# KLASIFIKASI SENTIMEN PERGELARAN MOTOGP DI INDONESIA MENGGUNAKAN ALGORITMA CORRELATED NAÏVE BAYES CLASIFIER

**Ridwan Indransyah[1], Yulison Herry Chrisnanto, S.T., M.T[2], Puspita Nurul Sabrina, S.Kom., M.T[3]**
*[1,2,3]Program Studi Informatika, Fakultas sains dan Informatika, Universitas Jenderal Achmad Yani*
*Email: [1]ridwani18@if.unjani.ac.id, [2] yhc@if.unjani.ac.id,*
*[3] puspita.sabrina@lecture.unjani.ac.id.*

**ABSTRAK**
*Knowing the public's sentiment towards the international MotoGP event which has been held in Indonesia in 2022 is very necessary because the role of the community is very influential in the implementation and public interest in visiting an international event is still few and difficult because the information is still limited. Tweets, comments, reviews, and opinions of people using social media play an important role in determining whether a particular population is satisfied with products, performances, and services. The method used in this study is the Correlated Naïve Bayes Classifier (CNBC). The Correlated Naive Bayes Classifier (CNBC) method recalculates the correlation value for each attribute of the dataset to that class. There are several processes carried out in this study including data acquisition, data labeling, data preprocessing, feature extraction, classifying data using the Correlated Naive Bayes Classifier (CNBC) method, visualizing data, and finally evaluating the results. This study resulted in an accuracy of 82%.*

*Kata Kunci: Klasifikasi sentimen,Correlated Naïve Bayes,Mandalika,MotoGp.*

## 1. PENDAHULUAN

### 1.1. Latar Belakang

Pada tahun 2022 Indonesia telah menjadi tuan rumah penyelenggara kejuaraan olah raga balap motor dunia MotoGp setelah yang terakhir digelar di Sirkuit Sentul Bogor 24 tahun yang lalu kini Sirkuit Mandalika di Lombok Nusa Tenggara Barat menjadi venue representative sekaligus agenda mendongkrak sektor pariwisata dan investasi baru di Indonesia.(Kurniawan 2022)

Tingginya opini masyarakat tentang pergelaran MotoGp di sirkuit Mandalika tersebut sehingga pemerintah harus merancang sebuah konsep pembangunan yang melibatkan masyarakat setempat ketika event internasional itu diselenggarakan dan rencana besar itu tidak boleh mengungtungkan segilitintir orang tapi mengabaikan masyarakat banyak. (https://ekbis.sindonews.com/ 2021)

Klasifikasi sentimen merupakan suatu proses untuk menentukan suatu isi dari dataset yang berbentuk text (kalimat, dokumen, kata, paragraf,dll).(Chandani and Wahono 2015) Klasifikasi sentimen ini bertujuan untuk mengetahui subjektivitas opini, hasil review atau tweet. Berdasarkan klasifikasi sentimen, opini dari seseorang dapat diklasifikasikan ke dalam berbagai kategori diantaranya positif, negatif, atau netral berdasarkan data tekstual.(Dwianto and Sadikin n.d.) Salah satu teknik yang dapat digunakan pada klasifikasi sentimen adalah menggunakan metode Correlated Naïve Bayes Classifier dan Naïve Bayes Classifier.

Metode Correlated Naïve Bayes Classifier adalah pengembangan dari algoritma Naïve bayes Classifier yang sering digunakan dalam mengkategorikan teks dengan peningkatan akurasi sebesar 3% dari metode Naïve Bayes Classier dengan menunjukkan bahwa akurasi metode yang menggunakan metode Naive Bayes Classifier (NBC) adalah 64,33%, sedangkan dari Correlated Naive Bayes Classifier (CNBC) adalah 67,15%(Satya Nugraha, Nurkholis Abdillah, and Innuddin n.d.) Dapat dilihat penggunaan algoritma Correlated Naïve Bayes Classifier memiliki tingkat akurasi yang baik untuk melakukan klasifikasi Maka klasifikasi sentimen menggunakan Correlated Naïve Bayes Classifier (CNBC) merupakan salah satu cara terbaik untuk melihat sentiment masyarakat mengenai antusias terhadap MotoGp yang diselengarakan di Indonesia tahun 2022.

### 1.2. Tinjuan Pustaka

a. Klasifikasi Sentimen

Klasifikasi sentimen adalah kegiatan yang dilakukan untuk melihat tingkat sentimen publik atau opini publik yang berkaitan dengan suatu kegiatan yang diselenggarakan oleh pemerintah. Sentimen adalah istilah yang dapat menggambarkan topik yang objektif dan subjektif serta topik non-faktual ataupun faktual yang memiliki hasil berbeda diantaranya topik positif atau topik negatif. (Wongkar and Angdresey 2019). Klasifikasi sentimen yaitu suatu bidang yang sedang berlangsung dalam penelitian berbasiskan teks. Klasifikasi sentimen memiliki beberapa macam bentuk teks yang berbeda-beda diantaranya seperti seluruh dokumen, paragraf, kalimat atau hanya berupa teks.(Previtali, Arrieta, and Ermanni 2015) Klasifikasi sentimen atau penggalian opini, adalah studi tentang bagaimana suatu entitas memecahkan masalah dari opini publik, sikap, dan emosi entitas

yang dapat mewakili individu, peristiwa, atau masalah.(Rahmasari and Andini n.d. 2021)

b. Klasifikasi Correlated Naïve Bayes

Pengklasifikasi Naïve Bayes yang berkorelasi merupakan pengembangan lebih lanjut dari pengklasifikasi naif Bayes. Korelasi Pengklasifikasi naive Bayes menghitung ulang nilai korelasi (RSquare) antara variabel independen (X) dan variabel dependen (Y). Penambahan parameter korelasi digunakan untuk mengukur derajat hubungan antara variabel bebas (X) dan variabel terikat (Y) dan bilangan lapcian. Bilangan laplacian digunakan untuk menghindari terjadi zero probability. Pada penelitian lain juga menunjukan akurasi terbaik diperoleh metode Correlated Naive Bayes dengan seleksi fitur pada dataset Pima Indian Diabetes sebesar 71,4%, sedangkan pada dataset Thyroid akurasinya sebesar 79,38%. Dengan demikian, penggunaan seleksi fitur Wrapper dapat meningkatkan akurasi metode Correlated Naive Bayes sebesar 4,1% untuk dataset Pima Indian Diabetes dan sebesar 0,48% untuk dataset Thyroid (Hairani and Innuddin n.d.)

Rumus metode *Correlated Naive Bayes Classifier* ditunjukkan pada persamaan 1.

$$P(H|X) = \frac{P(X|H)P(H) * r^2}{P(X)} \quad \text{.....(1)}$$

Keterangan :

$P(H|X)$ : Probabilitas kelas H dari dokumen (*tweet*) yang diinputkan, *posterior probability*

$P(X|H)$ : Probabilitas *term X* dalam kelas H, *conditional probability*

$P(H)$ : Probabilitas kemunculan kelas H dalam dataset, *class* p*rior probability*

$P(X)$ : Probabilitas kemunculan *term* X dalam dataset, *predictor prior probability*

$r$ = merupakan nilai korelasi fitur antar kelasnya.

$r^2$ = *R Square* setiap atribut dari data X berdasarkan kondisi hipotesis Y.

$$r = \frac{n.(\sum(X.Y) - (\sum \boldsymbol{X}).(\sum \boldsymbol{Y})}{\sqrt{(n.\sum \boldsymbol{X^2} - (\sum \boldsymbol{X})^2)}\sqrt{(n.\sum \boldsymbol{Y^2} - (\sum \boldsymbol{Y})^2)}} \quad (2)$$

*R* merupakan *R-Square* fitur antar kelasnya, sedangkan *r* merupakan nilai korelasi fitur antar kelasnya. n merupakan total data pada *dataset*. $\sum XY$ merupakan total perkalian fitur (X) dengan kelasnya (Y). $\sum X$ merupakan total dari fitur X, sedangkan $\sum$Y merupakan total dari fitur Y. $\sum X2$ merupakan total dari fitur X yang dikuadratkan, sedangkan $(\sum X)2$ merupakan kuadrat total fitur X. $\sum Y2$ merupakan total fitur Y yang dikuadratkan, sedangkan $(\sum Y)2$ merupakan kuadrat total fitur Y.

c. Preprocessing Data

Text processing atau pengolah kata bertujuan untuk mengubah dokumen teks yang tidak terstruktur menjadi data terstruktur untuk diproses lebih lanjut. (Amrizal 2018) Fase-fase pengolah kata meliputi:

1. Case folding

Pada tahap case folding akan merubah semua huruf menjadi huruf kecil

2. Tokenization

Pada tahap tokenization memotong string dokumen menjadi kata atau huruf sesuai dengan persyaratan sistem.

3. Filtering

Filtering adalah proses mengambil kata kata penting token berdasarkan stopwords. Stopwords terdiri dari menghilangkan huruf, tanda baca, dan kata umum yang tidak memiliki arti atau informasi yang diperlukan.

4. Stemming

Pada proses stemming akan mengubah token yang asalnya memiliki imbuhan menjadi kata dasar dengan menghilangkan imbuhannya

d. Term Frequency - Inverse Document Ferquency (TF-IDF)

Setelah dilakukan preprocessing dilakukannya feature extraction. Fitur extraction pada penelitian ini menggunakan TF-IDF. Data yang berupa istilah akan diubah ke pada bentuk nomor menggunakan dilakukan proses pembobotan. Dalam fase ini, ada dua bagian proses: TF (Term Frequency) dan IDF (Reverse Document Frequency).(Amrizal 2018) TF adalah jumlah kemunculan setiap kata dalam dokumen. Semakin banyak kata yang terkandung dalam setiap dokumen, semakin tinggi nilai TF. IDF adalah jumlah nilai dokumen untuk setiap kata dan dibandingkan secara terbalik. Artinya, jika sebuah kata jarang muncul dalam sebuah dokumen, nilai IDF akan lebih tinggi dari kata yang sering muncul.

Rumus TF :

$$TF_{ij} = tf_{ij} \quad (1) \text{ next } \text{.....(3)}$$

Keterangan:

$TF_{ij}$ : Nilai *Term Frequency* pada *term i* dalam dokumen *j*

$tf_{ij}$ : Frekuensi kemunculan *term i* dalam dokumen *j*

Rumus IDF :

$$IDF_{ij} = \log\left(\frac{N}{n_i}\right) + 1 \text{ …….(4)}$$

Keterangan :

$IDF_i$ : Nilai *Invers Document Frequency* pada *term i* dalam dokumen *j*

$N$ : Jumlah keseluruhan dokumen (*tweet*) dalam dataset

$n_i$ : Jumlah dokumen yang memiliki kemunculan *term i*

Sehingga didapat untuk persamaan dari TF-IDF adalah sebagai berikut :

$$TF_{ij} - IDF_{ij} = tf_{ij}\ x\ \log\left(\frac{N}{n_i}\right) + 1 \text{ …….(5)}$$

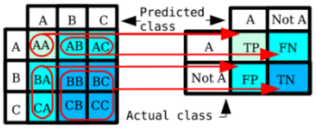
e. Confusion Matrix

Confusion Matrix merupakan sebuah metode untuk evaluasi yang menggunakan tabel matrix seperti pada Gambar 2.1.

| | Prediksi | |
|---|---|---|
| Aktual | TRUE | FALSE |
| TRUE | TP | FP |
| FALSE | FN | TN |

Gambar 1 . 1 Confusion Matrix

Pada gambar 1.1. dapat kita lihat bahwa jika dataset terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif. kemudian jika terdapat terdapat kasus confusiom matrix berukuran 3 * 3 maka dapat dirubah menjadi berukuran 2 * 2 dengan kelas ‘A’ sebagai kelas positif dan kelas ‘Not A’ sebagai kelas negatif. Dapat dilihat pada gamabar dibawah ini.




Gambar 1 . 2 Confusion Matrix 3 * 3

Confusion Matrix ini biasa digunakan untuk mengukur kinerja dari machine learning dan dapat digunakan sebagai alat visual untuk mengevaluasi hasil klasifikasi (Nabillah, Alam, and Resmi 2022). Confusion matrix merupakan indikator performa untuk masalah klasifikasi machine learning yang outputnya bisa di lebih dari satu kelas. Confusion matrix ini adalah tabel yang berisi empat kombinasi nilai prediksi dan nilai aktual yang berbeda.

Ada empat istilah yang merupakan representasi hasil proses klasifikasi pada confusion matrix dapat dilihat Pada Gambar 2.1 dan Gambar 1.2. True Positive (TP) menyatakan data positif yang diprediksi benar, True Negatif (TN) menyatakan data negatif yang diprediksi benar. kemudian False Positive (FP) sebagai kesalahan tipe 1 merupakan data negatif namun diprediksi sebagai data positif, sebaliknya False Negatif (FN) sebagai kesalahan tipe 2 merupakan data postif namun diprediksi sebagai data negatif. Nilai Prediksi adalah keluaran dari program dimana nilainya positif dan negatif sedangkan Nilai Aktual adalah nilai sebenarnya dimana nilainya true dan false. Selain itu kita dapat menghitung nilai tersebut di antaranya Accuracy, Precision, Recall, dan F1-Score (Nabillah, Alam, and Resmi 2022).

$$\text{Accuracy} = \frac{TN+TP}{TP+FP+FN+TN} \quad (6)$$

Accuracy 2 kelas: Untuk mengukur tingakat akurasi dari pengklasifikasi data yang akan dievaluasi pada 2 kelas . [24]

$$Precision = \frac{TP}{TP+FP} (7)$$

Presisi: Ini adalah rasio jumlah sentimen yang diprediksi secara akurat dengan jumlah total sentimen yang diprediksi. [24]

$$Recall = \frac{TP}{TP+FN} (8)$$

Recall : Ini adalah rasio jumlah sentimen yang diprediksi secara akurat dengan jumlah total sentimen aktual. [24]

$$F-1\ Score = \frac{2*(Recall*Precision)}{Recall+Precision} (9)$$

F-1 Score: ini merupakan perbandingan rata-rata precision dan recall yang dibobotkan. Accuracy tepat kita gunakan sebagai acuan performansi algoritma ika kumpulan data berisi jjumlah data False Negatif dan False Positif yang sangat mirip (simetris), gunakan akurasi yang akurat sebagai ukuran kinerja algoritme. Namun, jika angkanya tidak dekat, Anda harus menggunakan skor F1 sebagai referensi.

$$Accuracy = \frac{TP\ tiap\ kelas}{Total\ Data} \quad (10)$$

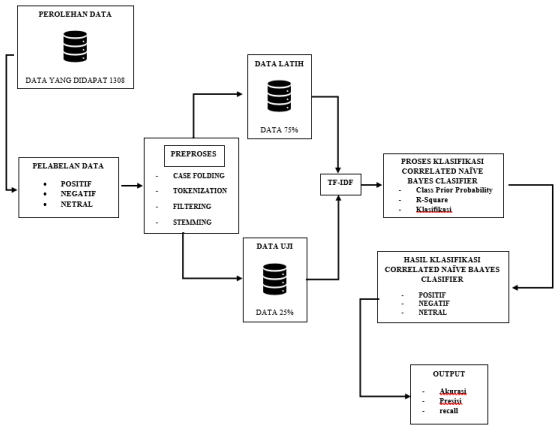
Accuracy 3 kelas : Untuk mengukur tingakat akurasi dari pengklasifikasi data yang akan dievaluasi pada 3 kelas

f. K-Fold CrossValidation

K-fold adalah salah satu metode Cross Validation yang populer dengan melipat data sebanyak k dan mengulangi (iterasi) eksperimennya sebanyak k juga (Setyo Nugroho, Kuncahyo & Istiadi, Istiadi & Marisa, Fitri. (2020).

### 1.3. Metodologi Penelitian

Dalam metode penelitian tugas akhir ini akan diuraikan kedalam beberapa bagian seperti yang tertera pada gambar alur penelitian berikut




Ada beberapa proses yang dilakukan pada penelitian ini diantaranya perolehan data, pelabelan data, preprocessing data, feature extraction(TF-IDF), mengklasifikasikan data menggunakan metode Correlated Naive Bayes Classifier (CNBC) dan evaluasi hasil menggunakan Confusion Matrix dan K-Fold Crossvalidation.

## 2. PEMBAHASAN

A. Perolehan Data

Perolehan data yang digunakan pada penelitian ini diambil dari kumpulan tweet yang berasal dari data public yang ada di twitter. Data tweet ini diperoleh dengan menggunakan proses scraping data menggunakan tools rapid miner Data yang diambil sebanyak 1308 data dengan kata kunci “MotoGp Indonesia” dan “Mandalika” yang dimulai dari tanggal 1 Maret 2022 sampai 30 juni 2022. dilakukannya filter pada data yang telah didapat agar tidak adanya data yang sama atau duplikat.

B. Pelabelan Data

Pelabelan data merupakan tahapan yang digunakan untuk memberi kelas atau label pada data tweet yang mentah agar dapat digunakan untuk proses selanjutnya. Pelabelan ini dibedakan dalam beberapa kategori diantaranya positf negatif ataupun netral. Pelabelan dilakukan oleh seorang ahli dibidangnya secara manual atau satu persatu dengan 3 orang narasumber untuk menentukan makna dari setiap tweetnya

C. Preprocessing Data

Text processing atau pengolah kata bertujuan untuk mengubah dokumen teks yang tidak terstruktur menjadi data terstruktur untuk diproses lebih lanjut. Fase-fase pengolah kata meliputi Case folding pada tahap case folding akan merubah semua huruf menjadi huruf kecil selanjutnya tahap *tokenization* memotong string dokumen menjadi kata atau huruf sesuai dengan persyaratan sistem, lalu tahap *filtering* adalah proses mengambil kata kata penting token berdasarkan *stopwords. stopwords* terdiri dari menghilangkan huruf, tanda baca, dan kata umum yang tidak memiliki arti atau informasi yang diperlukan dan terakhir proses *stemming* akan mengubah token yang asalnya memiliki imbuhan menjadi kata dasar dengan menghilangkan imbuhannya.

D. Pembagian Data

Pada tahap ini setelah dilakukan tahap prepocessing data, data lalu akan di bagi menjadi 2 bagian yaitu data latih dan data uji dengan pembagian data latih sebesar 75% dan data uji sebesar 25%

E. Feature Extraction (TF-IDF)

Dalam pembobotan TF-IDF terdapat 3 langkah yang harus dilalui diantaranya mendapatkan vector dari hasil perhitungan TF-IDF, mendapatkan vector nilai dari hasil perhitungan IDF dan yang terakhir mengalikan hasil dari perhitungan TF dan IDF. Dataset yang akan digunakan dalam proses TF-IDF adalah data yang telah melalui tahapan praproses

F. Klasifikasi Correlated Naïve Bayes

Setelah mendapatkan hasil pembobotan TF-IDF maka tahap selanjutnya yaitu melakukan proses klasifikasi dengan menggunakan metode Correlated Naïve Bayes. Pada proses klasifikasi ini terdapat tiga tahapan untuk mendapatkan nilai probabilitas tertinggi berdasarakan data latih yang dimiliki dan menabahkan nilai korelasi, korelasi disini yaitu dengan menambahkan paramater korelasi antar atribut terhadap kelas. Dengan memperhitungkan nilai korelasi dari masing-masing atribut vektor X terhadap kelas Y. Untuk hasil klasifikasi Correlated naïve bayes dapat dilihat pada tabel 1 Conditional probability berikut

**Tabel 1 Conditional Probability**

| Conditional probability | Hasil |
|---|---|
| $\prod_{k=1}^{n=3} P(x_k \mid H_{Positif})$ | $6{,}795309 * 10^{14}$ |
| $\prod_{k=1}^{n=3} P(x_k \mid H_{negatif})$ | $1{,}535956 * 10^{31}$ |
| $\prod_{k=1}^{n=3} P(x_k \mid H_{Netral})$ | $2{,}12985 * 10^{31}$ |

**Tabel 2 Posterior probability**

| **Posterior probability** | **Class prior probability * Conditional probability** | **Hasil *Correlated Naïve Bayes*** | **Hasil *Naïve Bayes*** |
|---|---|---|---|
| $P(H_{Positif}\|X$ | 0.4 ∗6,795309 ∗ $10^{14}$ | **2,718123* 10$^{14}$** | 2,12445 6 * $10^{10}$ |
| $P(H_{Netral}\|X$ | 0.4 ∗ 7,633912 ∗ $10^{13}$ | 3,053565 ∗ $10^{13}$ | 2,76029 6 * $10^{9}$ |
| $P(H_{Negatif}\|$ | 0.2 ∗ 3,943998 ∗ $10^{14}$ | 7,887995 ∗ $10^{13}$ | 2,10411 9 * $10^{9}$ |

G. Evaluasi Hasil

Pada tahap ini dilakukan pengukuran kinerja suatu model terhadap pengujian data setiap kelas yang bertujuan untuk mengetahui hasil akurasi terbaik dari setiap kelas. Pengujian akurasi merupakan persentase dari total data yang diidentifikasi dan dinilai benar menggunakan Confusion Matrix berbentuk tabel matriks

**Tabel 3 Confusion Matriks**

| # | Positif | Negatif | Netral | X |
|---|---|---|---|---|
| Positif | 148 True Positif Positif | 5 Positif False Negatif | 29 Positif False Netral | 182 |
| Negatif | 3 Negatif False Positif | 43 True Negatif Negatif | 2 Negatif False Netral | 48 |
| Netral | 17 Netral False Positif | 5 Netral False Negatif | 74 True Netral Netral | 96 |
| Total | 168 | 53 | 105 | 326 |

Dari proses confusion matriks diatas diperoleh hasil akurasi, precision, recall dan f1-score seperti gambar 1 berikut

| | |
|---|---|
| Akurasi | 81.29 % |
| Presisi | 88.10 % |
| Recall | 81.318681318681 % |
| F1 | 84.57 % |

**Gambar 1 Hasil Confusion Matriks**

Nilai rata-rata precision yang diperoleh adalah :

$$\frac{148}{148+43+17}\frac{129}{155} = 0{,}88 * 100 = \mathbf{88{,}10\%}$$

Nilai rata-rata recall yang diperoleh adalah :

$$\frac{148}{148+5+27}\frac{\mathbf{129}}{\mathbf{158}} = 0{,}81 * 100 = \mathbf{81{,}31\%}$$

Nilai rata-rata f1-score yang diperoleh adalah :

$$\frac{2*(0{,}81*0{,}88)}{0{,}81+0{,}88}\frac{128}{169} = 84*100 = \mathbf{84{,}54\%}$$

Nilai rata-rata akurasi Correlated naïve bayes yang diperoleh adalah :

$$\frac{\mathbf{129+9+73}}{\mathbf{261}} = 0{,}80 * 100 = \mathbf{81{,}29\%}$$

Dari Hasil tabel perhitungan Akurasi, Precision, Recall dan F1-Score diatas dapat disimpulkan bahwa tingkat akurasi menghasilkan akurasi sebesar 81,29%. Kemudian untuk data uji yang digunakan sebanyak 326 data tweet dengan pembagian untuk kelas (Negatif) sebanyak 48 data, untuk kelas (Positif) sebanyak 182 data dan untuk kelas (Netral) sebanyak 96.

## 3. ALGORITMA ATAU PROGRAM

H. K-Fold CrossValidation

K-fold adalah salah satu metode Cross Validation yang populer dengan melipat data sebanyak k dan mengulangi (iterasi) eksperimennya sebanyak k juga. Untuk menghitung Cross Validation atau representasi hasil proses klasifikasi dari beberapa data uji dapat dijelaskan dari total dataset sebesar 1308 dataset dibagi menjadi 4 bagian atau k=4 data uji diantaranya, Data1,2,3 dan 4 berisi 327 record dengan data latih toal 981record dengan pembagian data uji sebesar 25% dan untuk data latih sebesar 75% dari total dataset. Pengujian menggunakan data yang sudah dipartisi akan diulang sebanyak 4 kali (k=4) dengan posisi data tes berbeda disetiap iterasinya. Misalkan iterasi pertama data tes pada posisi awal, iterasi kedua data tes di posisi kedua begitu seterusnya. Hingga iterasi keempat.

| | | | | | |
|---|---|---|---|---|---|
| DATA1 | 327 | 327 | 327 | 327 | 1308 |
| | | | | | |
| DATA2 | 327 | 327 | 327 | 327 | 1308 |
| | | | | | |
| DATA3 | 327 | 327 | 327 | 327 | 1308 |
| | | | | | |
| DATA4 | 327 | 327 | 327 | 327 | 1308 |

= Data Uji
= Data Latih

Untuk pengujian fold ke-1 sampai fold ke-4 jumlah data tes yang dimasukkan adalah 327 *tweet* dengan posisi data tes dapat dilihat pada tabel 4 hingga 7

**Tabel 4 Iterasi K-Fold Cross Validation 1**

| **Data uji** | **Akurasi** |
|---|---|

| | |
|---|---|
| Data1 | 81,29% |

**Tabel 5 Iterasi K-Fold Cross Validation 2**

| **Data uji** | **Akurasi** |
|---|---|
| Data2 | 79,72% |

**Tabel 6 Iterasi K-Fold Cross Validation 3**

| **Data uji** | **Akurasi** |
|---|---|
| Data3 | 86,50% |

**Tabel 7 Iterasi K-Fold Cross Validation 4**

| **Data uji** | **Akurasi** |
|---|---|
| Data4 | 83,79% |

Untuk hasil pengujian keseluruhan *fold* dapat dilihat pada Tabel 8

**Tabel 8 Hasil K-Fold Cross Validation**

| **Data uji** | **Akurasi** |
|---|---|
| Data1 | 81,29% |
| Data2 | 79,72% |
| Data3 | 86.50% |
| Data4 | 83,79% |
| **Average** | 82,82% |

Dari Hasil tabel perhitungan 4-Fold CrossValidation, diatas akurasi yang didapat dari pengujian data tes yang tertinggi pada saat pengujian fold ke-3 yaitu 86,50%, diikuti fold ke-4 dengan akurasi sebesar 83,79%, fold ke-1 dengan akurasi sebesar 81,29% dan akurasi terendah pada fold ke-2 dengan akurasi 79,72 dengan rata-rata akurasi pengujian keempat fold adalah 82,82%.

Dapat disimpulkan bahwa tingkat akurasi mengalami kenaikan jika menggunakan pengujian 4-Fold CrossValidation menghasilkan akurasi dengan rata-rata sebesar 82,82%. Mengalami peningkatan akurasi sebesar 1,53% dibanding tidak menggunakan CrossValidation

## 4. KESIMPULAN

A. Kesimpulan

Berdasarkan hasil penelitian tugas akhir, dapat disimpulkan bahwa opini publik cenderung melakukan tweet positif tentang pergelaran MotoGp di Indonesia. Model klasifikasi Correlated Naïve Bayes dalam penelitian ini mencapai skor akurasi (82%). Data yang digunakan sebesar 1308 data tweet yang dikumpulkan secara langsung menggunakan teknik scraping menggunakan tools Rapid miner. Masyarakat cenderung melakukan tweet sentimen positif adalah (45%) tweet, sedangkan sekitar (22%) sentimen negatif, dan (31%) melakukan tweet dengan sentimen netral. Dengan tingkat akurasi sebesar (82%) menunjukan bahwa sentimen analisis menggunakan metode Correlated Naïve Bayes dengan data tweet MotoGp di Indonesia cukup baik akan tetapi karena dataset memiliki jumlah sentimen positif yang lebih mendominasi oleh karena itu hasil yang didapat lebih kecil dari penelitian sebelumnya.

B. Saran

Berdasarkan hasil penelitian ini, ada beberapa saran untuk pengembangan lebih lanjut, diantaranya:

1. Menggunakan data yang lebih banyak untuk data latih agar akurasinya lebih akurat.

2.Melakukan filter data agar data seimbang.

## PUSTAKA

Amrizal, Victor. 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: HADITS SHAHIH BUKHARI-MUSLIM)." *JURNAL TEKNIK INFORMATIKA* 11(2): 149–64.

Chandani, Vinita, and Romi Satria Wahono. 2015. "Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection Pada Analisis Sentimen Review Film." *Journal of Intelligent Systems* 1(1). http://journal.ilmukomputer.org.

Dwianto, Enos, and Mujiono Sadikin. *Analisis Sentimen Transportasi Online Pada Twitter Menggunakan Metode Klasifikasi Naïve Bayes Dan Support Vector Machine*.

Hairani, Data Kesehatan, and Dan Muhammad Innuddin. *Kombinasi Metode Correlated Naive Bayes Dan Metode Seleksi Fitur Wrapper Untuk Klasifikasi Data Kesehatan*.

https://ekbis.sindonews.com/. 2021. "Sirkuit Mandalika Diyakini Bangkitkan Ekonomi RI." https://ekbis.sindonews.com/read/460412/34/sirkuit-mandalika-diyakini-bangkitkan-ekonomi-ri-selengkapnya-pride-of-indonesia-sabtu-pukul-1800-wib-1624079209 (January 7, 2022).

Kurniawan, Agung. 2022. "Jadwal MotoGP 2022 - Debut Indonesia, Mandalika Jadi Yang Pertama Di Asia Tenggara." https://www.bolasport.com/read/313075080/jadwal-motogp-2022-debut-indonesia-mandalika-jadi-yang-pertama-di-asia-

tenggara (January 7, 2022).

Nabillah, Asyfah, Syariful Alam, and Mochzen Gito Resmi. 2022. “Twitter User Sentiment Analysis Of TIX ID Applications Using Support Vector Machine Algorithm.” 3(1): 14–27.

Previtali, Francesco, Andres F. Arrieta, and Paolo Ermanni. 2015. “Double-Walled Corrugated Structure for Bending-Stiff Anisotropic Morphing Skins.” *Journal of Intelligent Material Systems and Structures* 26(5): 599–613.

Rahmasari, Gina, and Rizkiki Andini. “ANALISIS RESPON MASYARAKAT PADA PLATFORM MEDIA SOSIAL TWITTER TERHADAP TOKOH POLITIK, JENDERAL TNI (PURN.) GATOT NURMANTYO.”

Satya Nugraha, Gibran, Mokhammad Nurkholis Abdillah, and Muhammad Innuddin. *KOMPARASI AKURASI METODE CORRELATED NAIVE BAYES CLASSIFIER DAN NAIVE BAYES CLASSIFIER UNTUK DIAGNOSIS PENYAKIT DIABETES*.

Setyo Nugroho, Kuncahyo & Istiadi, Istiadi & Marisa, Fitri. (2020). Optimasi naive Bayes classifier untuk klasifikasi teks pada e-government menggunakan particle swarm optimization. Jurnal Teknologi dan Sistem Komputer. 8. 21-26. 10.14710/jtsiskom.8.1.2020.21-26

Wongkar, Meylan, and Apriandy Angdresey. 2019. “Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter.” In *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*, Institute of Electrical and Electronics Engineers Inc.