e-ISSN 2962-6129
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 Agustus 2022

KLASIFIKASI DAN ANALISIS SENTIMEN PADA DATA TWITTER MENGGUNAKAN ALGORITMA NAÏVE BAYES. (STUDI KASUS: PEKAN OLAHRGA NASIONAL XX 2021)

Krisna Jonathan Sitorus[1], Anita Muliawati[2], Sarika[3]
Informatika / Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Daerah Khusus Ibukota Jakarta 12450
krisnajs@upnvj.ac.id[1], anitamuliawati2017prodi@gmail.com[2], sarika.afrizal@upnvj.ac.id[3]

Abstrak. Banyak sekali pengguna Twitter mencurahkan pendapatkan melalui tweet-tweet yang mereka kirimkan pada media sosial tersebut, khususnya mengenai Pekan Olahraga Nasional (PON) XX 2021. Banyak sekali tweet yang bersifat mendukung, tetapi tidak jarang juga ada tweet yang bersifat keluhan mengenai penyelenggaraan PON XX tersebut. Dari masalah tersebut, dilaksanakanlah penelitian mengenai analisis sentimen pada data twitter yang berkaitan dengan PON XX dan mempergunakan metode Naïve Bayes. Jumlah data yang digunakan sebanyak 218 data tweet dan belum terlabelkan. Lalu data dilakukan pemberian label dan masuk tahapan text processing, selanjutnya data diberi bobot pada tiap kata dengan Term Frequency– Inverse Document Frequency (TF-IDF) yang akan kedepannya kata tersebut akan dijadikan sebagai fitur. Karenak ketidakseimbangan data, digunakanlah metode Synthetic Minority Oversampling Technique (SMOTE) guna melaksanakan penyeimbangan terhadap datanya. Tahapan selanjutnya dilaksanakan pembagian data yang besarannya yakni 80% 20% dan diklasifikasikan dengan metode Naive Bayes. Hasil yang diperolehkan dari pelaksanaan penelitiannya tersebut ialah diperoleh bahwa data uji memperolehkan accuracy yang besaran persentasenya yakni 99%, precision dengan besaran persentasenya 100%, recall dengan besaran persentasenya 98%.
Kata Kunci: Analisis Sentimen, Klasifikasi, Twitter, *Naïve Bayes*, Pekan Olahraga Nasional XX 2021

1 Pendahuluan

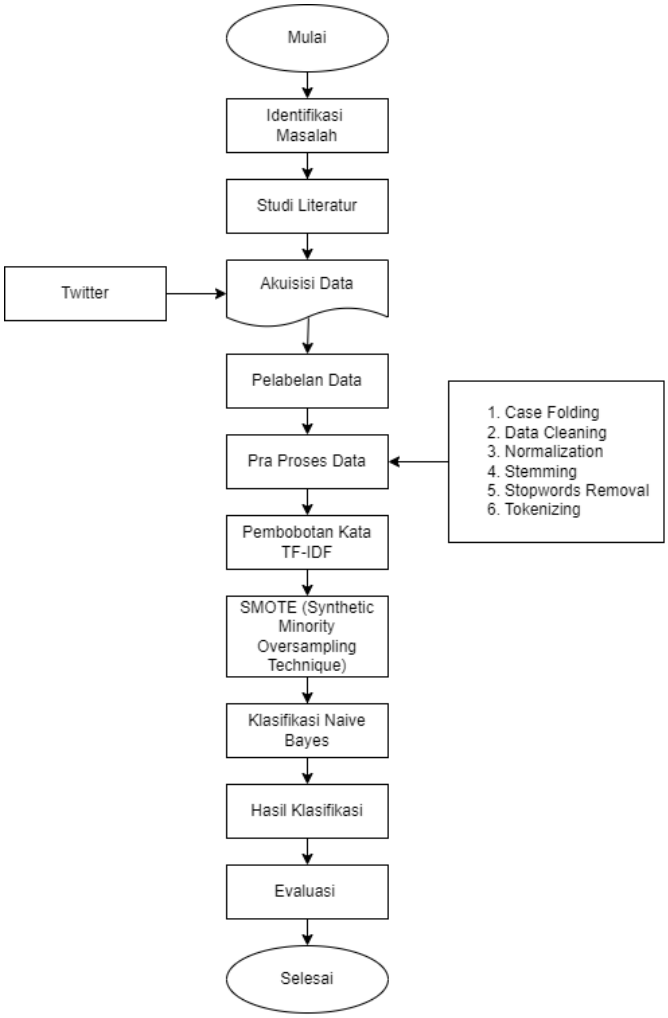
Di masa transformasi digital yang sedang terjadi sekarang, tidak bisa dihindari bahwasanya media sosial telah menjadi elemen untuk bersosialisasi dalam kehidupan masyarakat. Banyak sekali fungsi media sosial, bukan hanya menjadi media berkomunikasi, namun media sosial juga termasuk ke dalam salah satu sarana untuk menampungkan seluruh opini ataupun pendapat yang dipunyai oleh masyarakat. Dari sekian banyaknya media sosial yang menjadi tempat masyarakat bertukar pendapat, Twitter termasuk ke dalam salah satu media sosial yang kerap dimanfaatkan untuk menampung seluruh pendapat yang telah disebutkan sebelumnya. Selain sering memanfaatkan jejaring sosial Twitter untuk bertukar opini ataupun pendapat, Twitter juga menjadi tempat untuk masyarakat memperolehkan berita terkini mengenai apa yang tengah dialami masyarakat di lingkungan sekitarnya.

Pada akhir tahun 2021 tepatnya di tanggal 2 sampai 15 Oktober 2021 sedang berlangsung ajang perlombaan olahraga yang diadakan secara rutin tiap 4 tahun sekali yang dinamakan Pekan Olahraga Nasional ke-XX. Banyak dari pengguna media sosial Twitter yang membahaskan perihal ajang perlombaan yang telah disebutkan sebelumnya, Berbagai jenis pendapat terkait dengan Pekan Olahraga Nasional XX tahun 2021 ini bisa diperhatikan pada tweet yang tertera di media sosial Twitter

2 Metodologi Penelitian

2.1 Tahapan Penelitian

Dalam melakukan penelitian, ada beberapa prosedur yang dilakukan sebagai berikut:



Gambar 1. *Tahap Penelitian*

2.2 Identifikasi Masalah

Identifikasi masalah termasuk ke dalam suatu tahapan untuk memperjelaskan permasalahan yang hendak diangkatkan pada kajian ini yakni permasalahan dalam mengklasifikasikan tweet yang mengandung opini tentang Pekan Olahraga Nasional XX tahun 2021 dengan metode klasifikasi dan algoritma Naïve Bayes.

2.3 Studi Literatur

Studi literatur dimanfaatkan sebagai acuan dan dasar pengetahuan penulis dalam proses analisis sentimen dengan mengumpulkan jurnal- jurnal terkait mengenai masalah yang diangkat penulis, yakni mengenai analisis sentimen, metode klasifikasi, text mining, serta algoritma Naïve Bayes sebagai sumber pustaka. Sumber pustaka yang dijadikan dasar pengetahuan oleh penulis seperti website, jurnal, literatur, serta e-book yang berhubungan dengan penelitiannya. Setelah melaksanakan hal yang telah disebutkan sebelumnya studi pustaka diharapkan bisa membantu penulis pada kajian ini .

2.4 Pengumpulan Data

Proses pengumpulan data dilaksanakan untuk mencari dan memperolehkan data yang diperlukan untuk kajian ini, data yang akan diambil dan dimanfaatkan pada kajian ini ialah tweet pengguna twitter yang mention Pekan Olahraga Nasional bulan Oktober sampai November tahun 2021 yang akan digunakan sebagai data latih serta data ujinya. Data tweet diperolehkan melalui proses pemanfaatan crawling dari API (Application Programming Interface) yang telah twitter sediakan.

Setelah data tweet berhasil didapatkan, penulis melaksanakan labeling terhadap data tweet dengan memanfaatkan 3 anatator untuk melakukan penentuan terhadap sentimen negatif serta positif dari data tweet yang sudah didapatkan.

2.5 Pelabelan Data

Pelabelan data dikerjakan secara manual oleh 3 orang penilai ke dalam 2 kategori, kelas berlabel positif dan negatif. Berikut pada Tabel 1 adalah contoh hasil pelabelan secara manual oleh 3 orang penilai :

Tabel 1. Contoh Pelabelan Data *Tweet*:

Data *Tweet*	Penilai 1	Penilai 2	Penilai 3	Label Akhir
Adanya penyelenggaraan Pekan Olahraga Nasional XX Papua, Mengakibatkan penggunaan QRIS di Papua mengalami peningkatan drastis. #PapuaBangkitPapuaMaju #PapuaIndonesia https://t.co/dsH2pZdXvd	Positif	Positif	Positif	Positif
@Ar07Pangeran @caesar_emil Goblok nih org, mantan gubernur Riau Rusli Zainal ditangkap KPK krn korupsi dana pekan olahraga nasional (PON).	Negatif	Negatif	Negatif	Negatif
Tunggakan pembayaran penyelenggaraan Pekan Olahraga Nasional (PON) XX masih terus mengemuka. Semoga cepat terselesaikan.PapuaMaju BerkatOtsus #PapuaIndonesia https://t.co/n0QXn5h0kg	Negatif	Negatif	Negatif	Negatif

Dari hasil penilaian label tersebut masih terdapat perbedaan pendapat antara masing-masing penilai dalam mengkategorikan label tweet, maka diperlukan hasil yang menunjukan persetujuan antar penilai dengan perhitungan *Kappa Value*. Berikut hasil perhitungan *kappa value* untuk data tweet yang sudah diberi nilai oleh penilai:

Rumus Persamaan (1) dengan *Kappa Value*

$$Kappa = \frac{P_0 - P_e}{1 - P_e} \quad \textbf{(1)}$$

Keterangan:
Kappa : Koefisien dari nilai kesepakatan dimana 0 untuk persetujuan secara kebetulan tidak satu penilaian, dan 1 untuk persetujuan yang mutlak.
P_0 : Proporsi frekuensi penilai yang penilaian sama
P_e : Peluang kesepakatan antar penilai yang penilaiannya berbeda.

Dimana persamaan (2) dengan P_e:

$$P_e = P(positif)^2 + P(negatif)^2 \quad \textbf{(2)}$$

Hasil kappa value yang digunakan dapat dikatakan objektif untuk menilai sebuah kesepakatan[1]. Hasil kesepakatan dari 3 orang penilai dengan pengukuran kappa value dapat dikategorikan pada tabel 2:

Tabel 2. Nilai Kesepakatan *Kappa Value*:

Kesepakatan	Nilai k
Rendah (*poor*)	k < 0.00
Kurang (*deficientt*)	0.00 - 0.20
Lumayan (*fair*)	0.21 – 0.40
Cukup (*moderate*)	0.41 – 0.60

Baik (*good*)	0.61 – 0.80
Sangat Baik (*Very Good*)	k > 0.81

2.6 Praproses Data

Dalam text mining, tahapan pra proses ataupun disebutkan juga Text Preprocessing termasuk ke dalam tahapan awal dalam memproses teks guna mempersiapkan serta membersihkan data dari data yang tidak berstruktur. Tahap preprocessing dilaksanakan sebelum tahap modelling ataupun pemodelan data, dimana terjadi proses pembentukan dan pembersihan informasi sehingga bisa diolah pada tahap berikutnya[2]. Sebelum data tweet diklasifikasi, perlu dilakukan pra proses terlebih dahulu karena data belum berstruktur dan memiliki banyak noise. Tahapan dari pra proses ini terdiri dari case folding, pembersihan data, normalisasi bahasa, stemming, stopword removal, dan tokenisasi.

2.5.1 Case Folding

Tahap pertama pada pra proses data yaitu Case Folding, Case Folding ialah rangkaian proses menyeragamkan ataupun menyamakan seluruh teks [3], dimana seluruh data tweet yang didapat dari yang memiliki huruf kapital (uppercase) dikonversi menjadi huruf non-kapital (lowercase), bertujuan untuk mencegah terjadinya case sensitive.

2.5.2 Pembersihan Data

Data Cleaning ialah tahapan untuk menghilangkan noise yang kurang penting dalam [3]. Noise yang dimaksud seperti menghapus tags, angka, tanda baca, URL, dan lain-lain.

2.5.3 Normalisasi Bahasa

Tahap Spelling Normalization ataupun normalisasi termasuk ke dalam tahapan untuk melakukan perbaikan terhadap berbagai kata-kata yang terdapat pada teks dokumen yang salah eja, disingkatkan, maupun kata-kata yang tidak sesuai dengan KBBI. Spelling Normalization dilaksanakan untuk memperkecil dimensi kata yang memiliki ejaan berbeda namun mengandung makna yang sama [4].

2.5.4 Stemming

Stemming termasuk ke dalam proses pencarian kata dasar dengan menghilangkan imbuhan yang terdapat pada kata tersebut sehingga tersisa kata dasar [3].

2.5.5 Stopwords Removal

Stopword removal ialah proses menghilangkan kata yang tidak terkait ataupun tanpa makna. Stopword ialah kata-kata yang acap kali terlihat namun tidak mempengaruhi sentimen, maka kata tersebut diklasifikasikan sebagai noise untuk dihilangkan [5].

2.5.6 Tokenisasi

Tokenizing adalah proses sebuah kalimat atau paragraf dipecah menjadi sebuah kata. [6].

2.7 Pembobotan Term (TF-IDF)

Setelah pra proses data dilakukan, maka perhitungan bobot term dengan mengubah kata-kata pada data tweet menjadi sebuah angka sehingga kata/term dapat dikenali sebagai fitur untuk klasifikasi nanti. Term Frequency ialah frekuensi satuan term yang terdiri dari kata, frasa ataupun elemen dari indexing pada naskah. Semakin besar frekuensi kemunculan term, semakin besar bobotnya dan *Inverse Document Frequency* digunakan dalam mengurangi bobot satuan *term* di berbagai dokumen yang acap kali terlihat. Situasi ini dikarenakan *term* tersbut memiliki frekuensi kemunculan terbanyak di berbagai dokumen, akan menjadi *term* jamak yang menyebabkan nilainya tidak terlalu penting (Praptiwi, 2018). Metode perhitungan Term Frequency (TF) dan Inverse Document Frequency (IDF) menggunakan rumus:

$$W_{t,d} = tf_{t,d} \times idf_t = tf_{t,d} \times \log\frac{N}{df_t} \quad \textbf{(3)}$$

Keterangan :

$W_{t,d}$	: Bobot TF-IDF
$tf_{t,d}$	: Jumlah frekuensi kata
idf_t	: Jumlah inverse frekuensi dokumen tiap kata
df_t	: Jumlah frekuensi dokumen tiap kata
N	: Jumlah total dokumen

2.8 Synthetic Minority Oversampling Technique

SMOTE adalah sebuah algoritma yang berfungsi untuk melakukan oversample data yang termasuk ke dalam kelas minor. SMOTE melakukan oversampling dengan cara mengambil k data dari k-NN untuk setiap data di kelas minor [7]. Dikarenakan pada saat pelabelan didapati bahwa data tidak seimbang maka kelas minority akan ditambahkan jumlahnya dengan tujuan untuk menyeimbangkan data yang akan diolah.

2.9 Klasifikasi dengan Naïve Bayes

Klasifikasi menggunakan Support Vector Machine akan dilakukan dengan kernel linear dan radial basis function (RBF). Data akan dibagi menjadi data latih sebesar 80% dan data uji sebesar 20%. Proses klasifikasi menggunakan algoritma Support Vector Machine dihitung dengan :

Rumus Persamaan (6) Probabilitas Kata

$$P(Wi \mid c) = (count(Wi, c) + 1)/(|c| + |V|) \quad \textbf{(6)}$$

Keterangan :

P(Wi|c)	*: Probabilitas kata Wi pada kelas c*
Count(Wi,c)	*: jumlah kemunculan kata Wi pada kelas c.*
|c|	*: jumlah semua kata pada kelas c*
|V|	*: jumlah keseluruhan kata*

Rumus Persamaan (7) Peluang Kemunculan Dokumen

$$P\ (c) = \frac{|doc\ c|}{|document|} \quad \textbf{(7)}$$

Keterangan :

P(c)	*:Peluang kemunculan suatu dokumen yang memiliki kategori j.*
doc c	*: Jumlah dari dokumen untuk tiap kategori j*
|document|	*: Jumlah dokumen dari setiap kategori.*

Rumus Persamaan (8) Naïve Bayes

$$c_{MAP} = \underset{c \in V}{\arg\max}\ P(c) \prod_i P(W_i|c) \quad \textbf{(8)}$$

Keterangan :

P(c)	: Peluang kemunculan suatu dokumen yang memiliki class c.
P(W_i |c)	: Peluang kemunculan Wi pada class c.

2.10 Evaluasi

Pada tahap akhir penelitian, model klasifikasi akan melewati tahap evaluasi menggunakan metode confusion matrix yang sudah dijelaskan pada tabel untuk menganalisis hasil performanya. Berikut rumus yang digunakan untuk menghitung evaluasi pada penelitian ini:

Persamaan rumus (9) untuk menghitung nilai akurasi:

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} x\ 100\% \quad \textbf{(10)}$$

Persamaan rumus (10) untuk menghitung nilai *recall* (*sensitivity*):

$$Recall = \frac{TP}{TP+FN} x\ 100\% \quad \textbf{(11)}$$

Persamaan rumus (11) untuk menghitung nilai presisi:

$$Presisi = \frac{TP}{TP+FP} x\ 100\% \quad \textbf{(12)}$$

Persamaan rumus (12) untuk menghitung nilai *Specificity*:

$$Specificity = \frac{TN}{TN+FP} x\ 100\% \quad \textbf{(13)}$$

2.11 Visualisasi

Hasil sentimen pengguna twitter terhadap PON XX 2021 berdasarkan data tweet yang didapat akan divisualisasi dalam bentuk wordcloud yang berisi frekuensi kemunculan kata terbanyak pada masing-masing sentimen positif ataupun negatif.

3 Hasil Pembahasan

Penelitian ini menggunakan data tweet yang diperoleh dengan mengaplikasikan metode crawling dengan bahasa R memanfaatkan API (Application Programming Interface) full archieve yang disediakan oleh twitter untuk dapat mengakses tweet lama. Data hasil crawling disimpan dalam bentuk file .xlsx yang berisikan record tweet. Data tweet yang diambil mulai dari tanggal 01 Agustus 2021 sampai 31 Desember 2021 dengan kata kunci "pekan olahraga nasional" dan dari hasil crawling data tersebut melalui proses penyaringan sehingga terkumpul sebanyak 218 data tweet yang berisikan sentimen opini publik.

Pelabelan data dikerjakan secara manual oleh 3 orang penilai ke dalam 2 kategori, label positif dan negatif. Dari hasil penilaian label tersebut masih terdapat perbedaan pendapat antara masing-masing penilai dalam mengkategorikan label tweet yang hasil yang perlu perhitungan Kappa Value untuk menunjukan tingkat persetujuan antar penilai. Setelah data tweet dilakukan voting hasil penilaian labelnya, hasil dari pelabelan data tweet sebanyak 218 tweet adalah 199 tweet berlabel positif dan 19 tweet berlabel negatif.

Setelah data yang diperoleh diberi label, dilakukan beberapa tahapan praproses data, yaitu case folding, pembersihan data, normalisasi bahasa, stemming, stopwords removal, dan tokenisasi. Hasil dari praproses data dapat dilihat pada Tabel 3

Tabel 3. Hasil praproses data

Data *tweet*
['pagelaran', 'pekan', 'paralimpik', 'nasional', 'jayapura', 'hasil', 'sukses', 'laksana', 'tutup', 'hebat']
['selenggara', 'pekan', 'olahraga', 'nasional', 'xx', 'papua', 'akibat', 'qris', 'papua', 'alami', 'tingkat', 'drastis']

Kemudian data diatas dilakukan perhitungan pembobotan kata dengan variabel kata atau term yang diperoleh dari hasil praproses data sebanyak 732 kata (term) dari jumlah data tweet sebanyak 218 tweet menggunakan metode perkalian Term Frequency - Inverse Document Frequency (TF-IDF) seperti pada Tabel 4 :

Tabel 4. Perhitungan Pembobotan Kata (TF-IDF)

Term	Dokumen		DF	IDF	TF-IDF	
	D1	D2			D1	D2
akibat	0	1	1	0,602	0,000	0,602
alami	0	1	1	0,602	0,000	0,602
drastis	0	1	1	0,602	0,000	0,602
hasil	1	0	1	0,602	0,602	0,000
hebat	1	0	1	0,602	0,602	0,000
jayapura	1	0	1	0,602	0,602	0,000
laksana	1	0	1	0,602	0,602	0,000
nasional	1	1	2	0,301	0,301	0,301
olahraga	0	1	1	0,301	0,000	0,301
pagelaran	1	0	1	0,602	0, 602	0,000
papua	0	2	1	0,602	0,000	1,204

paralimpik	1	0	1	0,602	0,000	0,301
pekan	1	1	2	0,301	0,000	0,301
qris	0	1	1	0,602	0,000	0,602
selenggara	0	1	1	0,602	0,000	0,602
sukses	1	0	1	0,602	0,602	0,000
tingkat	0	1	1	0,602	0,000	0,602
tutup	1	0	1	0,602	0,602	0,000
xx	0	1	1	0,602	0,000	0,602

Data tweet yang sudah diberi bobot dengan TF-IDF, kelas yang minor akan disamakan jumlahnya menggunakan algoritma SMOTE agar model mendapatkan hasil yang lebih baik, hasil dari penyeimbangan data dapat dilihat pada Tabel 5 berikut:

Tabel 5. Penyeimbangan Data

	Sebelum		Sesudah	
	Positif	Negatif	Positif	Negatif
Total Data	199	19	199	199

Selanjutnya data dibagi menjadi data latih (training) yang diambil 80% secara acak, sedangkan data uji (testing) diambil 20% yang berisi hasil sisa dari proses pembagian data latih (training). Perbandingan pembagian data latih (training) dan data uji (testing) secara bebas dan imbang seperti pada Tabel 6 berikut :

Tabel 6. Pembagian Data

	Total
Data Latih (Training)	318
Data Uji (Testing)	80
Total	398

Setelah itu dilakukan pengklasifikasian menggunakan metode Naïve Bayes, sebagai contoh data uji yang diambil adalah seperti pada Tabel 7 berikut :

Tabel 7. Contoh Data Uji

Data *tweet*
['selenggara', 'pekan', 'olahraga', 'nasional', 'pekan', 'olahraga', 'nasional', 'xx', 'papua', 'oktober', 'sisa', 'organisasi', 'papua', 'merdeka', 'hambat', 'maju']

Dari hasil percobaan dapat disimpulkan bahwa data uji tersebut masuk kedalam kelas negatif karena nilai probabilitas tertinggi pada kelas negatif dengan nilai 8.42×10^{-21}. Lalu untuk mengevaluasi kinerja model menggunakan Confusion Matrix yang dapat dilihat pada Tabel 8 berikut:

Tabel 8. Confusion Matrix dari Model Klasifikasi SVM

Aktual	**Prediksi**	
	Positif	**Negatif**
Positif	41 (TP)	0 (FN)
Negatif	1 (FP)	38 (TN)

Maka dari Tabel 8 Confusion Matrix dapat dihitung hasil evaluasi model Naïve Bayes menggunakan rumus (10), (11), (12), dan (13) sebagai berikut :

Akurasi	$= \frac{TP+TN}{TP+FN+FP+TN} x\ 100\%$	$= \frac{41+38}{41+0+1+38} x\ 100\%$	= 99%
Recall	$= \frac{TP}{TP+FN} x\ 100\%$	$= \frac{41}{41+1} x\ 100\%$	= 98%
Presisi	$= \frac{TP}{TP+FP} x\ 100\%$	$= \frac{41}{41+0} x\ 100\%$	= 100%
Specificity	$= \frac{TN}{TN+FP} x\ 100\%$	$= \frac{38}{39+0} x\ 100\%$	= 100%

Kemudian tahapan terakhir yaitu visualisasi berdasarkan sentimen data tweet dengan label positif dan negatif secara manual, dengan tujuan menggambarkan hasil penelitian sentimen terhadap tweet mengenai Pekan Olahraga Nasional. Visualisasi akan digambarkan oleh wordcloud.

Gambar 2. Wordcloud Sentimen Positif

Berdasarkan Gambar 2 tersebut, kata yang kemunculannya sering pada sentimen positif adalah ‘atlet’, ‘hebat’, ‘selenggara’, dan lain-lain.

Gambar 3. Wordcloud Sentimen Negatif

Berdasarkan Gambar 4 , kata yang kemunculannya sering pada sentimen negatif adalah ‘otonomi’, ‘korupsi’, ‘usut’, ‘dana’, dan lain-lain.

4 Kesimpulan dan Saran

4.1 Kesimpulan

Berdasarkan hasil analisis dan pembahasan pada penelitian yang telah dilaksanakan, dapat ditarik kesimpulan bahwa data yang dimanfaatkan termasuk ke dalam data tweet dari media sosial twitter mengenai Pekan Olahraga Nasional XX yang di crawling dari tanggal 01 Agustus 2021 hingga 31 Desember 2021 sebanyak 1000 ulasan. Setelah itu data di lakukan pembersihan dengan menghilangkan data duplicate sehingga memperolehkan hasil 218 data. Data dilaksanakan pelabelan dengan cara manual dan didapatkan jumlah kelas masing-masing sebanyak 199

positif dan 19 negatif. Kemudian data dilaksanakan pembersihan terlebih dahulu di praproses sebelum dilaksanakan pembobotan, kemudian data yang sudah bersih diberikan bobot setiap kata dengan Term Frequency-Invers Document Frequency (TF-IDF) yang nantinya akan dijadikan sebagai fitur. Sebelum fitur tersebut akan dilaksanakan pembagian data menjadi data latih dan data uji, dilaksanakan tahapan penyeimbangan data agar data berlabel positif dan negatif simbang untuk diolah dengan metode Synthetic Minority Oversampling Technique (SMOTE). Setelah fitur simbang, lalu fitur tersebut dibagi dua menjadi data training dan data testing untuk membentuk suatu model dengan memanfaatkan metode Naïve Bayes lalu dijalankan pengujian model yang terbentuk dengan metode Naïve Bayes dilaksanakan dengan membandingkan hasil dari model tersebut dengan data uji sebanyak 20% dari keseluruhan data. Selanjutnya dilaksanakan evaluasi tersebut didapatkan hasil akurasi yang besarannya yakni 99%, dengan nilai presisi 100%, recall 98%, dan specitifity 100%.

4.2 Saran

Terdapat beberapa saran dari hasil penelitian untuk pengembangan penelitian selanjutnya agar menjadi lebih baik, yaitu :

- Diharapkan untuk penelitian berikutnya untuk meningkatkan pada proses pre- processing dan pemodelan data yang lebih baik sehingga bisa menganalisis dengan baik dan evaluasi model yang sangat baik.
- Pada tahap normalisasi data diharapkan untuk menambahkan lagi kosa kata pada kamus untuk menghindari adanya singkatan kata dan slang word sehingga makna kata dapat dimengerti dan formal.
- Label yang dimanfaatkan tidak hanya positif dan negatif saja, tetapi menambahkan label netral.
- Penelitian selanjutnya bisa memanfaatkan algoritma klasifikasi lainnya seperti Support Vector Machine, K-Nearest Neighbor dan lain-lain sebagai perbandingan untuk performa model.

Referensi

[1] F. Adams, L. Ernawati, and N. Chamidah, "Analisis Sentimen Vaksin COVID-19 pada Twitter Menggunakan Algoritma Support Vector Machine," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, pp. 221–231, 2021.

[2] N. Herlinawati, Y. Yuliani, S. Faizah, W. Gata, and Samudi, "ANALISIS SENTIMEN ZOOM CLOUD MEETINGS DI PLAY STORE MENGGUNAKAN NAÏVE BAYES DAN SUPPORT VECTOR MACHINE," *CESS (Journal Comput. Eng. Syst. Sci.*, pp. 293–298, 2020.

[3] F. F. Irfani, "ANALISIS SENTIMEN REVIEW APLIKASI RUANGGURU MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE," *JBMI (Jurnal Bisnis Manaj. dan Inform.*, pp. 258–266, 2020.

[4] D. Y. Praptiwi, "ANALISIS SENTIMEN ONLINE REVIEW PENGGUNA E-COMMERCE MENGGUNAKAN METODE SUPPORT VECTOR MACHINE DAN MAXIMUM ENTROPY," *J. Univ. Islam Indones.*, 2018.

[5] L. B. I. M. A. Mudeb, "Perbandingan Metode Klasifikasi Support Vector Machine dan Naïve Bayes untuk Analisis Sentimen pada Ulasan Tekstual di Google Play Store," *Ilk. J. Ilm.*, pp. 154–161, 2020.

[6] Fatayat and R. A. Nugroho, "ANALISA PENENTUAN DOSEN PEMBIMBING TUGAS AKHIR MAHASISWA MENGGUNAKAN NAIVE BAYES CLASSIFIER," *J. SIMTIKA*, vol. 4, pp. 1–7, 2021.

[7] W. Satriaji and R. Kusumaningrum, "Effect of Synthetic Minority Oversampling Technique (SMOTE), Feature Representation, and Classification Algorithm on Imbalanced Sentiment Analysis," *Int. Conf. Informatics Comput. Sci.*, vol. 2, 2018.