

Klasifikasi Sentimen Terhadap Penyelenggaraan Haji Indonesia Tahun 2023 Melalui Media Sosial Twitter Menggunakan Teknik *Balancing* Dan Metode *Gaussian Naïve Bayes*

Classification Of Sentiments Towards The Organization Of The Indonesian Hajj In 2023 On Twitter Using Balancing Techniques And Naïve Bayes Gaussian Methods

Juniartha Indra Pratama¹⁾, Agung Nilogiri²⁾, Ari Eko Wardoyo³⁾

¹Mahasiswa Fakultas Teknik, Universitas Muhammadiyah Jember
email: juniarthapratama1997@gmail.com

²Dosen Fakultas Teknik, Universitas Muhammadiyah Jember
email: agungnilogiri@unmuhjember.ac.id

³Dosen Fakultas Teknik, Universitas Muhammadiyah Jember
email: arieko@unmuhjember.ac.id

Abstrak

Haji merupakan salah satu ibadah umat Islam dan termasuk salah satu bagian dari rukun islam selain syahadat, salat, puasa, dan zakat. Di media sosial seperti Twitter berisi komentar mengenai haji di Indonesia, mulai dari apresiasi serta kritik bahkan hinaan terhadap penyelenggara dan para jamaah. Hal ini memicu respon positif dan negatif antar pengguna twitter mengenai haji di Indonesia tahun 2023. Mengamati Hal tersebut peneliti tertarik untuk melakukan klasifikasi sentimen. Namun dalam melakukan klasifikasi, terdapat permasalahan yang sering ditemui oleh peneliti yaitu ketidakseimbangan data. Maka penelitian ini menggunakan teknik *balancing* untuk mengatasi ketidakseimbangan data. Penelitian ini bertujuan untuk mengetahui performa akurasi, presisi, dan *recall* sebelum dan sesudah menggunakan teknik *balancing*(*Random Oversampling*, *Random Undersampling*, *SMOTE*). Hasil Performa sebelum menggunakan teknik *balancing* yaitu akurasi sebesar 77%, presisi sebesar 86%, *recall* sebesar 79%, untuk *Random Oversampling* diperoleh akurasi sebesar 85%, presisi sebesar 98%, *recall* sebesar 73%, untuk *Random Undersampling* diperoleh akurasi sebesar 85%, presisi sebesar 86%, *recall* sebesar 86% dan untuk *SMOTE* diperoleh akurasi sebesar 90%, presisi sebesar 98%, *recall* sebesar 83%.

Keywords: *Analisis Sentimen, Gaussian Naïve Bayes, Random Oversampling, Random Undersampling, SMOTE*

Abstract

Hajj is one of the Muslim pilgrimages and it is part of the pillars of Islam apart from the shahada, prayer, fasting, and zakat. Social media such as Twitter contains comments about the Hajj in Indonesia, from appreciation and criticism even an insult to the organisers and pilgrims. This triggered positive and negative responses among Twitter users regarding the Hajj in Indonesia in 2023. Observing this, researcher is interested in classifying sentiment. However, in carrying out classification, there is a problem that researchers often encounter, namely data imbalance. So this research uses balancing techniques to overcome data imbalance. This research aims to determine the performance of accuracy, precision, and recall before and after using balancing techniques (Random Oversampling, Random Undersampling, SMOTE). Performance results before using the balancing technique are an accuracy is 77%, Precision is 86%, and Recall is 79%, for Random Oversampling, accuracy is 85%, Precision is 98%, and Recall is 73%, for Random Undersampling, accuracy is 85%, Precision is 86%, Recall is 86% and for SMOTE, accuracy is 90%, Precision is 98%, Recall is 83%.

Keywords: *Gaussian Naïve Bayes, Random Oversampling, Random Undersampling, Sentiment analysis, SMOTE*

1. PENDAHULUAN

Haji merupakan salah satu ibadah umat Islam dan termasuk salah satu bagian dari rukun islam selain syahadat, salat, puasa, dan zakat. Haji memiliki pengertian datang ke rumah Allah atau Baitullah untuk melakukan ibadah pada waktu dan aturan yang telah ditentukan oleh penyelenggara bagi muslim dan muslimat yang mampu secara fisik dan finansial. Aturan mengenai ibadah haji telah tertulis dalam Alquran Surat Al-Imran ayat ke-97 yang jika diterjemahkan artinya adalah melaksanakan ibadah haji ke Baitullah merupakan salah satu kewajiban umat muslim terhadap Allah SWT. Ayat tersebut juga menyebutkan siapa saja yang wajib melakukan haji, salah satunya adalah orang-orang yang mampu mengadakan perjalanan ke rumah Allah atau Baitullah.

Berdasarkan Keputusan Menteri Agama nomor 189 Tahun 2003 tentang Kuota Haji Indonesia 1444 H/2023 M ditetapkan bahwa kuota haji Indonesia berjumlah 221.000 orang dengan adanya tambahan kuota 8.000 maka total kuota haji Indonesia menjadi 229.000 dengan jadwal lengkap yaitu dari tanggal 23 Mei 2023 – 3 Agustus 2023. Di media sosial seperti Twitter berisi komentar komentar mengenai haji di Indonesia, mulai dari apresiasi serta kritik bahkan hinaan terhadap penyelenggara dan para jamaah. Hal ini memicu respon positif dan negatif antar pengguna twitter mengenai haji di Indonesia tahun 2023. Mengamati Hal tersebut peneliti tertarik untuk melakukan klasifikasi sentimen terhadap penyelenggaraan Haji Indonesia tahun 2023.

Namun dalam melakukan klasifikasi terdapat permasalahan yang sering ditemui oleh peneliti yaitu ketidakseimbangan data. Data tidak seimbang adalah keadaan dimana distribusi kelas data tidak seimbang, hal ini terjadi dikarenakan jumlah kelas data lebih banyak atau lebih sedikit dengan jumlah kelas data lainnya. Kelompok kelas data yang lebih banyak disebut dengan *majority* sedangkan kelompok kelas data yang lebih sedikit disebut dengan *minority*. Apabila melakukan klasifikasi dengan data yang tidak seimbang maka akan menyebabkan akurasi yang jauh lebih tinggi terhadap data yang memiliki jumlah data lebih

banyak (*majority*) daripada data yang memiliki jumlah data yang lebih sedikit (*minority*). Perbandingan antara kelas *minority* dan *majority* disebut dengan *Imbalance Ratio* (IR) atau rasio ketidakseimbangan. Semakin besar perbedaan antara kelas *majority* dan *minority* maka akan semakin besar nilai dari *Imbalance Ratio* (IR) atau rasio ketidakseimbangan. Maka perlu adanya cara untuk menyeimbangkan jumlah distribusi data seperti *Random Oversampling*, *Random Undersampling*, dan *SMOTE*. Dalam melakukan klasifikasi terdapat beberapa metode salah satunya adalah *Gaussian Naïve Bayes*.

Gaussian Naïve Bayes adalah salah satu turunan dari algoritma *Naïve Bayes* yang terdapat pada data *mining* dimana penggunaannya tergolong mudah dan pemrosesannya juga memiliki waktu yang cepat, selain itu metode ini juga mudah diimplementasikan dengan struktur sederhana dan memiliki tingkat efektifitas tinggi.

Berdasarkan latar belakang tersebut, peneliti akan melakukan perbandingan performa model klasifikasi *Gaussian Naïve Bayes* untuk melakukan klasifikasi sentimen terhadap penyelenggaraan Haji Indonesia tahun 2023 dan menggunakan teknik *balancing* yaitu *Random Oversampling*, *Random Undersampling* dan *SMOTE*, dengan judul adalah “Klasifikasi Sentimen Terhadap Penyelenggaraan Haji Indonesia Tahun 2023 Melalui Media Sosial Twitter Menggunakan Teknik *Balancing* Dan Metode *Gaussian Naïve Bayes*”.

A. RUMUSAN MASALAH

Berdasarkan uraian latar belakang di atas, maka peneliti merumuskan permasalahan sebagai berikut :

1. Berapa tingkat akurasi, presisi, dan *recall* yang diperoleh dalam menggunakan algoritma *Gaussian Naïve Bayes* pada klasifikasi sentimen penyelenggaraan haji Indonesia 2023 pada twitter?
2. Berapa tingkat akurasi, presisi, dan *recall* yang diperoleh dalam menggunakan algoritma *Gaussian Naïve Bayes* pada klasifikasi sentimen penyelenggaraan haji Indonesia 2023 pada twitter setelah menggunakan Teknik *balancing* (*Random*

Oversampling, Random Undersampling dan SMOTE) ?

B. TUJUAN PENELITIAN

Berdasarkan rumusan masalah diatas, adapun Tujuan yang ingin dicapai dari penelitian ini adalah sebagai berikut :

1. Mengetahui tingkat akurasi, presisi, dan *recall* yang diperoleh dalam menggunakan algoritma *Gaussian Naïve Bayes* pada klasifikasi sentimen haji Indonesia 2023 pada twitter.
2. Mengetahui tingkat akurasi, presisi, dan *recall* yang diperoleh dalam menggunakan algoritma *Gaussian Naïve Bayes* pada klasifikasi sentimen haji Indonesia 2023 pada twitter setelah menggunakan Teknik *balancing (Random Oversampling, Random Undersampling dan SMOTE)*.

C. BATASAN MASALAH

Batasan masalah pada penelitian ini adalah sebagai berikut :

1. Mengetahui jumlah sentimen yang bernilai positif dan negatif pada penyelenggaraan haji Indonesia 2023 di twitter pada periode tertentu sehingga menjadi dataset untuk dikelola menggunakan metode *Gaussian Naïve Bayes*.
2. Diharapkan menjadi referensi bagi peneliti selanjutnya dalam melakukan klasifikasi sentimen dengan menggunakan metode *Gaussian Naïve Bayes* serta menggunakan Teknik *balancing (Random Oversampling, Random Undersampling dan SMOTE)*.

D. MANFAAT PENELITIAN

Manfaat dari penelitian ini adalah sebagai berikut :

1. Data diperoleh dari Twitter berdasarkan masa haji Indonesia pada tahun 2023 yaitu pada tanggal 23 Mei 2023 – 3 Agustus 2023
2. Proses pencarian data dengan menggunakan kata kunci “haji”.
3. penelitian menggunakan API atau *Application Programming Interface Twitter* dalam melakukan *crawling* yang didapatkan dari *Twitter Development*.

4. Menggunakan bahasa pemrograman *python* dan *software* yang digunakan berupa *Jupyter Notebook*.

2. TINJAUAN PUSTAKA

A. Twitter

Twitter merupakan platform yang menyediakan beragam fitur. Salah satunya fitur yang sering digunakan adalah fitur “*tweet*”, fitur tersebut digunakan oleh pengguna untuk membagikan cerita, pengalaman dan hal menarik sebagai status di twitter. Sebuah tweet hanya memuat 140 karakter.. (Kotwal et al., 2016). Pengguna aktif Twitter di Indonesia menduduki peringkat 5 di dunia dengan 18,45 juta pengguna per Januari 2022(Statista, 2022).

B. Text Mining

Text mining adalah sebuah variasi dari *data mining* untuk menemukan pola menarik dari kumpulan data tekstual berskala besar (Sunoto, 2017). *Text Mining* adalah salah satu teknik dalam melakukan klasifikasi dokumen.

C. Preprocessing Text

Pada tahapan ini, data yang telah diperoleh merupakan data tidak terstruktur sehingga perlu adanya *preprocessing* untuk mengubah menjadi data terstruktur. *Preprocessing* merupakan langkah awal dalam melakukan analisis sentimen sehingga tahap ini harus dilakukan (Tuhuteru, 2020). Dilakukannya sebuah tahapan *preprocessing* pada data training melalui tahapan seperti berikut:

1. Cleansing

Cleansing merupakan proses penghilangan semua karakter bukan alfabet sehingga data yang berhasil dikumpulkan akan dihilangkan beberapa karakter yang tidak sesuai dan tidak memiliki arti dalam analisis sentimen. (Tuhuteru, 2020).

2. Case Folding

Pada tahap ini, setiap huruf yang terdapat dalam kalimat akan diubah menjadi huruf kecil atau *lowercase*. (Tuhuteru, 2020).

3. Tokenizing

Pada tahap ini, komentar yang telah melalui 2 tahapan di awal akan dipisahkan dari

kalimatnya sehingga menjadi per kata atau token. (Tuhuteru, 2020).

4. Stopword Removal

Pada proses *stopword removal* kata-kata penting yang dipilih dari hasil token berupa kata-kata yang digunakan dalam dokumen atau disebut dengan *filtering*. (Tuhuteru, 2020).

5. Stemming

Stemming yaitu mengembalikan kata menjadi kata dasar dengan cara menghilangkan kata imbuhan berupa awalan, akhiran atau sisipan. (Tuhuteru, 2020).

D. Analisis Sentimen

Analisis sentimen adalah sebuah teknik untuk mengekstrak sebuah data berbentuk teks. Tujuan analisis sentimen adalah untuk memperoleh sebuah informasi tentang suatu hal atau suatu kalimat. Hasil dari analisis sentimen berupa nilai positif atau negatif. (Sari, 2019).

E. TF-IDF

TF-IDF atau *Term Frequency-Inverse Document Frequency* adalah suatu proses dalam melakukan transformasi data dari data berupa teks menjadi data numerik. Tujuannya untuk melakukan pembobotan pada tiap kata atau fitur. Metode ini terkenal efisien, mudah dan akurat (Ria, 2018). *TF-IDF* adalah sebuah ukuran statistik yang digunakan untuk mengevaluasi seberapa penting sebuah kata di dalam sebuah dokumen.

TF adalah frekuensi kemunculan kata untuk setiap dokumen. *TF* menunjukkan seberapa penting kata itu dalam setiap dokumen. Dalam menghitung *TF* ditunjukkan pada persamaan berikut ini :

$$tf_{t,d} = \frac{n_{t,d}}{N} \quad (2.1)$$

Keterangan :

$n_{t,d}$: nilai istilah yang muncul
 N : semua term dalam dokumen
 $tf_{t,d}$: frekuensi kemunculan kata pada sebuah dokumen.

Setelah berhasil mendapatkan nilai *TF*, selanjutnya adalah menghitung nilai *IDF*, dimana nilai *IDF* digunakan untuk mengukur seberapa pentingnya sebuah kata. Berikut

adalah rumus yang digunakan untuk menentukan nilai *IDF* yang ditunjukkan pada persamaan berikut ini :

$$idf_d = \log\left(\frac{N_d}{df}\right) \quad (2.2)$$

Keterangan :

N_d : total dokumen
 df : banyak dokumen yang mengandung term tersebut
 idf_d : jumlah kemunculan suatu kata dalam dokumen.

Setelah nilai *IDF* berhasil ditemukan. Maka kita dapat menentukan nilai sebuah *TF-IDF*. Hasil *TF-IDF* ini adalah hasil perkalian dari *TF* dikalikan dengan *IDF* yang ditunjukkan pada persamaan berikut ini :

$$tf.idf_{t,d} = tf_{t,d} \times idf_d \quad (2.3)$$

Keterangan :

$tf.idf_{t,d}$: *Term Frequency – Inverse Document Frequency*

$tf_{t,d}$: frekuensi kemunculan kata pada sebuah dokumen.

$idf_{t,d}$: jumlah kemunculan suatu kata dalam dokumen.

Bobot suatu istilah semakin besar apabila istilah tersebut sering muncul dalam suatu dokumen dan sebaliknya, bobot istilah semakin kecil apabila istilah tersebut muncul dalam banyak dokumen (Grossman, 1998)

F. Python

Python adalah sebuah bahasa pemrograman interpretatif yang multiguna dengan memiliki filosofi perancangan yang fokus pada tingkat keterbacaan kode. (Ljubomir Perkovic, 2012).

G. Jupyter Notebook

Jupyter Notebook adalah sebuah website untuk mengintegrasikan antara kode dengan *output* di dalam satu dokumen secara interaktif. *Jupyter Notebook* akan otomatis terinstall ketika kita telah menginstall *python* dengan

anaconda. Jika menginstall *python* tidak melalui *anaconda* maka kita terlebih dahulu harus menginstal melalui pip atau conda (Faqih, 2019)

H. Gaussian Naïve Bayes

Gaussian Naïve Bayes memiliki pengklasifikasi *Naive Bayes* paling sederhana yang memiliki asumsi bahwa data dari masing-masing label diambil dari distribusi Gaussian sederhana. (dqlab, 2021). Perhitungan *Gaussian Naïve bayes* seperti persamaan berikut :

$$P(X_i = x_i | Y = y_i) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}} \quad (2.4)$$

Keterangan :

- P : Peluang
- X_i : Atribut ke i
- x_i : Nilai atribut ke i
- Y : Kelas yang dicari
- y_i : Sub-kelas yang dicari
- σ : standar deviasi
- μ : nilai rata-rata hitung / *mean*

Perhitungan standar deviasi memiliki persamaan seperti berikut :

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1}} \quad (2.5)$$

Keterangan :

- σ : Standar Deviasi
- μ : nilai rata – rata hitung (mean)
- x_i : nilai x ke- i
- n : jumlah sampel

I. Pengujian

Pengujian adalah sebuah proses dalam mengkaji sebuah program dengan cara mengumpulkan dan menganalisis sebuah data tentang aktivitas program, karakteristik, dan hasil program. Tujuan pengujian adalah untuk meningkatkan keefektifan, membuat penilaian, dan menginformasikan hasil keputusan program (Patton, 1987). Confusion matrix adalah pengujian paling optimal untuk mengklasifikasikan multiclass yang dalam

penelitian ini kelas positif atau negatif (Hossin & Sulaiman, 2015).

Tabel 1. Tabel Confusion Matrix

		Nilai Sebenarnya	
		Positif	Negatif
Nilai Prediksi	Positif	TP	FP
	Negatif	FN	TN

Sumber : Hossin & Sulaiman, 2015

Keterangan :

1. TP (*True Positif*) = Jumlah prediksi yang positif dengan kejadian yang sebenarnya juga positif.
2. TN (*True Negatif*) = Jumlah prediksi yang negatif dengan kejadian yang sebenarnya juga negatif.
3. FP (*False Positif*) = Jumlah prediksi yang positif dengan kejadian yang sebenarnya negatif.
4. FN (*False Negatif*) = Jumlah prediksi yang negatif dengan kejadian yang sebenarnya positif.

1. Akurasi merupakan kesesuaian antara nilai hasil prediksi pengujian dengan nilai aktual. Tujuannya adalah untuk mengetahui berapa persen kalimat yang benar diprediksi positif dan negatif dari keseluruhan kalimat. (Arthana Resika, 2019). Akurasi memiliki persamaan sebagai berikut :

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.6)$$

2. Presisi merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Tujuannya adalah untuk mengetahui berapa persen kalimat yang benar positif dari keseluruhan kalimat yang diprediksi positif. (Arthana Resika, 2019). Presisi memiliki persamaan sebagai berikut :

$$\text{Presisi} = \frac{TP}{TP + FP} \quad (2.7)$$

3. *Recall* adalah Merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Tujuannya adalah untuk

mengetahui berapa persen kalimat yang diprediksi positif dari keseluruhan kalimat yang sebenarnya positif. (Arthana Resika, 2019). Recall memiliki persamaan sebagai berikut :

$$Recall = \frac{TP}{TP + FN} \quad (2.8)$$

J. *Balancing Data*

Balancing data adalah tindakan untuk menyeimbangkan sebuah data yang memiliki distribusi tiap kelas yang berbeda atau salah satu kelas dari data tersebut memiliki dominasi yang lebih apabila dibandingkan dengan kelas lainnya sehingga perlu penyesuaian. Hal ini sering terjadi pada proses klasifikasi yang disebut dengan *imbalance data*. Terdapat beberapa metode untuk mengatasi ketidakseimbangan data diantaranya dengan teknik pengambilan sampel dasar yaitu *Random Undersampling*, *Random Oversampling* dan *SMOTE* (Almie, 2023).

K. *Random Oversampling*

Random Oversampling adalah sebuah metode *non-heuristik*. Tujuan *Random Oversampling* adalah menyeimbangkan distribusi kelas melalui replikasi acak contoh kelas minoritas. Cara *Random Oversampling* bekerja adalah menggandakan kelas minoritas secara acak sehingga seimbang dengan kelas mayoritas. (Kotsiantis et al., 2006).

L. *Random Undersampling*

Random Undersampling adalah sebuah metode *non-heuristik*. Tujuan *Random Undersampling* adalah menyeimbangkan distribusi kelas melalui penghapusan acak contoh kelas mayoritas. Cara *Random Undersampling* bekerja adalah menghapus kelas mayoritas secara acak sehingga seimbang dengan kelas minoritas. (Kotsiantis et al., 2006).

M. *SMOTE*

SMOTE (Synthetic Minority Oversampling Technique) adalah sebuah teknik pendekatan *oversampling* yang secara sintesis menghasilkan *instance* dengan memilih secara acak *instance* dari kelas minoritas dan menggunakan metode interpolasi untuk menghasilkan *instance* antara titik yang dipilih dan *instance* yang berdekatan.

(Kaur & Gosain, 2018). *SMOTE* bekerja dengan cara memilih satu atau lebih tergantung dari rasio pengambilan sampel berlebih dari tetangga terdekat k dari *instance* kelas minoritas setelah itu melakukan interpolasi acak dua *instance* untuk membuat *instance* sintesis baru.

N. *K-Fold Cross Validation*

K-fold cross validation adalah salah satu cara untuk mengevaluasi keakuratan suatu model. Metode *k-fold cross validation* bekerja dengan cara melakukan generalisasi dengan membagi data ke dalam k bagian yang berukuran sama. Di dalam proses tersebut, salah satu dari partisi dipilih untuk *testing* atau pengujian, dan sisanya digunakan untuk *training* atau latihan (Torres, et al., 2012). Tahapan tersebut dilakukan secara berulang sebanyak k kali sehingga setiap partisi digunakan untuk testing tepat satu kali. Rumus perhitungan untuk *k-fold cross validation* adalah sebagai berikut:

Keterangan :

D : Hasil pembagian data dalam file dokumen

n : jumlah data

k : nilai *k-fold cross validation*

$$D = \frac{n}{k} \quad (2.9)$$

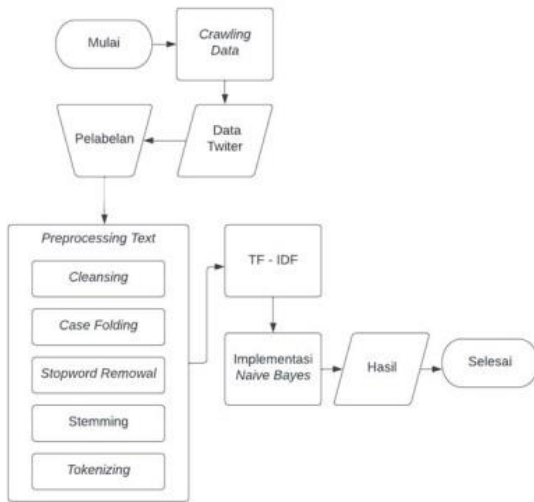
3. METODOLOGI PENELITIAN

A. *Pengumpulan Data*

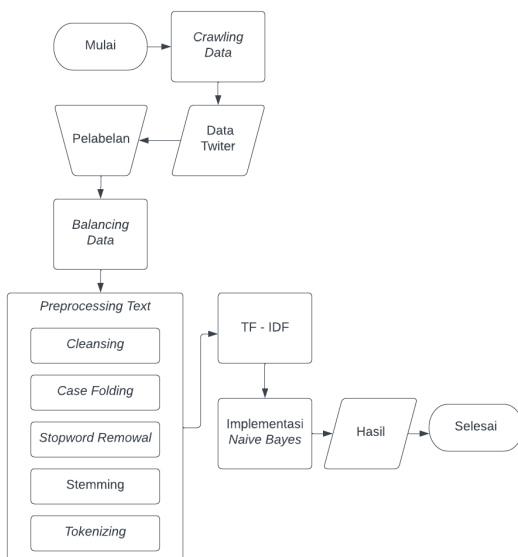
Data diperoleh dengan cara *crawling* menggunakan *API* Twitter menggunakan kata kunci “haji” dari tanggal penyelenggaraan haji hingga selesai yaitu pada tanggal 23 Mei 2023 – 3 Agustus 2023. Total 1000 data dengan data positif sebanyak 660 data dan data negatif sebanyak 340 data.

B. *Flowchart*

Flowchart pada penelitian ini ditunjukkan pada Gambar 1 dan Gambar 2 :



Gambar 1. Flowchart tanpa *balancing*
 Sumber : Rancangan Sendiri



Gambar 2. Flowchart menggunakan *balancing*
 Sumber : Rancangan Sendiri

C. Labelling Data

Pelabelan data atau *Labelling* data dilakukan secara manual. Pelabelan ini terbagi menjadi 2 yaitu kelas positif dan kelas negatif. Dimana pelabelan tersebut dilakukan oleh ahli bahasa yang berkompeten di bidangnya yaitu ibu Berlian Eka Kurnia, S.I.P., M.A. sebagai Kepala Perpustakaan Universitas Muhammadiyah Jember.

D. Balancing Data

Pada penelitian ini dikarenakan data sampel memiliki label positif dan negatif yang tidak seimbang, diperoleh dari 1000 data yang terdiri dari 660 data positif dan 340 data negatif maka untuk mengatasi ketidakseimbangan data dilakukan teknik *balancing* data yaitu *random oversampling*, *random undersampling*, dan *SMOTE*. Pada *random oversampling* dan *SMOTE* untuk menyeimbangkan data dengan menjadikan jumlah data terbesar sebagai patokan, sedangkan *random undersampling* untuk menyeimbangkan data dengan menjadikan jumlah data terkecil sebagai patokan. Sehingga didapatkan hasil data sebagai tabel berikut :

Tabel 2. Hasil *Balancing* Data

Label		<i>Random Oversampling</i>	<i>Random Undersampling</i>	<i>SMOTE</i>
Positif	660 Data	660 Data	340 Data	660 Data
Negatif	340 Data	660 Data	340 Data	660 Data
Total	1000 Data	1320 Data	680 Data	1320 Data

Sumber : Hasil Perhitungan

4. HASIL DAN PEMBAHASAN

Berdasarkan skenario pengujian yang telah dilakukan terhadap penyelenggaraan haji indonesia tahun 2023 melalui media sosial twitter menggunakan teknik *balancing* dan metode *gaussian naive bayes* diperoleh rekapitulasi seperti berikut :

A. Normal Data

Rekapitulasi hasil berdasarkan data normal ditunjukkan pada **Tabel 3** :

Tabel 3. Rekapitulasi hasil data normal

K-Fold	Langkah Uji	Normal Data		
		Akurasi	<i>Presisi</i>	<i>Recall</i>
2 – Fold	Langkah Uji 1	76%	83%	81%
	Langkah Uji 2	69%	76%	76%
5 – Fold	Langkah uji 1	74%	85%	77%
	Langkah uji 2	74%	82%	77%
	Langkah uji 3	71%	76%	78%

K-Fold	Langkah Uji	Normal Data		
		Akurasi	Presisi	Recall
	Langkah uji 4	60%	73%	65%
	Langkah uji 5	64%	72%	70%
10 – Fold	Langkah uji 1	77%	86%	79%
	Langkah uji 2	73%	84%	76%
	Langkah uji 3	69%	80%	70%
	Langkah uji 4	77%	82%	85%
	Langkah uji 5	69%	74%	74%
	Langkah uji 6	76%	80%	82%
	Langkah uji 7	57%	70%	61%
	Langkah uji 8	64%	77%	69%
	Langkah uji 9	73%	85%	72%
	Langkah uji 10	68%	69%	79%

Sumber : Hasil Perhitungan

Berdasarkan **Tabel 3** di atas dapat diperoleh nilai performa terbaik berdasarkan data normal yaitu pada 10-fold *cross validation* langkah uji 1 dengan akurasi sebesar 77%, presisi sebesar 85% dan *recall* sebesar 81% dan 10-fold *cross validation* langkah uji 4 dengan akurasi sebesar 77%, presisi sebesar 82% dan *recall* sebesar 85%.

B. Random Oversampling Data

Rekapitulasi hasil berdasarkan data *Random Oversampling* ditunjukkan pada **Tabel 4** :

Tabel 4. Rekapitulasi hasil data *Random Oversampling*

K-Fold	Langkah Uji	Random Oversampling Data		
		Akurasi	Presisi	Recall
2 – Fold	Langkah Uji 1	79%	83%	75%
	Langkah Uji 2	78%	84%	69%
5 – Fold	Langkah uji 1	81%	92%	69%
	Langkah uji 2	82%	87%	75%
	Langkah uji 3	81%	87%	76%
	Langkah uji 4	81%	86%	70%
	Langkah uji 5	81%	95%	67%
10 – Fold	Langkah uji 1	84%	96%	73%
	Langkah uji 2	79%	91%	64%
	Langkah uji 3	83%	98%	69%
	Langkah uji 4	84%	85%	78%
	Langkah uji 5	83%	89%	77%
	Langkah uji 6	83%	91%	74%

K-Fold	Langkah Uji	Random Oversampling Data		
		Akurasi	Presisi	Recall
	Langkah uji 7	80%	86%	72%
	Langkah uji 8	82%	90%	64%
	Langkah uji 9	85%	98%	73%
	Langkah uji 10	79%	91%	63%

Sumber : Hasil Perhitungan

Berdasarkan **Tabel 4** di atas dapat diperoleh nilai performa terbaik berdasarkan *Random Oversampling* data yaitu pada 10-fold *cross validation* langkah uji 9 dengan akurasi sebesar 85%, presisi sebesar 98% dan *recall* sebesar 73%.

C. Random Undersampling Data

Rekapitulasi hasil berdasarkan data *Random Undersampling* ditunjukkan pada **Tabel 5**:

Tabel 5. Rekapitulasi hasil data *Random Undersampling*.

K-Fold	Langkah Uji	Random Undersampling Data		
		Akurasi	Presisi	Recall
2 – Fold	Langkah Uji 1	74%	74%	73%
	Langkah Uji 2	76%	77%	74%
5 – Fold	Langkah uji 1	77%	78%	73%
	Langkah uji 2	71%	71%	65%
	Langkah uji 3	82%	84%	81%
	Langkah uji 4	70%	78%	67%
	Langkah uji 5	72%	70%	60%
10 – Fold	Langkah uji 1	76%	78%	78%
	Langkah uji 2	76%	75%	70%
	Langkah uji 3	68%	69%	73%
	Langkah uji 4	75%	76%	57%
	Langkah uji 5	85%	86%	86%
	Langkah uji 6	78%	79%	77%
	Langkah uji 7	65%	77%	63%
	Langkah uji 8	74%	79%	66%
	Langkah uji 9	76%	80%	57%
	Langkah uji 10	71%	69%	60%

Sumber : Hasil Perhitungan

Berdasarkan **Tabel 5** di atas dapat diperoleh nilai performa terbaik berdasarkan data *Random Undersampling* yaitu pada *10-fold cross validation* langkah uji 5 dengan akurasi sebesar 85%, presisi sebesar 86% dan *recall* sebesar 86%.

D. SMOTE DATA

Rekapitulasi hasil berdasarkan data *SMOTE* ditunjukkan pada **Tabel 6**:

Tabel 6. Rekapitulasi hasil data *SMOTE*.

K-Fold	Langkah Uji	<i>SMOTE</i> Data		
		Akurasi	Presisi	Recall
2 – Fold	Langkah Uji 1	82%	88%	74%
	Langkah Uji 2	79%	89%	68%
5 – Fold	Langkah uji 1	84%	91%	75%
	Langkah uji 2	84%	90%	76%
	Langkah uji 3	81%	92%	69%
	Langkah uji 4	81%	90%	69%
	Langkah uji 5	79%	86%	69%
10 – Fold	Langkah uji 1	85%	98%	75%
	Langkah uji 2	86%	86%	79%
	Langkah uji 3	81%	87%	72%
	Langkah uji 4	90%	98%	83%
	Langkah uji 5	85%	91%	74%
	Langkah uji 6	76%	92%	65%
	Langkah uji 7	83%	97%	69%
	Langkah uji 8	82%	94%	69%
	Langkah uji 9	83%	92%	73%
	Langkah uji 10	82%	79%	77%

Sumber : Hasil Perhitungan

Berdasarkan **Tabel 6** di atas dapat diperoleh nilai performa terbaik berdasarkan data *SMOTE* yaitu pada *10-fold cross validation* langkah uji 4 dengan akurasi sebesar 90%, presisi sebesar 98% dan *recall* sebesar 83%.

5. KESIMPULAN DAN SARAN

A. KESIMPULAN

Setelah dilakukan implementasi, pengujian, dan analisis hasil pada “Klasifikasi Sentimen

Terhadap Penyelenggaraan Haji Indonesia Tahun 2023 Melalui Media Sosial Twitter Menggunakan Teknik *Balancing* Dan Metode *Gaussian Naïve Bayes*” maka dapat diambil kesimpulan sebagai berikut :

1. Pada Penelitian ini performa terbaik berdasarkan data normal yaitu pada *10-fold cross validation* langkah uji 1 dengan akurasi sebesar 77%, presisi sebesar 85% dan *recall* sebesar 81% dan *10-fold cross validation* langkah uji 4 dengan akurasi sebesar 77%, presisi sebesar 82% dan *recall* sebesar 85%
2. Pada Penelitian ini performa terbaik berdasarkan data *Random Oversampling* yaitu pada *10-fold cross validation* langkah uji 9 dengan akurasi sebesar 85%, presisi sebesar 98% dan *recall* sebesar 73%.
3. Pada Penelitian ini performa terbaik berdasarkan data *Random Undersampling* yaitu pada *10-fold cross validation* langkah uji 5 dengan akurasi sebesar 85%, presisi sebesar 86% dan *recall* sebesar 86%.
4. pada Penelitian ini performa terbaik berdasarkan data *SMOTE* yaitu pada *10-fold cross validation* langkah uji 4 dengan akurasi sebesar 90%, presisi sebesar 98% dan *recall* sebesar 83%.
5. Pada Penelitian ini performa terbaik berdasarkan seluruh data yaitu pada *10-fold cross validation* teknik *balancing SMOTE* langkah uji 4 dengan akurasi sebesar 90%, presisi sebesar 98% dan *recall* sebesar 83%.
6. Pada Penelitian ini teknik *balancing* dapat meningkatkan performa akurasi, presisi dan *recall*.

B. SARAN

1. Penelitian selanjutnya dapat menambahkan metode klasifikasi sebagai perbandingan performa agar dapat mengetahui metode mana yang lebih baik.
2. Penelitian selanjutnya dapat menambahkan teknik *balancing* yang berbeda seperti *SMOTE-Text* yang dioptimalkan khusus untuk data teks sehingga mendapatkan hasil akurasi, presisi dan *recall* yang bervariasi.

6. DAFTAR PUSTAKA

- Arthana Resika.2019. “Mengenai Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning ”, <https://rey1024.medium.com/mengenai-accuracy-precision-recall-dan-specificity-septa-yang-diprioritaskan-b79ff4d77de8> , diakses pada 2 Juli 2023, pukul 02:20.
- Almie, A. U. 2023. Klasifikasi Sentimen Ulasan Pelanggan Layanan Ekspedisi Menggunakan Algoritma Naive Bayes. Universitas Muhammadiyah Jember.
- Dqlab.2021. “Belajar Algoritma Naive Bayes”, <https://www.dqlab.id/belajar-algoritma-naive-bayes> , diakses pada 23 November 2022, pukul 10.49.
- Faqih, F. 2019. “Belajar Python Dasar : Memahami Jupyter Notebook dan cara menggunakannya”, <https://ngodingdata.com/memahami-jupyter-notebook-dan-cara-menggunakannya/> , diakses pada 23 November 2022, pukul 08.43.
- Grossman, D., Ophir, F., (1998), Information Retrieval : Algorithm and Heuristics. Kluwer Academic Publisher, USA.
- Hossin, M. and Sulaiman, M. N. (2015). A Review on E-Valuation Metrics for Data Classification. International Journal of Data Mining & Knowledge Management Process (IJDKP), 5(2), 1–11.
- Kaur, P., & Gosain, A. (2018). Comparing The Behavior of Oversampling Versus Undersampling. Research Gate, 653(January). <https://doi.org/10.1007/978-981-10-6602-3>.
- Kotsiantis S. B., D. K. and P. E. P. (2006). Data Preprocessing for Supervised Learning.
- Patton, Michael Quinn. (1987.) “Qualitative Education Methods”. Beverly Hills: Sage Publication.
- Perkovic, Ljubomir. 2012. Introduction to Computing Using Python: An Application Development Focus.
- Ria, M. 2018. Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Syarah Umdatil Ahkam), Universitas Islam Negeri Syarif Hidayatullah Jakarta.
- Statista. (2020). “Leading countries based on number of Twitter users as of January 2022”, <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/> diakses pada 28 Oktober 2022, pukul 14.35.
- Sunoto, Y. 2017. Analisis Testimonial Wisatawan Menggunakan Text Mining Dengan Metode Naive Bayes Dan Decision Tree, Studi Kasus Pada Hotel – Hotel Di Jakarta. Institut Bisnis Dan Informatika Kwik Kian Gie.
- Torres, T. J. G. M., J. A. Saez., F Herrera. (2012). Study on the impact of partition-induced dataset shift on k-fold cross-validation. IEEE Transactions on Neural Networks and Learning Systems. 23(8): 1304– 1312.
- Tuhuteru. 2020. Analisis Sentimen Masyarakat Terhadap Pembatasan Sosial Berskala Besar Menggunakan Algoritma Support Vector Machine. Universitas Kristen Indonesia Maluku.