatif* 471 *records* dan *true positif* 461 *records*.



Gambar 8. Grafik ROC Algoritma *Naive Bayes*

Berdasarkan hasil pengujian *performance* menghasilkan *Kurva* ROC seperti pada gambar 8 dan *nilai* AUC yang didapatkan sebesar 0,820 (*good classification*).

4. Kesimpulan

Pengujian terhadap data hasil *crawling* dari media sosial *twitter* dengan *query #UsutTuntasTragediKanjuruhan* dengan algoritma *Naïve Bayes* telah berhasil dilakukan. Pendekatan menggunakan metode *text mining* dan algoritma *Naïve Bayes* terbukti efektif untuk mengklasifikasikan *perspektif tweet* Positif dan negatif, hal ini didukung dengan dihasilkannya nilai *accuracy 77,67%, precision 77,19%, recall 78,50%, dan* AUC yang didapat sebesar 0,820 yang dievaluasi dengan *confusion matrix*.

Referensi

Astiningrum Mungki, Hani'ah Mamluatul, & YP. Yanuar Rahmat. (2020). *Analisis Sentimen Tentang Opini Terhadap Performa Timnas Sepak Bola Indonesia Pada Twitter*. 2020.

Augustia, A. E., Taufan, R., Alkhalifi, Y., & Gata, W. (2021). Analisis Sentimen Omnibus Law Pada Twitter Dengan Algoritma Klasifikasi Berbasis Particle Swarm Optimization. *Paradigma - Jurnal Komputer Dan Informatika*, *23*(2). https://doi.org/10.31294/p.v23i2.10430

Buntoro, G. A. (2019). Analisis Sentimen Calon Gubernur Jawa Timur 2018 Dengan Metode Naive Bayes Classifier. In *Journal Of Informatics Pelita Nusantara* (Vol. 4, Issue 1).

Buslim, N., Busman, B., Sinatrya, N. S., & Kania, T. S. (2018). Analisa Sentimen Menggunakan Data Twitter, Flume, Hive Pada Hadoop dan Java Untuk Deteksi Kemacetan di Jakarta. *Jurnal Online Informatika*, *3*(1), 1. https://doi.org/10.15575/join.v3i1.141

Hakim, B. (2021). Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning. *JBASE - Journal of Business and Audit Information Systems*, *4*(2). https://doi.org/10.30813/jbase.v4i2.3000

Hartati, Hermawan Deni, Akshanal M, Wahyudi Zailani, Ariyanto Angga, & Saputra D. (n.d.). *Optimasi Analisis Sentimen Pada Twitter Olshop Tokopedia Menggunakan Textmining Dengan Algoritma Naïve Bayes & Adaboost*.

Hendika, F., & Nuraeni, D. (2020). Globalisasi Hooliganisme terhadap Suporter Sepak Bola di Indonesia. In *Jurnal Hubungan Internasional □ Tahun XIII* (Issue 1).

Imamah, I., & Siddiqi, A. (2019). Penerapan Teorema Bayes untuk Mendiagnosa Penyakit Telinga Hidung Tenggorokan (THT). *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *18*(2), 268–275. https://doi.org/10.30812/matrik.v18i2.398

Nurmalasari, M. D., Kusrini, K., & Sudarmawan, S. (2021). Komparasi Algoritma Naive Bayes dan K-Nearest Neighbor untuk Membangun Pengetahuan Diagnosa Penyakit Diabetes. *Jurnal Komtika (Komputasi Dan Informatika)*, *5*(1), 52–59. https://doi.org/10.31603/komtika.v5i1.5140

Patel, R., & Passi, K. (2020). Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning. *IoT*, *1*(2), 218–239. https://doi.org/10.3390/iot1020014

Prajamukti, R., & Mega Santoni, M. (2021). *Klasifikasi Dan Analisis Sentimen Pada Data Twitter Menggunakan Algoritma Naïve Bayes (Studi Kasus: Timnas Indonesia Senior, U-23, Dan U-19)*. https://t.co/SAbaU6Prrz

Pratama Putra, A., Pratama, Y., Kharisma Krisnadi, E., Purnamasari, I., & Dwi Saputra, D. (2022). Text Mining untuk Sentimen Analisis dengan Metode Naïve Bayes, SMOTE, N-Gram dan AdaBoost Pada Twitter CommuterLine. In *Jurnal Sains Komputer & Informatika (J-SAKTI* (Vol. 6, Issue 2).

Saputra, D. D., Pratama, B., Akbar, Y., & Gata, W. (2018). Penerapan Text Mining Untuk Assingment Complaint Handling Customer Terhadap Divisi Terkait Menggunakan Metode Decission Tree Algoritma C4.5 (Studi Case : Pt. XI Axiata, Tbk) Selection

and peer-review under responsibility of The 11th STIKOM CKI on SPOT. *CKI On SPOT*, *11*(2).

Susana, H., & Suarna, N. (2022). Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet. *Jurnal Sistem Informasi Dan Teknologi Informasi)*, *4*(1), 1–8.

Sutoyo, E., Asri Fadlurrahman, M., Telekomunikasi Jl Terusan Buah Batu, J., Dayeuhkolot, K., Bandung, K., & Barat, J. (2020). *JEPIN (Jurnal Edukasi dan Penelitian Informatika) Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network*.