Volume V No. 2Agustus 2019
P-ISSN 2442-2436, E-ISSN: 2550-0120
Akreditasi Ristekdikti, No: 21/E/KPT/2018
http://ejournal.bsi.ac.id/ejurnal/index.php/jtk

JURNAL TEKNIK KOMPUTER AMIK BSI

DOI: 10.31294/jtk.v4i2

# Komparasi Algoritma Naive Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Pengguna Busway

**[1]Riska Aryanti1, [2]Atang Saepudin2, [3]Eka Fitriani3,[4]Rifky Permana4, [5]Dede Firmansyah Saefudin**

[12]Ilmu Komputer, Teknologi Informasi, Universitas Bina Sarana Informatika,
Jl. Kamal Raya No.18 Ringroad Barat, Cengkareng, Jakarta Barat
Email: riska.rts@bsi.ac.id, atang.saepudin93@gmail.com

[34]Sistem Informasi, Teknologi Informasi, Universitas Bina Sarana Informatika,
Jl. Kamal Raya No.18 Ringroad Barat, Cengkareng, Jakarta Barat
Email: eka.ean@bsi.ac.id, rifky.rpp@bsi.ac.id

[5]Sistem Informasi, Teknologi Informasi, Universitas Bina Sarana Informatika PSDKU Purwokerto,
Jl. HR. Bunyamin 106, Sumampir Wetan, Pabuaran, Purwokerto Utara, Kabupaten Banyumas, Jawa Tengah
Email: dede.dfs@bsi.ac.id

***Abstract*** *- Congestion major cities in Indonesi caused by the proliferation of the use of private vehicles.Some expressing he thinks about busway user through the social media and other web site, This opinion can be used as a sentiment analysis to see if the user busway proposes a review of positive or negative.The results of the analysis sentiment can help in the sight of and evaluate the use of busway, also expected to improve and transjakarta facility from so they tend to have an opinion positive. Based on the results of the analysis, sentiment it is hoped people will switch to using the will of course will reduce congestion.In the study also added the stages preprocesing by using the framework gataframework to complete the process that cannot be done on tools rapidminer. The methodology that was used in this research was it is anticipated that analysis the sentiment of the by the application of an genetic algorithm for an election features with an algorithm naive bayes. From the results of the testing to the case in research it is found that classification algorithm naive bayesbasedgenetic algorithm having the kind of accuracy that good enough 88,55 % and value of auc reached 0,813 % with the level of the diagnosis classifications good.So that in this research classification algorithm naive bayesbasedgenetic algorithm can be recommended as algorithms classifications good enough to analyze the busway user sentimen. Based on analysis is expected to private transport users will switch to using the busway will reduce congestion.*

*Keywords: Busway, Sentiment Analysis, Genetic Algorithm.*

## PENDAHULUAN

Transjakarta adalah sebuah sistem transportasi Bus Rapid Transit (BRT) pertama di Asia Tenggara dan Selatan yang beroperasi sejak tahun 2004 di Jakarta, Indonesia. TransJakarta dirancang sebagai moda transportasi massal pendukung aktivitas ibukota yang sangat padat. Dengan jalur lintasan terpanjang di dunia (251.2 km), serta memiliki 260 halte yang tersebar dalam 13 koridor, Transjakarta yang awalnya beroperasi mulai Pkl. 05.00 – Pkl. 22.00 WIB, kini beroperasi 24 jam. Banyak negara telah mengadopsi serangkaian tindakan dan kebijakan untuk mengembangkan angkutan umum dan bahkan telah memperkenalkan strategi prioritas untuk mendorong pembangunan transportasi publik. Namun, meskipun banyak peluang untuk pembangunan menurut Murray dalam (Li, Bai, Song, Chen, & Wu, 2018) transportasi publik juga menghadapi banyak tantangan. Belum tersedianya Busway yang layak dan terintegrasi seluruh penjuru serta terjangkau secara ekonomi. Sehingga masyarakat tidak lagi menggunakan kendaraan pribadi sebagai alat transportasi sehari-hari, yang tidak kalah penting adalah pemerintah belum melakukan upaya untuk meningkatkan kesadaran masyarakat terhadap pentingnya budaya tertib dan taat terhadap rambu dan aturan lalu lintas. Solusi untuk mengurangi kemacetan tersebut adalah dengan peningkatan penggunaan Busway, yang nyatanya masih belum banyak diminati masyarakat.

Menurut A. M. Kaplan dalam (Kristiayanti, Umam, Wahyudi, Amin, & Marlinda, 2018) Akhir-akhir ini media sosial menjadi tren yang luar biasa. Peran media sosial sangat berpengaruh bagi perkembangan situasi global saat ini. Menurut Forrester Research, 75% dari peselancar internet telah menggunakan media sosial pada kuartal kedua tahun 2008 dengan bergabung ke situs jejaring sosial, membaca blog, atau cukup memberi ulasan untuk situs belanja online, terjadi kenaikan yang

signifikan dari 56% pada tahun 2007, Pertumbuhan ini tidak terbatas pada kelompok remaja, sebab mereka yang kini berada pada kelompok Generasi X (kini pada rentang umur 35-44 tahun) ternyata juga ikut di dalamnya, baik sebatas gabung saja, sekadar menyimak, atau kritikus di dalamnya.

Sebagian masyarakat menyampaikan pendapat dan opininya mengenai penggunaan Busway dalam kota, pendapat masyarakat tersebut dituangkan dalam opini di media sosial salah satunya adalah *twitter*. *Twitter* menjadi salah satu media sosial yang digunakan untuk mengutarakan*review* tentang berbagai isu atau topik yang sedang tren melalui kolom tweet. Namun membaca *review* tersebut secara keseluruhan dapat memakan waktu. Sementara jika hanya sedikit *review* yang dibaca, maka evaluasi akan bias. Analisa sentimen bertujuan untuk mengatasi masalah ini dengan secara otomatis mengelompokkan *review* pengguna menjadi opini positif atau negatif (Muthia, 2016).

Analisis sentimen dengan menggunakan review bahasa Indonesia masih memiliki kesulitan dalam proses preprocessing, yaitu proses dimana data yang diambil langsung dari twitter masih harus dilakukan penyeragaman text untuk proses selanjutnya. Dalam *tools* rapidminer belum tersedianya kamus bahasa Indonesia untuk *preprocessing* text dengan bahasa Indonesia. *Gataframework* adalah sebuah *frameworktextmining* dalam bahasa Indonesia, pada *gataframework* memiliki *feature* untuk memproses text agar sebuah kalimat dapat di *preprocessing* dengan baik.

Analisis sentimen adalah semacam klasifikasi teks yang mengklasifikasikan teks berdasarkan orientasi sentimen pendapat yang dikandungnya. Hal ini juga dikenal sebagai opinion mining, ekstraksi pendapat dan mempengaruhi analisis dalam literatur (Govindarajan, 2013). Terdapat beberapa algoritma klasifikasi yang dapat digunakan untuk analisis sentimen klasifikasi teks diantaranya *Naive Bayes* (NB), *Suppport Vector Machine*(SVM), dan *K-Nearest neighbor* (KNN).

Penelitian sebelumnya telah banyak dilakukan dengan menerapkan *Genetic Algorithm* (GA) untuk seleksi fitur dalam menganalisa sentimen menggunakan algoritma klasifikasi seperti Naive Bayes, dan *Suppport Vector Machine* yang paling umum digunakan. Diantaranya seperti penelitian yang dilakukan oleh (Govindarajan, 2013) pada penelitian ini menganalisa sentimen berdasarkan *review* film dengan menggunakan teknik *ensemble* dengan klasifikasi sentimen menggunakan Naive Bayes dan Algoritma Genetika. Penelitian lainnya yaitu dilakukan oleh Dinda Ayu Muthia tentang *review* restoran menggunakan algoritma Naive Bayes(Muthia, Dinda Ayu, 2017).

Terdapat beberapa teknik seleksi fitur yang dapat digunakan untuk memecahkan masalah optimasi, diantaranya GA. GA memiliki potensi untuk menghasilkan fitur yang lebih baik dan menjadi parameter optimal pada waktu yang sama (Wahyudi & Putri, 2016). Sehingga pada penelitian ini penulis memilih teknik fitur seleksi dengan menggunakan Algoritma Genetika yang akan penulis bandingkan dengan algoritma naive bayes untuk diterapkan dalam mengklasifikasikan teks pada *review* pengguna busway, yang mana pada hasilnya dapat ditentukan dari penerapan Algoritma Genetika untuk seleksi fitur dengan komparasi algoritma klasifikasi manakah yang terbaik untuk diterapkan dalam rangka meningkatkan akurasi analisa sentimen.

Tinjauan pustaka dilakukan dengan menggunakan referensi dari buku-buku, jurnal ataupun artikel yang didapatkan melalui media internet sebagai acuan penulisan ini, berikut adalah pengertian-pengertian mengenai penulisan yang akan dibahas.

**A. Review Text**

*Review text* adalah teks yang ditujukan untuk meninjau suatu karya, baik film, buku dan sebagainya, untuk mengetahui kualitas, kelebihan serta kekurangan yang dimiliki oleh karya tersebut. Tujuan koomunikatif dari *Review text* adalah *to criticise an art work, event for a public audience* (melakukan kritik terhadap peristiwa atau karya seni ataupun lainnya untuk khalayak umum).

Busway, atau disebut juga angkutan kota, adalah angkutan dari suatu tempat ke tempat lain dalam wilayah kota dengan menggunakan mobil bus dan atau mobil penumpang umum yang terkait dalam trayek tetap dan teratur. Ketersediaannya Busway yang baik tentunnya akan memberikan manfaat bagi masyarakat secara umum. Masyrakat akan terbantu jika ingin berpergian ke suatu tempat tujuan. Apabila Busway dijaga dan dirawat dengan baik, tentu akan memberikan dampak baik, yaitu masyarakat akan lebih memilih menggunakan Busway dibandingkan kendaraat pribadi. (Novantirani, Sabariah, dan Effendy, 2015).

**B. Analisa Sentimen (*Sentiment Analysis*)**

Analisis sentimen adalah "Riset komputasional dari opini, sentimen dan emosi yang diekspresikan secara tekstual dilakukan untuk melihat pendapat terhadap sebuah masalah, atau untuk identifikasi kecenderungan hal di pasar. Saat ini pendapat masyarakat menjadi sumber yang penting dalam pengambilan keputusan akan suatu produk." (Ipmawati, Kusrini, dan Luthfi, 2017).

**C. Seleksi Fitur(*Feature Selection*)**

*Feature Selection* atau seleksi fitur adalah sebuah proses yang biasa digunakan pada *Machine Learning* dimana sekumpulan dari fitur yang dimiliki oleh data digunakan untuk pembelajaran algoritma. *Feature Selection* menurut (Nugroho dan Wibowo, 2017).

Seleksi fitur adalah kemampuannya untuk mendeteksi hubungan nonlinier antar variabel, hal ini memungkinkan pengambilan relevansi dan redundansi fitur secara bersamaan (Mira et al., 2018).

Masalah dalam seleksi fitur adalah pengurangan dimensi, dimana awalanya semua atribut diperlukan untuk memperoleh akurasi yang maksimal.

Adapun tujuan seleksi fitur menurut (Wahyuni, 2016) adalah "Mengurangi fitur data yang berdimensi tinggi", untuk tujuan dari kegiatan data mining dan fitur metode seleksi.

**D. Algoritma *Naive Bayes***

*Naive bayesian* adalah metode klasifikasi yang berdasarkan probabilitas, dengan asumsi bahwa setiap variabel X bersifat bebas (independent). Dengan kata lain, *Naïve Bayesian* mengansumsikan bahwa keberadaan sebuah atribut tidak ada kaitannya dengan beradaan atribut yang lain. Jika diketahui X adalah data sampel dengan kelas (label) yang tidak diketahui, H merupakan hipotesa bahwa X adalah data dengan klas (label) C, P(H) adalah peluang dari hipotesa H, P(X) adalah peluang data sampel yang diamati, maka P(X|H) adalah peluang data sampel X, bila diasumsikan bahwa hipotesa H benar (valid) (Wahyuni, 2016).

**E. Algoritma Genetika *(Genetic Algorithm)***

Algoritma genetika digunakan untuk mengoptimasi parameter yang optimal dengan ruang lingkup yang besar, dengan pemilihan parameter yang tepat algoritma genetika akan lebih optimal (Wang et al., 2013). Algoritma genetika memiliki kelemahan yaitu pemilihan parameter yang salah dapat mengurangi akurasi yang dihasilkan.

**F. Validasi dan Evaluasi Algoritma Text Mining**

Validasi merupakan proses mengevaluasi akurasi dari suatu model. Dalam mengevaluasi model klasifikasi berdasaran perhitungan objek data testing mana yang diprediksi benar dan tidak benar. Perhitungan tersebut akan ditabulasikan kedalam tabel yang disebut confusiom matrix (Gorunescu, 2011).

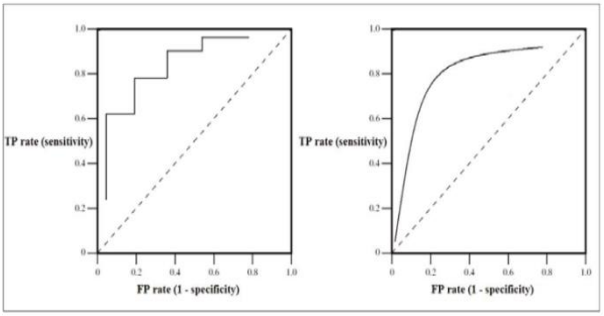
**Tabel.1. *Confusion Matrix***

| Classification | Predicted Class | | |
|---|---|---|---|
| | | Class = Yes | Class = No |
| Observed Class | Class = Yes | A (true positive – TP) | b (false negative – FN) |
| | Class = No | C (false positive – FP) | d (true negative – TN) |

Sumber : (Gorunescu, 2011)

Kurva ROC (*Receiver Operating Characteristic*) digunakan untuk mengevaluasi akurasi classifier dan untuk membandingkan klasifikasi yang berbeda model (Vercellis, 2009). Kurva ROC digunakan untuk mengukur AUC (*Area Under Curve*). Kurva ROC membagi hasil positif dalam sumbu y dan hasil negatif dalam sumbu x (Witten, Frank, dan Hall, 2011b). Sehingga semakin besar area yang berada dibawah kurva. semakin baik pula hasil prediksi.

Permasalahan dalam klasifikasi kurva ROC dapat digunakan untuk menguji dan menilai hasil kinerja pengklasifikasian secara visual dan yang digunakan untuk mengekspresikan *confusionmatrix*. Kurva ROC merupakan grafik dua dimensi dengan *falsepositive* sebagai garis *horizontal* dan *truepositive* sebagai garis *vertikal* (Vecellis, 2009).




Sumber : Gorunescu (2011)

Gambar 1. Kurva ROC

Berikut panduan untuk mengklasifikasikan keakuratan diagnosa menggunakan AUC (Gorunescu, 2011) :

1. 0.90-1.00 = *excellent classification*;
2. 0.80-0.90 = *good classification*;
3. 0.70-0.80 = *fair classification*;
4. 0.60-0.70 = *poor classification*;
5. 0.50-0.60 = *failure*.

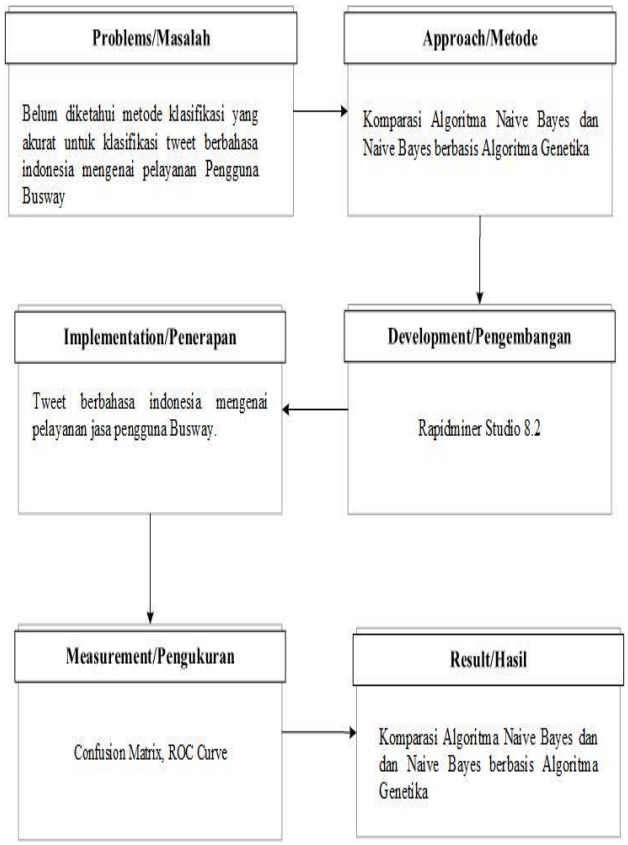
**G. *Gataframework***

*Gataframework* adalah *framework* yang digunakan untuk *preprocessing* text miningbahasa indonesia yang menyediakan fitur *indonesian stopword removal, indonesian stemming, tokenization: regexp, @anotation removal, Normalization:indonesianslank, Normalization:indonesian acronim, Normalization:emoticon, transformation:not(negative),* dan *words count.* Pada *framework* ini untuk melakukan tahapan *preprocessing* dapat dilakukan dengan *singletext,* dan *uploadfile*. Pada *upload file* umtuk *file excel* yang diupload dengan *formatexcel* 2003 (.XLS).Untuk menggunakan *Gataframework* dapat menggunakan link berikut: http://www.gataframework.com/textmining/

**H. Kerangka Pemikiran**

Dalam menyelesaikan penelitian ini, penulis membuat sebuah kerangka pemikiran yang digunakan sebagai acuan dalam penelitian ini

sehingga penelitian dapat dilakukan dengan baik. Permasalahan dalam penelitian ini adalah belum diketahui metode yang tepat dengan akurasi terbaik untuk klasifikasi teks pada review yang memiliki sentiment positive maupun negative terhadap penggunaan jasa Busway dengan menggunakan algoritma klasifikasi Naive Bayes (NB), dan penerapan Genetic Algorithm (GA) untuk seleksi fitur dengan menambahkan model pembobotan fitur yang akan digunakan adalah Term Frequency Invers Document Frequency (TF-IDF) dan pemilihan seleksi fitur menggunakan Genetic Algorithm (GA). Pengkalsifikasian yang digunakan adalah pertama menggunakan *Naive Bayes* dengan pengujian 10 *Fold Cross Validation*.




Sumber : Hasil Penelitian(2019)

Gambar 2. Kerangka Pemikiran

## METODOLOGI PENELITIAN

Metode Penelitian ekperimen adalah metode penelitian yang dilakukan penulis guna penyelesaian tesis ini, dengan tahapan sebagai berikut:

### A. Pengumpulan Data

Tahap pertama pada penelitian ini pengumpulan data *review* pengguna Busway dari sosial media. Dataset yang digunakan berfokus pada opini berbahasa indonesia yang membahas tentang penggunaan Busway.

Tahapan selanjutnya yaitu proses pelabelan data dilakukan dengan memberikan status *tweet* berdasarkan sentimen *positive* dan sentimen *negative*. Pada proses pelabelan ini dilakukan secara manual dengan memberikan nilai informasi terhadap masig-masing *tweet.* Berikut adalah contoh data yang telah diberikan status *positive* dan *negative.*

Tabel 1. Contoh data *tweet* yang sudah diberi Label

| Text | Status |
|---|---|
| @PT_TransJakarta Sampai saat ini belum ada busway yg lewat dr arah harmoni yg ke ancol @PT_TransJakarta penumpang nya udah… https://t.co/E1hnLapYps | Negative |
| Jam segini dapet duduk di busway tu bahagia ☺□ | Positive |
| @PT_TransJakarta mau tanya... transjakarta koridor 5a Kampung Melayu -Grogol masih ada ga ya? Saya udah nunggu setengah jam… https://t.co/L0QoLC1aaX | Negative |

Sumber : Hasil Penelitian(2019)

### B. Pengolahan Data Awal

Pengolahan data awal yaitu tahapan *preprocessing* yang dilakukan dengan proses *Tokenization, Transform Cases, Stopword Removal, Normalize, Stemming* dan *Generate N-Grams*. Pada tahapan ini penulis menggunakan tambahan *Gataframework* dikarenakan pada *tools rapidminer* masih tedapat kelemahan dalam text bahasa indonesia maka penulis menggunakan gataframework http://www.gataframework.com/textmining/ untuk pengolahan data awal seperti:*stopword removal, normalize indonesian slank,* dan *stemming.*

### C. Klasifikasi

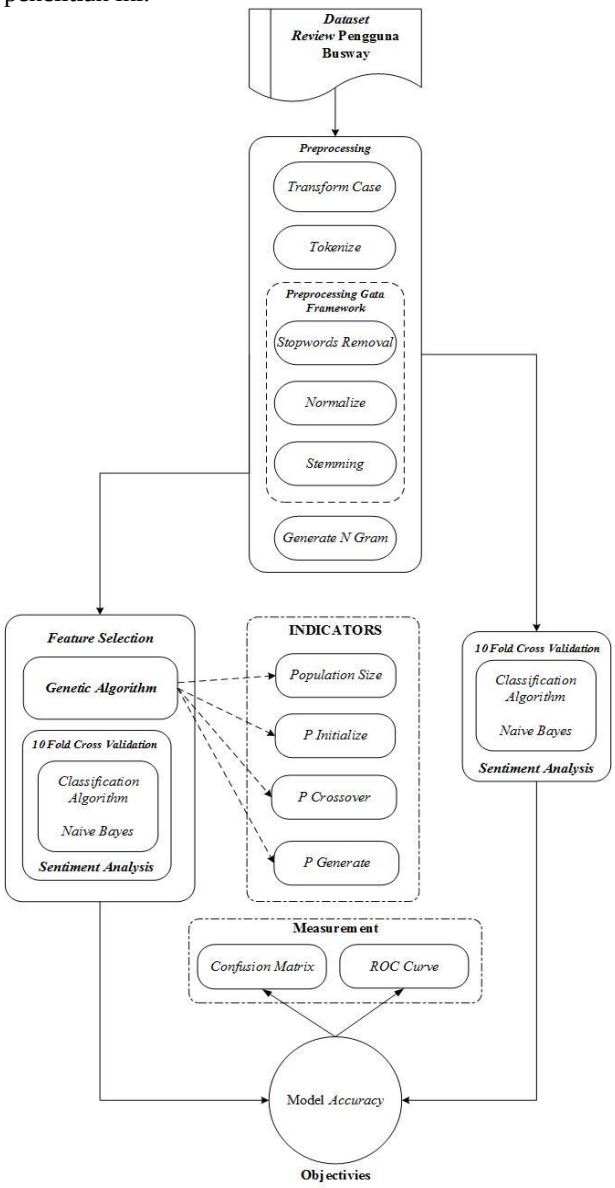
Tahap berikutnya proses klasifikasi dengan membandingkan model klasifikasi seperti Naive Bayes, untuk menentukan kesesuaian data dengan metode manakah yang terbaik dari beberapa metode pengklasifikasian teks yang digunakan oleh beberapa peneliti sebelumnya.

### D. Pembobotan dan Pemilihan Fitur Seleksi

Tahap selanjutnya menambahkan model pembobotan karena pengklasifikasian tersebut memiliki kekurangan terhadap masalah pemilihan parameter yang sesuai, karena dengan tidak sesuainya sebuah pengaturan parameter dapat menyebabkan hasil klasifikasi menjadi rendah. Metode pembobotan Fitur yang akan digunakan adalah *Term Frequency Invers Document Frequency* (TF-IDF) dan pemilihan seleksi fitur menggunakan Algoritma Genetika (GA). Pengklasifikasian yang digunakan adalah pertama menggunakan *Naive Bayes* kemudian klasifikasi yang kedua menggunakan *Naive Bayes* dengan pengujian 10 *Fold Cross Validation*.

**E. Validasi(*Validation)***

Hasil akurasi algoritma akan digambarkan ke dalam *Confusion Matrix* dan kurva ROC. RapidMiner digunakan sebagai alat bantu dalam mengukur akurasi data eksperimen yang dilakukan dalam penelitian. Gambar 3.menggambarkan metodologi penelitian yang penulis usulkan dalam penelitian ini.




Sumber : Hasil Penelitian(2019)

Gambar 3. Metodologi Penelitian

## HASIL DAN PEMBAHASAN

Berdasarkan metodologi penelitian yang telah dipaparkan, berikut implementasi metodologi yang dilakukan dalam penelitian ini.

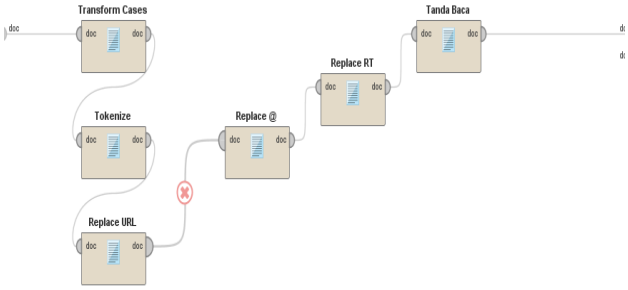
**A. Pengumpulan Data**

Pada tahapan awal ini, merupakan tahapan dimana data mentah dikumpulkan menggunakan operator search twitter yang tersedia pada tools rapidminer dengan bantuan menggunakan access twitter API sebagai konektor ke API Twitter tentang pengguna Busway berdasarkan keyword "Busway", dengan data diambil dari tanggal 15 Juli - 29 Juli data yang dikumpulkan dalam proses *crawling* data tweet ini kemudian disimpan dalam format file excel.

Setelah data twitter selesai untuk dikumpulkan tahap berikutnya adalah melakukan pelabelan tweet sesuai dengan kelas yang ditentukan sebelumnya. Dataset yang telah berhasil dikumpulkan dari twitter merupakan data yang tidak mempunyai label (unsupervised data), sehingga agar dapat diproses dengan menggunakan teknik supervised learning, maka data twitter yang telah dikumpulkan sebelumnya perlu untuk diberi label. Pada penelitian ini proses pelabelan dilakukan secara manual dengan memberikan status tweet positif dan negatif.

**B. Pengolahan Data Awal**

Pada tahapan ini diawali dengan tahapan preprocessing seperti transform cases, tokenize dengan menggunakan rapidminer. Dapat dilihat pada gambar 4.




Sumber : Hasil Penelitian(2019)

Gambar 4. Tahap *Preprocessing Tokenize* dan *Transform Cases*

1. *Transform Cases*

Pada tahapan transform cases tujuan tahapan ini mengubah semua huruf pada tweet menjadi huruf kecil semua atau menjadi huruf kapital semua. Pada penelitian ini mengubah semua huruf pada teks menjadi huruf kecil karena mayoritas teks berupa tulisan opini yang sebagian besar merupakan huruf kecil semua.

Tabel 2. Perbandingan Teks Sebelum dan Sesudah dilakukan Proses *Transform Cases*

| Teks sebelum dilakukan proses *Transform Cases* | Teks setelah dilakukan proses *Transform Cases* |
|---|---|
| RT @PT_TransJakarta Sampai saat ini belum ada busway yg lewat dr arah harmoni yg ke ancol @PT_TransJakarta penumpang nya udah… https://t.co/E1hnLapYps | rt @pt_transjakarta sampai saat ini belum ada busway yg lewat dr arah harmoni yg ke ancol @pt_transjakarta penumpang nya udah… https://t.co/e1hnlapyps |
| @PT_TransJakarta mau tanya... transjakarta koridor 5a Kampung Melayu -Grogol masih ada ga ya? Saya udah nunggu setengah jam… https://t.co/L0QoLC1aaX | @pt_transjakarta mau tanya... transjakarta koridor 5a kampung melayu -grogol masih ada ga ya? saya udah nunggu setengah jam… https://t.co/l0qolc1aax |

Sumber : Hasil Penelitian(2019)

2. *Tokenization*

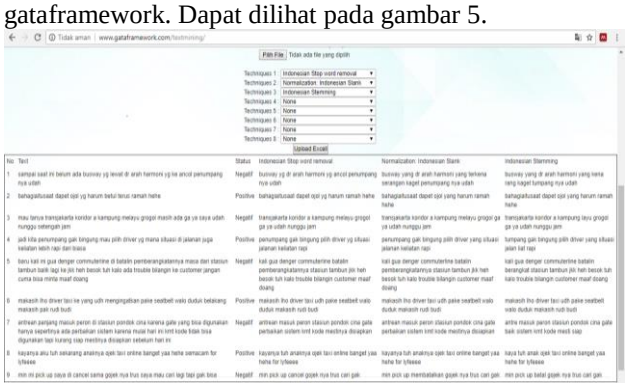
Proses *tokenization* berfungsi untuk menghilangkan tanda baca, simbol dan karakter yang bukan huruf pada setiap dokumen *review*. Semua karakter yang tidak diperlukan akan dibuang. Termasuk *white space* yang berlebihan dan semua tanda baca.

Tabel 3. Perbandingan Teks Sebelum dan Sesudah dilakukan Proses *Tokenization*

| Teks sebelum dilakukan proses pembersihan dari *Mentioned* | Teks setelah dilakukan proses pembersihan dari *mentioned* |
|---|---|
| rt @pt_transjakarta sampai saat ini belum ada busway yg lewat dr arah harmoni yg ke ancol @pt_transjakarta penumpang nya udah… https://t.co/e1hnlapyps | sampai saat ini belum ada busway yg lewat dr arah harmoni yg ke ancol penumpang nya udah |
| @pt_transjakarta mau tanya... transjakarta koridor 5a kampung melayu -grogol masih ada ga ya? saya udah nunggu setengah jam… https://t.co/l0qolc1aax | mau tanya transjakarta koridor lima a kampung melayu grogol masih ada ga ya saya udah nunggu setengah jam |

Sumber : Hasil Penelitian(2019)

Tahapan preprocessing selanjutnya yaitu proses Stopword Removal, Normalize Indonesian Slank dan Stemming dengan menggunakan gataframework. Dapat dilihat pada gambar 5.



Sumber : Hasil Penelitian(2019)

Gambar 5. Tahap *Preprocessing Gataframework*

3. *Stop Word Removal*

*Stop Word Removal* yaitu penghapusan kata-kata yang tidak relevan, seperti kata sambung dan lainnya misalnya tetapi, guna, untuk, agar, supaya, yang merupakan kata-kata yang tidak mempunyai makna tersendiri jika dipisahkan dengan kata yang lain dan tidak terkait dengan kata sifat yang berhubungan dengan sentiment. Pada tahapan ini akan menyempurnakan tahap sebelumnya.

Tabel 4. Perbandingan Teks Sebelum dan Sesudah dilakukan Proses *Stop Word Removal*

| Teks sebelum dilakukan proses *Stop Word Removal* | Teks setelah dilakukan proses *Stop Word Removal* |
|---|---|
| sampai saat ini belum ada busway yg lewat dr arah harmoni yg ke ancol penumpang nya udah | busway yg dr arah harmoni yg ancol penumpang nya udah |
| mau tanya transjakarta koridor lima a kampung melayu grogol masih ada ga ya saya udah nunggu setengah jam | transjakarta koridor lima a kampung melayu grogol ga ya udah nunggu jam |

Sumber : Hasil Penelitian(2019)

4. *Normalize Indonesian Slank*

*Normalize indonesian slank* menormalkan kalimat indonesia yang alay atau gaul. Seperti: yg menjadi yang, gw menjadi saya, otw menjadi dalam perjalanan dan lain sebagainya.

Tabel 5. Perbandingan Teks Sebelum dan Sesudah dilakukan Proses *Normalize Indonesian Slank*

| Teks sebelum dilakukan proses *Transform Cases* | Teks setelah dilakukan proses *Transform Cases* |
|---|---|
| busway yg dr arah harmoni yg ancol penumpang nya udah | busway yang dr arah harmoni yang terkena serangan kaget penumpang nya udah |
| transjakarta koridor lima a kampung melayu grogol ga ya udah nunggu jam | transjakarta koridor lima a kampung melayu grogol ga ya udah nunggu jam |

Sumber : Hasil Penelitian(2019)

5. *Stemming*

Proses pengelompokkan kata ke dalam beberapa kelompok yang memiliki kata dasar yang sama dan melakukan transformation untuk proses pembobotan dengan melakukan penghitungan terhadap kehadiran atau ketidakhadiran sebuah kata di dalam dokumen. Dengan tujuan semua kata yang telah telah dipilih menjadi token pada tahapan sebelumnya, akan diubah ke dalam bentuk asal (kata dasar).

Tabel 6. Perbandingan Teks Sebelum dan Sesudah dilakukan Proses *Stemming*

| Teks sebelum dilakukan proses *Stemming* | Teks setelah dilakukan proses *Stemming* |
|---|---|
| busway yang dr arah harmoni yang terkena serangan kaget penumpang nya udah | busway yang dr arah harmoni yang kena rang kaget tumpang nya udah |
| transjakarta koridor lima a kampung melayu grogol ga ya udah nunggu jam | transjakarta koridor lima a kampung layu grogol ga ya udah nunggu jam |

Sumber : Hasil Penelitian(2019)

6. *Generate N-Gram*

Generate N-Gram yaitu proses menghitung probabilitas bersyarat untuk sebuah kata dari urutan kata sebelumnya. Sebuah n-gram adalah sebuah kumpulan kata dengan masing-masing memiliki panjang n kata. N-gram ukuran 1 disebut sebagai unigram, ukuran 2 sebagai bigram, ukuran 3 sebagai trigram, dan seterusnya. Proses ini dimulai dengan memecah kata per kata dan mengelompokkan hasil pemecahan kata tersebut kedalam n-gram ukuran 1, n-gram ukuran 2, n-grams ukuran 3 dan sterusnya untuk dilakukan perhitungan dari kata yang sering muncul pada suatu kalimat.

Tabel 7. Perbandingan Teks Sebelum dan Sesudah dilakukan Proses *Generate N-Gram*

| Teks sebelum dilakukan proses T*okenization* | Teks setelah dilakukan proses T*okenization* |
|---|---|
| busway yang dr arah harmoni yang kena rang kaget tumpang nya udah | busway busway_yang yang yang_dr dr dr_arah arah arah_harmoni harmoni harmoni_yang yang yang_kena kena kena_rang rang rang_kaget kaget kaget_tumpang tumpang tumpang_nya nya nya_udah udah |
| transjakarta koridor lima a kampung layu grogol ga ya udah nunggu jam | transjakarta transjakarta_koridor koridor koridor_lima lima lima_a a a_kampung kampung kampung_layu layu layu_grogol grogol grogol_ga ga_ya ya ya_udah udah_nunggu nunggu nunggu_jam jam |

Sumber : Hasil Penelitian(2019)

**C. Klasifikasi**

Tahap pengklasifikasian teks menggunakan data sampel kendaraan busway dengan menggunakan model klasifikasi Naive Bayes, dipilih untuk menentukan kesesuaian data dengan metode manakah yang terbaik dari beberapa metode pengklasifikasian teks yang digunakan peneliti, dengan tujuan untuk mengetahui pola sentimen dari komentar-komentar yang telah di preprocessing.

**D. Pembobotan dan Pemilihan Fitur Seleksi**

Metode pembobotan fitur yang akan digunakan adalah Term Frequency Invers Document Frequency (TF-IDF) dan pemilihan seleksi fitur yang akan diujicoba menggunakan GA. Ujicoba pertama kali yaitu fitur seleksi GA dengan model NB dan pengujian 10 Fold Cross Validation. Ujicoba selanjutnya yaitu fitur seleksi GA dengan model SVM dan pengujian 10 Fold Cross Validation.

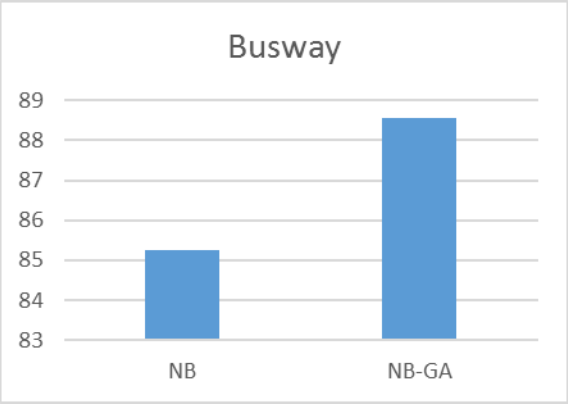
**E. Hasil**

Berikut adalah hasil perbandingan dari percobaan pengujian menggunakan model *Naive Bayes* dan *Naive Bayes* berbasis Algoritma Genetika.

Tabel 8. Hasil Akurasi dan Nilai AUC Komparasi Algoritma Naive Bayes dan Naive Bayes berbasis Algoritma Genetika

| **Akurasi** | | |
|---|---|---|
| **Dataset** | **NB** | **NB-GA** |
| Busway | 85,27 | 88,55 |
| **AUC** | | |
| **Dataset** | **NB** | **NB-GA** |
| Busway | 0,819 | 0,813 |

Sumber : Hasil Penelitian(2019)




Sumber : Hasil Penelitian(2019)

Gambar 6. Grafik Hasil Komparasi Algoritma Klasifikasi

Dapat disimpulkan bahwa bedasarkan hasil dari pengolahan data yang telah dilakukan, terbukti setelah penambahan fitur seleksi Algoritma Genetika akurasi untuk algoritma Naive Bayes berbasis Algoritma Genetika hasil akurasi terjadi peningkatan sebesar 3,28%, sehingga hasil menjadi 88,55% dan nilai AUC mencapai 0.813%. sehingga pada penelitian ini dapat direkomendasikan bahwa algoritma Naive Bayes berbasis Algoritma Genetika dapat direkomendasikan sebagai klasifikasi yang baik dalam analisis sentimen terhadap pengguna busway.

## KESIMPULAN

Dalam penelitian ini penulis melakukan pengklasifikasian teks untuk menganalisa sentiment dari opini masyarakat melalui media sosial twitter mengenai penggunaan Busway. Adapun dalam penelitian ini menggunakan algoritma klasifikasi, Naive Bayes dan Naive Bayes berbasis algoritma genetika.

Terdapat peningkatan dengan menambahkan fitur naive bayes berbasis algoritma genetika, dengan akurasi yang dihasilkan dari hasil pengujian algoritma Naive Bayes dengan metode pengujian confusion matrix dan kurva AUC maka dihasilkan dengan nilai akurasi mencapai 85,27% dan nilai AUC mencapai 0.819% sehingga termasuk *good classification*. Sedangkan setelah ditambahkan fitur Algoritma Genetika maka Naive Bayes berbasis Algoritma Genetika menghasilkan 88,55% dan nilai AUC mencapai 0.813% sehingga termasuk *good classification*.Sehingga dapat direkomendasikan bahwa algoritma klasifikasi Naive Bayes berbasis Algoritma Genetika sebagai algoritma klasifikasi yang baik dibandingkan dengan Naive Bayes.

Pada penelitian ini juga menemukan bahwa dengan menggunakan framework gataframework dapat membantu pada tahapan preprocessing text mining untuk Bahasa Indonesia. Feature yang ada pada gataframework mampu membuat teks yang ditarik dari social media twitter menjadi data yang dapat diolah dalam tools rapidminer.

## REFERENSI

Govindarajan, M. (2013). Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm. *International Journal of Advanced Computer Research*, *3*(13), 139–145.

Ipmawati, J., Kusrini, & Luthfi, E. T. (2017). Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen. *Ijns.Org Indonesian Journal on Networking and Security*, *6*(1), 28–36.

Kristiayanti, D. A., Umam, A. H., Wahyudi, M., Amin, R., & Marlinda, L. (2018). Comparison of SVM & Naïve Bayes Algorithm for Sentiment Analysis Toward West Java Governor Candidate Period 2018-2023 Based

on Public Opinion on Twitter. In *International Conference on Cyber and IT Service Management (CITSM 2018)* (pp. 3–8). Medan.

Li, L., Bai, Y., Song, Z., Chen, A., & Wu, B. (2018). Public transportation competitiveness analysis based on current passenger loyalty. *Transportation Research Part A: Policy and Practice*, *113*(April), 213–226. https://doi.org/10.1016/j.tra.2018.04.016

Mira, A., Izzaty, K., Mubarok, M. S., Informatika, F., Bandung, U. T., & Information, M. (2018). Klasifikasi Multi-Label pada Topik Ayat Qur'an Bahasa Inggris Menggunakan Tree Augmented Naive Bayes ( TAN ), *5*(1), 1–6.

Muthia, D. A. (2016). Opinion Mining Pada Review Buku Menggunakan Algoritma Naive Bayes. *Jurnal TEKNIK KOMPUTER*, *II*(1), 1–8. Retrieved from http://ejournal.bsi.ac.id/ejurnal/index.php/jtk/article/view/357

Novantirani, A., Sabariah, M. K., & Effendy, V. (2015). Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine (Vol. 2, pp. 1177–1183).

Nugroho, M. F., & Wibowo, S. (2017). Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes. *Jurnal Informatika Upgris*, *3*(1), 63–70.

Wahyudi, M., & Putri, D. A. (2016). Algorithm Application Support Vector Machine With Genetic Algorithm Optimization Technique for Selection Features for the Analysis of Sentiment on Twitter. *Journal of Theoretical and Applied Information Technology*, *84*(3), 321–331. Retrieved from www.jatit.org

Wahyuni, E. S. (2016). Penerapan metode seleksi fitur untuk meningkatkan hasil diagnosis kanker payudara. *Jurnal SIMETRIS*, *7*(1), 283–294.

**PROFIL PENULIS**

Riska Aryanti M.Kom.
Tahun 2016 lulus dari Program Strata Satu (S1) Jurusan Sistem Informasi STMIK Nusa Mandiri Jakarta. Tahun 2018 lulus dari Program Strata Dua (S2) Jurusan Magister Ilmu Komputer STMIK Nusa Mandiri Jakarta

Atang Saepudin, M.Kom.
Tahun 2016 lulus dari Program Strata Satu (S1) Jurusan Sistem Informasi STMIK Nusa Mandiri Jakarta. Tahun 2018 lulus dari Program Strata Dua (S2) Jurusan Magister Ilmu Komputer STMIK Nusa Mandiri Jakarta

Eka Fitriani, M.Kom.
Tahun 2016 lulus dari Program Strata Satu (S1) Jurusan Sistem Informasi STMIK Nusa Mandiri Jakarta. Tahun 2018 lulus dari Program Strata Dua (S2) Jurusan Magister Ilmu Komputer STMIK Nusa Mandiri Jakarta

Rifky Permana M.Kom.
Tahun 2014 lulus dari Program Strata Satu (S1) Jurusan Sistem Informasi STMIK Nusa Mandiri Jakarta. Tahun 2018 lulus dari Program Strata Dua (S2) Jurusan Magister Ilmu Komputer STMIK Nusa Mandiri Jakarta

Dede Firmansyah Saefudin, M.Kom.
Tahun 2013 lulus dari Program Strata Satu (S1) Jurusan Sistem Informasi STMIK Nusa Mandiri Jakarta. Tahun 2015 lulus dari Program Strata Dua (S2) Jurusan Magister Ilmu Komputer STMIK Nusa Mandiri Jakarta