

# PENERAPAN ALGORITMA *NAÏVE BAYES* DALAM ANALISIS SENTIMEN PENGGUNA TWITTER TERHADAP APLIKASI M-TIX 21 CINEPLEX

Reza Nur Rahman<sup>1\*</sup>, Yuliazmi<sup>2</sup>, Ady Widjaja<sup>3</sup>, Rusdah<sup>4</sup>

<sup>1,2,3,4</sup> Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, DKI Jakarta, Indonesia

Email: <sup>1\*</sup>1712500972@student.budiluhur.ac.id, <sup>2</sup>yuliazmi@budiluhur.ac.id, <sup>3</sup>ady.widjaja@budiluhur.ac.id, <sup>4</sup>rusdah@budiluhur.ac.id  
(\* : corresponding author)

**Abstrak**-M-Tix merupakan layanan pembelian tiket bioskop dari 21 Cineplex yang dimana ini merupakan layanan jarak jauh yang membuat konsumen membeli tiket bioskop tanpa harus mengantri dan dapat dilakukan dimana saja. Pada penelitian ini penulis bermaksud menganalisis sentimen terhadap M-tix 21 Cineplex. Lalu dijadikan sebagai bahan identifikasi opini negatif dan positif dengan menggunakan algoritma *naïve bayes* untuk mengetahui performa pelayanan dari aplikasi M-tix 21 Cineplex. Peneliti juga menggunakan CRISP-DM (*Cross-Industry Standard Process for Data Mining*) sebagai metode pengolahan data serta menggunakan aplikasi *rapidminer* untuk mendapatkan, mengolah dan menghasilkan klasifikasi positif dan negatif. Dalam penelitian ini mengambil 749 tweet yang membahas tentang M-tix 21 Cineplex pada media Twitter dengan menggunakan keyword "Mtix". Hasil dari analisis komparasi algoritma klasifikasi *data mining* yaitu KNN (*K-Nearest Neighbor*) dan *Naïve Bayes* menghasilkan *Naïve Bayes* menghasilkan nilai uji AUC (*Area Under Curve*) dan akurasi yang lebih baik dibandingkan dengan *K-Nearest Neighbor* yang menjadikan *Naïve Bayes* sebagai model terbaik untuk penelitian ini. Hasil dari penelitian ini menghasilkan nilai sentiment positif sebanyak 59% dan negatif sejumlah 41%. *Accuracy* tertinggi yaitu 65.29%, yang dihasilkan dengan pemodelan menggunakan algoritma *Naïve Bayes* yang menggunakan perbandingan 80:20 untuk data *training* dan data *testing*.

**Kata Kunci:** Analisis Sentimen, Penambangan Teks, Aplikasi M-Tix 21, Twitter, *Naïve Bayes*.

## *Application of the Naïve Bayes Algorithm in the Analysis of Twitter User Sentiments for the M-Tix 21 Cineplex Application*

**Abstract**-M-Tix is a cinema ticket purchase service made by 21 Cineplex which is a long-distance service that allows consumers to buy cinema tickets without having to queue and can conduct it anywhere. In this study the authors intend to analyze sentiment towards M-tix 21 Cineplex. Then used as material for identifying negative and positive opinions using the *Naïve Bayes* algorithm to determine the service performance of the M-tix 21 Cineplex application. Researchers also use CRISP-DM (*Cross-Industry Standard Process for Data Mining*) as a data processing method and use the *rapidminer* application to obtain, process and produce positive and negative classifications. In this study, we took 749 tweets that discussed the M-tix 21 Cineplex on Twitter using the keyword "Mtix". The results of the comparative analysis of the data mining classification algorithms, namely KNN (*K-Nearest Neighbor*) and *Naïve Bayes*, resulted in *Naïve Bayes* producing AUC (*Area Under Curve*) test values and better accuracy compared to *K-Nearest Neighbor*, which made *Naïve Bayes* the best model for this study. The results of this study resulted in positive sentiment values of 59% and negative sentiments of 41%. The highest accuracy is 65.29%, which is generated by modeling using the *Naïve Bayes* algorithm which uses a 80:20 ratio for training data and testing data.

**Keywords:** *Sentiment Analysis, Text Mining, M-Tix 21 Application, Twitter, Naïve Bayes.*

## 1. PENDAHULUAN

Seiring kemajuan teknologi, 21 Cineplex mulai mengembangkan sebuah inovasi yang dapat mempermudah pemesanan tiket bioskop melalui sistem pembelian tiket bioskop online menggunakan aplikasi Mtix. Inovasi ini diimplementasikan oleh 21 Cineplex berdasarkan berjalannya waktu banyak orang yang mulai mengenal dan menggunakan teknologi internet. Tentunya dengan inovasi tersebut membuat banyak orang memperbincangkan aplikasi ini di berbagai media sosial salah satunya Twitter. Salah satunya adalah mengenai aplikasi yang sering error ataupun daftar bioskop yang terlalu sedikit.

Twitter merupakan *platform* media sosial yang disukai oleh orang-orang di seluruh dunia. Menilik hasil survei yang dilakukan Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) Triwulan 2019-2020, jumlah pengguna internet di Indonesia mencapai 196,7 juta orang. Angka itu meningkat 23,5 juta atau 8,9% dibandingkan tahun 2018. Data laporan keuangan Twitter kuartal III 2019 menunjukkan jumlah pengguna aktif harian di platform Twitter meningkat 17% menjadi 145 juta pengguna [1].

Tweet yang dikumpulkan dan kemudian dianalisis disebut analisis sentimen. Analisis sentimen adalah subbidang penelitian opini publik. Tweet yang dikumpulkan dan kemudian dianalisis disebut analisis sentimen. Ketika analisis opini dilakukan gunanya untuk melihat opini tentang sesuatu, atau bisa juga digunakan untuk mengidentifikasi tren pasar karena mampu mengambil opini umum tentang topik, produk, atau layanan tertentu yang disematkan dalam teks yang tidak terstruktur.

Penelitian ini membatasi analisis opini pengguna twitter terhadap Mtix dari 21Cineplex yang terdapat pada media sosial yaitu twitter. Menggunakan metode Naïve Bayes dengan dataset berupa teks Tweet dengan kata kunci “mtix”. sehingga dapat diketahui seberapa besar sentiment positif atau negative dari masyarakat terhadap aplikasi Mtix dari 21Cineplex tersebut. Dari penelitian ini, dapat diharapkan menjadi salah satu pengetahuan dan acuan masyarakat serta untuk developer tentang aplikasi Mtix dari 21Cineplex tersebut agar dapat menjadi lebih baik lagi.

Dari hasil proses analisis pada perbandingan algoritma klasifikasi terhadap Genose pada Twitter [2], diperoleh hasil kinerja algoritma yang cukup baik yaitu pada algoritma Naïve Bayes. Hasil prediksi pada *Naïve Bayes* cukup akurat dibandingkan dengan algoritma pembandingnya *K-Nearest Neighbor* dan *Decision Tree*. Dengan perolehan nilai *accuracy* sebesar 72.36%, *precision* 66.91%, *recall* 77.97% pada algoritma *Naïve Bayes* sedangkan kedua algoritma lainnya mendapatkan nilai *accuracy*, *precision* dan *recall* yang lebih rendah.

Pada analisis kesuksesan pada aplikasi M-Tix Cinema 21 di kota Malang, yang dimana pada analisis ini didasari oleh perspektif dari pengguna dilingkungan Kota Malang [3]. Pendekatan yang digunakan adalah *Delone and McLean Success Model*. Dengan pendekatan tersebut menunjukkan bahwa aplikasi M-tix Cinema 21 pada variable *System Quality* dan Variable *Use* mendapatkan nilai 75,8 dan 75,0% yang dimana nilai tersebut masuk kedalam nilai dengan kategori tinggi. Dengan dilandasi nilai perbandingan rata-rata dari persetase setiap variable, variable *Use* diprioritaskan untuk menjadi rekomendasi untuk perbaikan. Dari rekomendasi tersebut nantinya dapat meningkatkan variable *System Quality* maupun variable *Use* dengan *maintenance system* secara berkala serta diimbangi hasil yang maksimal dari strategi pemasaran aplikasi *mobile* M-Tix Cinema 21.

Pada penelitian analisis *sentiment* pada Tix ID dengan menggunakan algoritma *Support Vector Machine* (SVM) pada twitter [4], menyimpulkan bahwa akurasi tertinggi didapatkan dengan menggunakan *kernel dot* dan *scenario* pembagian data 80:20. Dari hal tersebut didapatkan nilai akurasi sebesar 74,17%, yang dimana dengan nilai tersebut disimpulkan bahwa algoritma tersebut memiliki nilai yang cukup baik. Dari proses algoritma tersebut menghasilkan nilai 146 sentimen positif dan 231 sentimen negatif.

## 2. METODE PENELITIAN

Metode penelitian berisi tentang teori-teori yang berkaitan dengan penelitian seperti analisis sentiment, *text mining*, *preprocessing*, *naïve bayes classification*, *cross validation*, dan CRISP-DM.

### 2.1 Analisis Sentimen

Analisis sentimen atau yang biasa dikenal dengan opinion mining merupakan salah satu bidang *text mining* yang tujuannya untuk mengetahui persepsi atau subjektivitas publik (*audience*) terhadap suatu topik pembahasan, peristiwa atau isu. Analisis sentimen mengevaluasi pendapat dan tren pendapat tentang suatu topik baik secara negatif maupun positif.

### 2.2 Text Mining

*Text mining* dapat juga diartikan sebagai penemuan informasi yang baru dan tidak diketahui sebelumnya oleh computer, dengan begitu secara otomatis dapat mengekstrak informasi dari sumber-sumber yang berbeda [5]. Adapun tugas khusus dari *text mining* antara lain, yaitu: pengkategorisasian teks (*text categorization*) dan pengelompokan teks (*text clustering*).

### 2.3 Preprocessing

Sebelum diproses *Data Mining* seringkali diperlukan sebelum proses penambangan data. *Data Preprocessing* menjelaskan jenis proses yang memproses data mentah sebagai persiapan untuk pemrosesan lainnya. Tujuannya adalah untuk mengonversi data ke format yang lebih mudah dan efisien bagi pengguna dalam beberapa langkah, antara lain:

- Remove Duplicates* adalah proses penghapusan data yang muncul secara berulang
- Replace* adalah proses untuk menghilangkan *link*, *symbol*, dan *mention* seperti *retweet*, *https*, dan *@username* yang terdapat pada *tweet*.
- Case Folding* adalah proses mengubah setiap huruf didalam post menjadi lowercase atau semuanya dijadikan huruf kecil.

- d. *Tokenizing* adalah proses tahapan pemotong *string* masukkan berdasarkan kata-kata yang menyusunnya atau dengan kata lain pemecahan kalimat menjadi kata.
- e. *Stopwords Removing* adalah proses menghapus kata-kata yang sangat umum contohnya seperti : dan, di, itu, dengan.
- f. *Stemming* merupakan proses mengubah kata atau term ke bentuk akar katanya (steam).
- g. *Term Weightting (TF-IDF)* merupakan suatu cara untuk memberikan bobot hubungan suatu kata (Term)

## 2.4 Naïve Bayes Classification

*Naive Bayes Classifier* (NBC) adalah metode pengenalan teks dan dokumen. Seperti pada salah satu kasus sebelumnya, metode ini digunakan untuk mempelajari klasifikasi pengaduan tentang jaringan. NBC merupakan klasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas suatu kelas dan kelebihan dari metode ini adalah akurasi yang tinggi dan waktu perhitungan yang lebih cepat [6].

Pada algoritma Naïve Bayesian, setiap dokumen direpresentasikan dengan menginputkan “a1, a2, a3,..., an” dimana nilai a1 adalah kata pertama dan kata berikutnya datang, sedangkan V adalah *class identifier*. Langkah selanjutnya adalah mencari nilai maksimum atau nilai tertinggi dari kategori teks yang diuji (VMAP). Persamaan VMAP dalam rumus (1) sebagai berikut:

$$VMAP = v_i \in \text{Vargmax} P_{v_j} \cap i P_{a_i v_j} \quad (1)$$

Nilai P(vj) dihitung pada saat data latih, dengan rumus 2.2 sebagai berikut :

$$P_{v_j} = \frac{|Docs_j|}{|training|} \cdot |docs_j| \quad (2)$$

adalah dokumen yang memiliki kategori j pada dokumen latih dan |training| adalah jumlah dokumen latih.



$$P_{a_i v_j} = \frac{|n_i + 1|}{|n + kosakata|} \quad (3)$$

1.  $N_i$  : Jumlah kemunculan kata  $a_i$  pada dokumen yang berkategori  $v_j$
2.  $n$  : jumlah seluruh kata pada dokumen yang berkategori  $v_j$
3. Kosakata: jumlah kata pada seluruh dokumen latih

## 2.5 Cross Validation

*Cross Validation* adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja suatu model atau algoritma, dimana data dibagi menjadi dua subset, yaitu data proses pembelajaran dan data validasi/evaluasi. *Cross Validation* berarti membagi dataset menjadi dua bagian, satu digunakan sebagai data pelatihan dan yang lainnya digunakan sebagai data uji [7].

|              | Dataset dibagi menjadi 10 bagian secara <i>random</i> (acak) |     |     |     |     |     |     |     |     |     | Akurasi |      |
|--------------|--|-----|-----|-----|-----|-----|-----|-----|-----|-----|---------|------|
|              | 10%  | 10% | 10% | 10% | 10% | 10% | 10% | 10% | 10% | 10% |         | 100% |
| Percobaan 1  | 10%  |     |     |     |     |     |     |     |     |     |         | a1   |
| Percobaan 2  |  | 10% |     |     |     |     |     |     |     |     |         | a2   |
| Percobaan 3  |  |     | 10% |     |     |     |     |     |     |     |         | a3   |
| Percobaan 4  |  |     |     | 10% |     |     |     |     |     |     |         | a4   |
| Percobaan 5  |  |     |     |     | 10% |     |     |     |     |     |         | a5   |
| Percobaan 6  |  |     |     |     |     | 10% |     |     |     |     |         | a6   |
| Percobaan 7  |  |     |     |     |     |     | 10% |     |     |     |         | a7   |
| Percobaan 8  |  |     |     |     |     |     |     | 10% |     |     |         | a8   |
| Percobaan 9  |  |     |     |     |     |     |     |     | 10% |     |         | a9   |
| Percobaan 10 |  |     |     |     |     |     |     |     |     | 10% |         | a10  |

Keterangan gambar:  
 = data testing  
 = data training

Gambar 1 10 Folds Cross Validation

## 2.6 Confusion Matrix

*Confusion matrix* adalah metode yang biasa digunakan untuk menghitung akurasi, *recall*, *precision* dan tingkat kesalahan. Sedangkan akurasi mengevaluasi kemampuan sistem untuk menemukan semua item yang

relevan dalam kumpulan dokumen dan didefinisikan sebagai persentase dokumen yang relevan dengan kueri. Akurasi mengacu pada perbandingan kasus yang teridentifikasi dengan benar dengan jumlah total kasus, dan tingkat kesalahan adalah jumlah kasus yang salah diidentifikasi dengan jumlah total kasus [8].

|                  |              | Actual Values |              |
|------------------|--------------|---------------|--------------|
|                  |              | Positive (1)  | Negative (0) |
| Predicted Values | Positive (1) | TP            | FP           |
|                  | Negative (0) | FN            | TN           |

**Gambar 2** Confusion Matrix

## 2.7 CRISP-DM

*Cross-Industry Standard Process for Data Mining* (CRISP-DM) dikembangkan pada tahun 1996 oleh analis dari berbagai industri seperti Daimler Chrysler, SPSS dan NCR. CRISP-DM menyediakan proses penambangan data standar sebagai strategi pemecahan masalah umum untuk setiap entitas bisnis atau penelitian. Dalam CRISP-DM, siklus hidup proyek data mining dibagi menjadi enam fase. Semua langkah sekuensial yang ada bersifat adaptif. Langkah selanjutnya dalam siklus tergantung pada output dari langkah sebelumnya [9].

## 3. HASIL DAN PEMBAHASAN

Pada hasil dan pembahasan berfokus pada proses bagaimana penelitian ini berjalan yang disesuaikan tahapan penelitian sesuai dengan CRISP-DM, dimana dari proses pengumpulan data sampai dengan proses evaluasi.

### 3.1 Data Penelitian

Data yang digunakan dalam bahan penelitian ini dikumpulkan dari Twitter sejak 9 Desember 2022 hingga 23 Desember 2022, kemudian dengan tambahan data sejak 22 Januari hingga 1 Februari 2023 dengan total 176 tanggal. Dari rentan waktu tersebut mendapatkan data dengan total sebanyak 749 *Record*, dengan melakukan proses *crawling* menggunakan *Rapidminer*. *Dataset* ini menggunakan kata kunci “mtix” yang merupakan kata kunci keterkaitan *tweet* dengan topik yang diambil yaitu tentang aplikasi Mtix dari 21 Cineplex.

### 3.2 Data Preparation

Pada fase *Data Preparation*, penulis akan melakukan pembersihan dan penyeleksian data dari hasil data yang telah di *Crawling* dengan beberapa tahapan seperti *cleansing*, *case folding*, *tokenizing*, *remove stopword*, *stemming* dan *term-weighting*. Dimana disini penulis membaginya menjadi 4 bagian sebagai berikut :

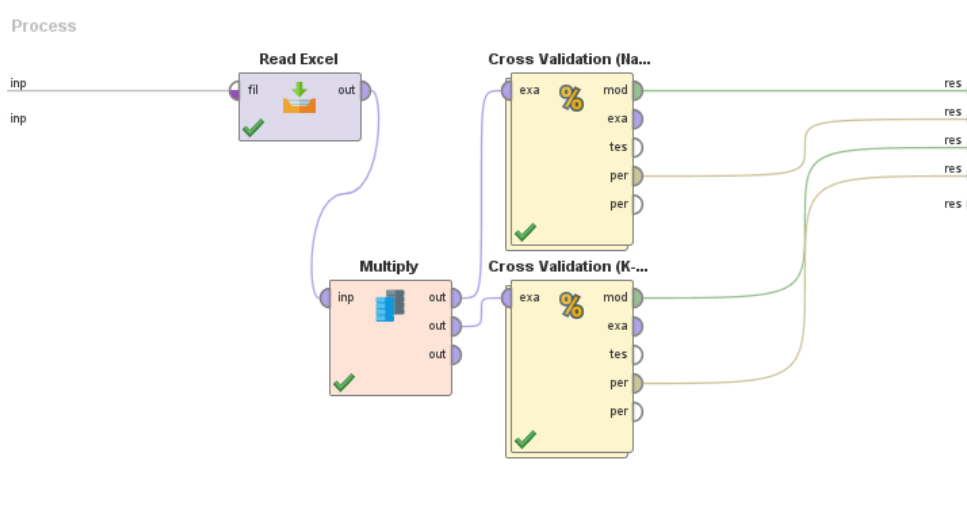
1. *Cleansing* dengan menggunakan *tools Rapidminer*. *Tools* ini nantinya akan berguna untuk membersihkan data yang telah di *crawling* dari *Twitter* agar tidak terdapat data yang duplikat dan tanda yang tidak diperlukan.
2. pemberian label pada sentimen berdasarkan *tweet* yang telah di *cleansing*. Pemberian label pada *tweet* peneliti dibantu oleh seorang pakar. Terdapat dua kelas atau label yang digunakan untuk melabelkan sentimen dari *tweet* pengguna twitter yaitu positif dan negatif.
3. Tahapan *Text Preprocessing* ditujukan guna mempersiapkan data yang bersih serta terbebas dari *noise* atau gangguan-gangguan seperti simbol, tanda baca, tautan, dan lain-lain.
4. Pembagian data training dan data testing dengan perbandingan 60:40, 70:30, dan 80:20.

### 3.3 Modelling

Tahapan selanjutnya adalah *Modelling*. Tahapan ini berhubungan langsung dengan teknik data mining. Pemilihan teknik data mining, algoritma dan parameter dengan nilai optimal [10]. Pemodelan dilakukan dengan tujuan untuk menguji keakuratan prediksi sistem berdasarkan data model yang dibuat. Dalam tahapan *Modelling* ini, penulis melakukan komparasi algoritma untuk mendapatkan model terbaik. Lalu menggunakan metode *split data* dengan perbandingan 80:20, 70:30 dan 60:40 dan metode lainnya yaitu *K-Fold Validation*.

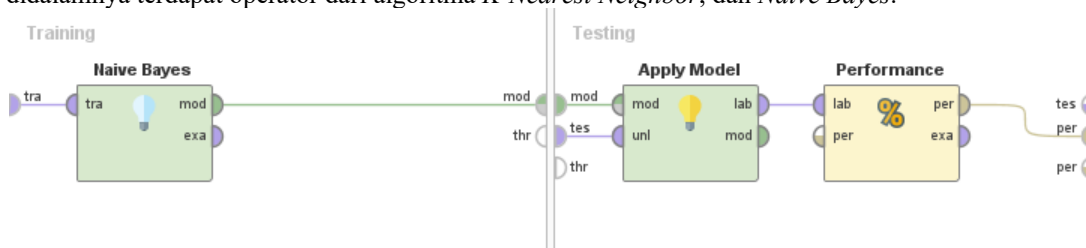
### 3.3.1 Komparasi Model

Pada penelitian ini dilakukan analisis komparasi algoritma klasifikasi *data mining* yaitu KNN (*K-Nearest Neighbor*) dan *Naïve Bayes*. Kemudian dari kedua metode tersebut dipilih yang memiliki tingkat akurasi terbaik dan hasil AUC terbaik.

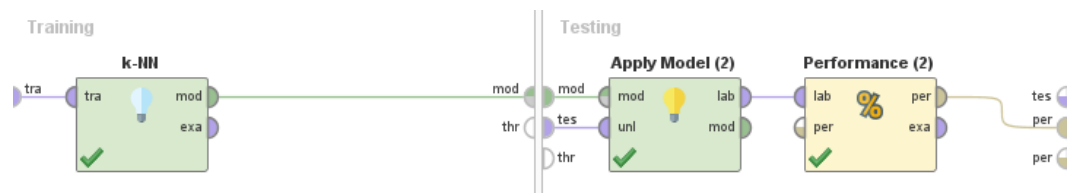


**Gambar 3** Proses Komparasi dengan Cross Validation KNN (*K-Nearest Neighbor*) dan *Naïve Bayes* Menggunakan Rapid Miner

Sebelum diproses pada tahapan *Cross Validation*, record data yang berupa *tweet* sebelumnya sudah diberi label oleh seorang pakar dan sudah melalui proses tahapan *cleasing* dan *text preprocessing*. Lalu dari data tersebut akan dibaca oleh operator *Read Excel* yang kemudian akan diproses dengan operator *Multiply*. Operator ini nantinya akan memproses 2 operator *Cross Validation* (*number of folds* = 10) dalam satu waktu yang bersamaan, yang didalamnya terdapat operator dari algoritma *K-Nearest Neighbor*, dan *Naïve Bayes*.



**Gambar 4** Operator di dalam *Cross Validation Naïve Bayes* Menggunakan Rapidminer



**Gambar 5** Operator didalam *Cross Validation K-NN* Menggunakan Rapidminer

Berdasarkan hasil dari analisis pengujian kedua algoritma tersebut maka dapat diamati hasilnya pada Tabel 1 berikut:

**Tabel 1.** Perbandingan *Performance* dari Kedua Algoritma

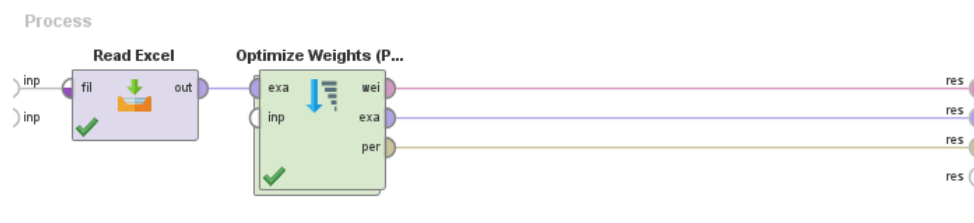
|                    | <i>Accuracy</i> | <i>AUC</i>   |
|--------------------|-----------------|--------------|
| K-Nearest Neighbor | 49,54%          | 0,524        |
| Naïve Bayes        | <b>54,79%</b>   | <b>0,599</b> |

Berdasarkan hasil tersebut dapat disimpulkan bahwa *Naïve Bayes* menghasilkan nilai uji AUC dan akurasi yang lebih baik dibandingkan dengan *K-Nearest Neighbor*. Yang dimana dapat disimpulkan juga bahwa algoritma *Naïve Bayes* lah yang dipilih untuk menjadi model terbaik untuk penelitian ini.

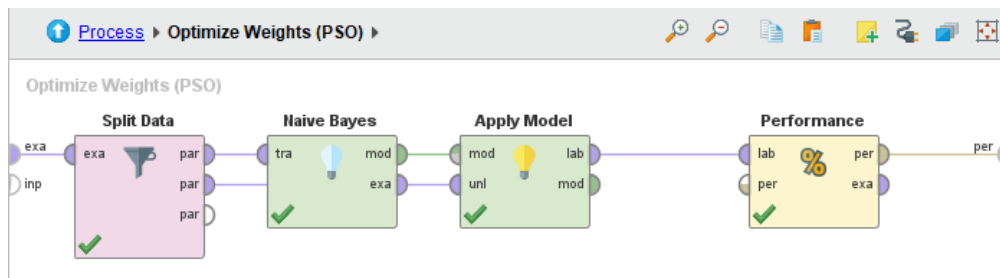
### 3.3.2 Modelling dengan Split Data

Pada tahapan *modelling* menggunakan *split data*, penulis menggunakan perbandingan 80:20 dengan algoritma *Naïve Bayes* sebagai contoh tahapan *modelling split data* dengan penambahan operator PSO (*Particle Swarm Optimization*).

*Particle Swarm Optimization* adalah teknik optimasi dengan cara menghitung secara terus menerus calon solusi dengan menggunakan suatu acuan kualitas. Algoritma ini mengoptimasi permasalahan dengan cara menggerakkan partikel / calon solusi di dalam ruang permasalahan menggunakan fungsi tertentu untuk posisi dan kecepatan dari partikel. Pergerakan partikel dipengaruhi oleh solusi terbaik partikel tersebut, dan solusi terbaik secara umum yang didapatkan dari partikel lain. Sekumpulan partikel ini dinamakan *swarm*, dan pada akhirnya *swarm* ini akan bergerak menuju kepada solusi terbaik [11].



Gambar 6 Proses *Modelling Split Data* Algoritma *Naïve Bayes* dengan PSO Menggunakan Rapidminer



Gambar 7 Operator *modelling split data* yang ada didalam operator PSO

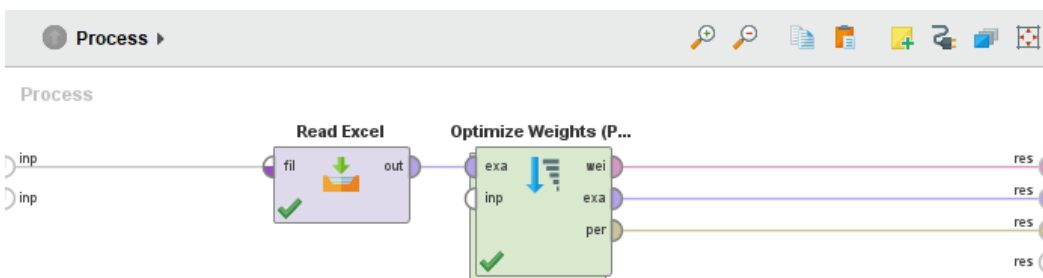
Dari hasil pengujian tersebut, hasil *modelling* menggunakan *split data* dengan ratio 80:20 menghasilkan tingkat akurasi sebesar 65.29%. Berikut ini adalah hasil *modelling* menggunakan *split data* dengan ratio 60:40, 70:30, dan 80:20. Maka nilai *accuracy* yang dihasilkan dapat dilihat pada Tabel 2 dengan nilai yang dicetak tebal merupakan nilai akurasi yang tertinggi.

Tabel 2 Hasil Modelling dengan Split Data

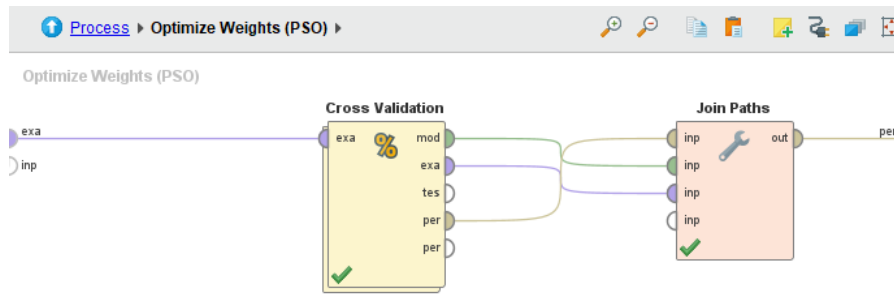
| <i>Split Data</i> | <i>Naïve Bayes</i> |
|-------------------|--------------------|
| 80 : 20           | <b>65.29%</b>      |
| 70 : 30           | 61.54%             |
| 60 : 40           | 60.33%             |

### 3.3.3 Modelling Dengan *K-Fold Cross Validation*

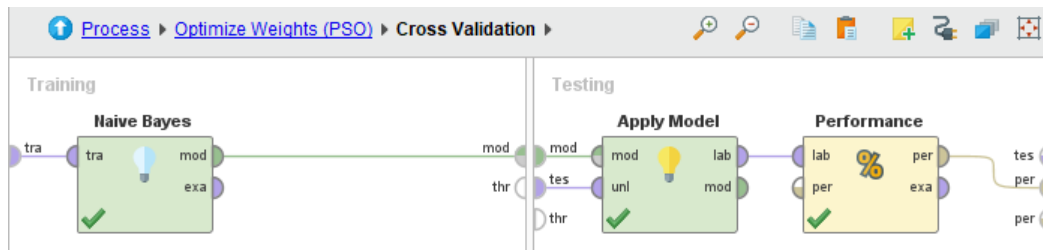
Dalam *modelling phase* kali akan akan menggunakan *k-fold cross validation*. Hal ini didasari untuk mengecek apakah nilai akurasi yang dihasilkan lebih tinggi atau lebih besar. Berikut adalah langkah pemodelan yang dilakukan dengan *k-fold cross validation*:



Gambar 8. *Modelling Cross Validation* Menggunakan Rapidminer



Gambar 9 Operator Cross Validation di dalam PSO



Gambar 10 Operator didalam Cross Validation Menggunakan Rapidminer

Setelah memproses tahapan *modelling* dengan penambahan operator *Particle Swarm Optimization* (PSO) dan dengan dataset yang telah melalui proses *preprocessing*, nilai *accuracy* yang di dapatkan dengan *k-fold cross validation* dengan nilai parameter *number of folds* = 10 didapatkan *Naïve Bayes* mendapatkan nilai akurasi tertinggi pada nilai sebesar 59.22%.

### 3.3.4 Perhitungan Algoritma *Naïve Bayes*

Untuk memproses perhitungan algoritma *Naïve Bayes* diperlukan sampel data. Sampel ini didapatkan dari dataset yang peneliti gunakan. Peneliti akan mengambil sebanyak 5 sampel dari dataset dengan secara acak.

Tabel 3 Sample Tweet

| Sampel | Tweet   | Label           |
|--------|---|-----------------|
| D1     | apakah ada promo tiket mtix pengen nonton avatar bawa rombongan ikih  | <i>positive</i> |
| D2     | Tadi di sebelah ada yang cetak tiket mtix gak bisa habis di cek ternyata beli nya buat di paragon semarang bukan paragon solo | <i>negative</i> |
| D3     | Sudah aman Kak Cuma mtix nya sedang dimatikan   | <i>negative</i> |
| D4     | Lebih suka mtix   | <i>positive</i> |
| D5     | kalau XXI cek di aplikasi Mtix aja  | <i>positive</i> |

Terhitung ada 3 record data positif dari total 5 record data maka nilai probabilitas dari record data positif sebesar  $P(\text{positif}) = \frac{3}{5} = 0.6$  dan probabilitas untuk record data negatif sebesar  $P(\text{negatif}) = \frac{2}{5} = 0.4$ . Untuk menghitung probabilitas setiap kata pada kelas ditentukan dengan rumus :

$$P(w_k|v_j) = \frac{n_{k+1}}{n + |\text{vocabulary}|} \quad (4)$$

$$P(\text{promosi}|\text{positif}) = \frac{1}{14} + \frac{1}{38} = 0,097744$$

$$P(\text{mtix}|\text{positif}) = \frac{3}{14} + \frac{1}{38} = 0,240602$$

$$P(\text{habis}|\text{negatif}) = \frac{1}{24} + \frac{1}{38} = 0,067982$$

$$P(\text{paragon}|\text{negatif}) = \frac{2}{24} + \frac{1}{38} = 0,109649$$

Berikut merupakan probabilitas setiap kata pada kelas.

**Tabel 4** Hasil Probabilitas Setiap Kata Pada Kelas

| Word     | Nk | N  | Vocabulary | P        |
|----------|----|----|------------|----------|
| promosi  | 1  | 14 | 38         | 0,097744 |
| tiket    | 1  | 14 | 38         | 0,097744 |
| mtix     | 3  | 14 | 38         | 0,240602 |
| nonton   | 1  | 14 | 38         | 0,097744 |
| avatar   | 1  | 14 | 38         | 0,097744 |
| bawa     | 1  | 14 | 38         | 0,097744 |
| kelompok | 1  | 14 | 38         | 0,097744 |
| lebih    | 1  | 14 | 38         | 0,097744 |
| suka     | 1  | 14 | 38         | 0,097744 |
| xxi      | 1  | 14 | 38         | 0,097744 |
| cek      | 1  | 14 | 38         | 0,097744 |
| aplikasi | 1  | 14 | 38         | 0,097744 |
| tadi     | 1  | 24 | 38         | 0,067982 |
| samping  | 1  | 24 | 38         | 0,067982 |
| cetak    | 1  | 24 | 38         | 0,067982 |
| bisa     | 1  | 24 | 38         | 0,067982 |
| habis    | 1  | 24 | 38         | 0,067982 |
| ternyata | 1  | 24 | 38         | 0,067982 |
| beli     | 1  | 24 | 38         | 0,067982 |
| untuk    | 1  | 24 | 38         | 0,067982 |
| paragon  | 2  | 24 | 38         | 0,109649 |
| semarang | 1  | 24 | 38         | 0,067982 |
| bukan    | 1  | 24 | 38         | 0,067982 |
| solo     | 1  | 24 | 38         | 0,067982 |
| sudah    | 1  | 24 | 38         | 0,067982 |
| aman     | 1  | 24 | 38         | 0,067982 |
| kakak    | 1  | 24 | 38         | 0,067982 |
| hanya    | 1  | 24 | 38         | 0,067982 |
| sedang   | 1  | 24 | 38         | 0,067982 |
| mati     | 1  | 24 | 38         | 0,067982 |

Sekarang proses dapat dilanjutkan ke proses klasifikasi untuk menentukan kelas dari dokumen baru. Sebagai contoh kita akan coba mengklasifikasikan data baru yang sentimennya belum di berikan.

**Tabel 5** Data Uji

| Tweet                    | Sentimen |
|--------------------------|----------|
| Lupa ada mtix penuh imax | ??       |

Sekarang dihitung nilai Naive Bayes untuk mendapatkan kelas positif dan negatif berdasarkan data tersebut:  
 Perhitungan Sentimen Positif :

$$p(\text{Positif}) \cdot p(\text{lupa}|\text{positif}) \cdot p(\text{ada}|\text{positif}) \cdot p(\text{mtix}|\text{positif}) \cdot p(\text{penuh}|\text{positif}) \cdot p(\text{imax}|\text{positif})$$

$$0.6 \times 0,097744 \times 0,097744 \times 0,097744 \times 0,097744 \times 0,097744 = 5,35305603235724 \times 10^{-6}$$

Perhitungan Sentimen Negatif :

$$p(\text{negatif}) \cdot p(\text{lupa}|\text{negatif}) \cdot p(\text{ada}|\text{negatif}) \cdot p(\text{mtix}|\text{negatif}) \cdot p(\text{penuh}|\text{negatif}) \cdot p(\text{imax}|\text{negatif})$$

$$0.4 \times 0,067982 \times 0,067982 \times 0,067982 \times 0,067982 \times 0,067982 = 5,80804105060018 \times 10^{-7}$$

Dari perhitungan diatas dapat disimpulkan bahwa nilai negatif lebih tinggi yaitu  $5,80804105060018 \times 10^{-7}$  sedangkan nilai dari positif sebesar  $5,35305603235724 \times 10^{-6}$ . maka sentimen dari tweet tersebut adalah negatif.



**Tabel 6** Hasil Dari Data Uji

| Tweet                    | Sentimen |
|--------------------------|----------|
| Lupa ada mtix penuh imax | Negatif  |

### 3.3.5 Confusion Matrix

Pada penelitian ini dilakukan *evaluation test* dengan *confusion matrix* yang dimana nilai yang digunakan didapat dari tahap pemodelan dengan algoritma *Naive Bayes* yang mendapat akurasi tertinggi menggunakan data split 80:20. Jadi *confusion matrix* yang ditampilkan pada Tabel 7 merupakan hasil dari *evaluation test* untuk *Naive Bayes*.

**Tabel 7** Confusion Matrix

|               | True Positive | True negative |
|---------------|---------------|---------------|
| Pred Positive | 39            | 32            |
| Pred Negative | 10            | 40            |

$$Recall = \frac{TP}{TP + FN} = \frac{39}{39 + 10} = 0.7959 = 79.59\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{39}{39 + 32} = 0.5493 = 54.93\%$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{39 + 40}{39 + 40 + 10 + 32} = 0.6529 = 65.29\%$$

Dilihat dari nilai yang ditampilkan oleh Tabel 7 didapatkan hasil dari pemrosesan data sentimen dengan *Naive Bayes* menghasilkan *recall* sebesar 79.59%, *precision* sebesar 54.93%, dan *accuracy* sebesar 65.29%. Adapun uraian hasil proses pembentukan confusion matrix sebagai berikut :

1. Hasil aktual positif yang diprediksi oleh *Naive Bayes* (TP) sebanyak 39 data.
2. Hasil aktual positif yang diprediksi negatif oleh *Naive Bayes* (TN) sebanyak 40 data.
3. Hasil aktual negatif yang diprediksi positif oleh *Naive Bayes* (FP) sebanyak 32 data.
4. Hasil aktual negatif yang diprediksi negatif oleh *Naive Bayes* (FN) sebanyak 10 data.

## 4. KESIMPULAN

Hasil dari penelitian ini menghasilkan nilai sentiment positif sebanyak 59% dan negatif sejumlah 41%. Hasil dari analisis komparasi algoritma klasifikasi *data mining* yaitu KNN (*K-Nearest Neighbor*) dan *Naive Bayes* menghasilkan *Naive Bayes* menghasilkan nilai uji AUC dan akurasi yang lebih baik dibandingkan dengan *K-Nearest Neighbor* yang menjadikan *Naive Bayes* sebagai model terbaik untuk penelitian ini. Berdasarkan hasil analisa dan pengujian bila dilihat dari tabel hasil *preprocessing* dataset menunjukkan jumlah komentar positif lebih banyak diberikan yang dimana dapat disimpulkan bahwa pengguna merasa puas dengan fitur serta layanan yang diberikan oleh aplikasi Mtx 21Cineplex. Lalu *Accuracy* tertinggi yaitu 65.29%, yang dihasilkan dengan pemodelan menggunakan algoritma *Naive Bayes* yang menggunakan perbandingan 80:20 untuk data *training* dan data *testing*.

## DAFTAR PUSTAKA

- [1] D. Yordan Sihombing and Y. Nataliani, "SISTEMASI: Jurnal Sistem Informasi Analisis Interaksi Pengguna Twitter pada Strategi Pengadaan Barang Menggunakan Social Network Analysis," 2021. [Online]. Available: <http://sistemasi.ftik.unisi.ac.id>
- [2] A. J. Syahid and D. Mahdiana, "Terakreditasi 'Peringkat 4 (Sinta 4)' oleh Kemenristekdikti Perbandingan Algoritma Untuk Klasifikasi Analisis Sentimen Terhadap Genose Pada Media Sosial Twitter," *semanTIK*, vol. 7, no. 1, pp. 1–5, 2021, doi: 10.5281/zenodo.5034916.
- [3] Y. Fairnando Augusto, A. Rachmadi, and A. D. Herlambang, "Analisis Kesuksesan Aplikasi Mobile Pemesanan Tiket Bioskop M-Tix Cinema 21 Berdasarkan Perspektif Pengguna di Kota Malang Menggunakan Pendekatan Delone and McLean Success Model," 2019. [Online]. Available: <http://j-ptiik.ub.ac.id>
- [4] A. Nabillah, S. Alam, and M. G. Resmi, "Twitter User Sentiment Analysis Of TIX ID Applications Using Support Vector Machine Algorithm," 2022.
- [5] P. Maulana, "Text Network Analysis Terhadap Layanan Smartfren Pada Media Sosial Twitter," 2022.

- [6] Y. B. Widodo, S. A. Anggraeini, and T. Sutabri, “Perancangan Sistem Pakar Diagnosis Penyakit Diabetes Berbasis Web Menggunakan Algoritma Naive Bayes,” *Jurnal Teknologi Informatika dan Komputer*, vol. 7, no. 1, pp. 112–123, Mar. 2021, doi: 10.37012/jtik.v7i1.507.
- [7] K. Sihotang and R. Ghaniy, “Penerapan Metode Naïve Bayes Classifier Untuk Penentuan Topik Tugas Akhir Pada Website Perpustakaan STIKOM Binaniaga,” *Jurnal Ilmiah Teknologi-Informasi dan Sains (TeknoIS)*, vol. 9, pp. 63–72, 2019.
- [8] A.- Arini, L. K. Wardhani, and D.- Octaviano, “Perbandingan Seleksi Fitur Term Frequency & Tri-Gram Character Menggunakan Algoritma Naïve Bayes Classifier (Nbc) Pada Tweet Hashtag #2019gantipresiden,” *KILAT*, vol. 9, no. 1, pp. 103–114, Apr. 2020, doi: 10.33322/kilat.v9i1.878.
- [9] M. F. Rifai, H. Jatnika, and B. Valentino, “Penerapan Algoritma Naïve Bayes Pada Sistem Prediksi Tingkat Kelulusan Peserta Sertifikasi Microsoft Office Specialist (MOS),” *PETIR*, vol. 12, no. 2, pp. 131–144, Sep. 2019, doi: 10.33322/petir.v12i2.471.
- [10] W. Astuti, E. Firasari, F. Lia Dwi Cahyanti, F. Sarasati, B. Digital, and S. Informasi, “Maret 2022 | Techno Nusa Mandiri,” *Journal of Computing and Information Technology As an Accredited Journal Rank*, vol. 19, no. 1, pp. 15–21, 2022, doi: 10.33480/techno.v19i1.2980.
- [11] H. A. D. Rani, “Optimasi Particle Swarm Optimization Pada Naïve Bayes Untuk Prediksi Kondisi Kelahiran Bayi,” *Jurnal Dialektika Informatika (Detika)*, vol. 2, no. 1, pp. 28–33, Dec. 2021, doi: 10.24176/detika.v2i1.6964.