JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER

VOL. 5. NO. 2 FEBRUARI 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864

ALGORITMA KLASIFIKASI *NAIVE BAYES* DAN *SUPPORT VECTOR MACHINE* DALAM LAYANAN KOMPLAIN MAHASISWA

Hermanto[1]; Ali Mustopa[2], Antonius Yadi Kuntoro[3]

Program Studi Teknologi Komputer[1]
Universitas Bina Sarana Informatika
www.bsi.ac.id
hermanto.hmt@bsi.ac.id

Program Studi Ilmu Komputer[2]
STMIK Nusa Mandiri
www.nusamandiri.ac.id
alimustopa.aop@bsi.ac.id; antonius.aio@nusamandiri.ac.id

Abstract— *Service in the world of education is an important element for the creation of an academic atmosphere that is conducive to the implementation of a successful teaching and learning process. The process of service to students there is a tendency to be implemented not following the minimum service standards that must be provided to students so that students tend to complain about the services provided. Submission of criticism, complaints, input, or suggestions for dissatisfaction and problems that exist in the university environment is still very limited. Complaints can be constructive if submitted to the right place and party. In this research the data processing of email complaints from students conducted at the academic student body (students.bsi.ac.id). Student complaint data that will be processed is data in the form of * .xls complaint file. Before text data is analyzed using text mining methods, the pre-processing text needs to be done including tokenizing, case folding, stopwords, and stemming. After pre-processing, the classification method is then performed in classifying each complaint category and dividing the status into two parts, namely complaint and not complaint so that the status becomes a normal condition in text mining research. The purpose of this study is to obtain the most accurate algorithm in the classification of student complaints and can find out the results of the classification of the Naïve Bayes algorithm method and Support vector Machine used and compared. In this study, the results of testing by measuring the performance of these two algorithms using Cross-Validation, Confusion Matrix, and ROC Curves. The obtained Support vector Machine algorithm has the highest accuracy value compared to Naïve Bayes. AUC value = 0.922. for the Support vector machine method using the student academic data collection dataset (students.bsi.ac.id) has 84.45%, from the Naïve Bayes algorithm has an accuracy rate of about 69.75% and AUC value = 0.679.*

Keywords: *Text Mining, Complaints Service, students, Support Vector Machine (SVM) algorithm, Naive Bayes.*

Intisari— Pelayanan dalam dunia pendidikan merupakan unsur penting untuk terciptanya suasana akademik yang kondusif untuk terlaksananya proses belajar mengajar yang sukses. Proses pelayanan terhadap mahasiswa ada kecenderungan dilaksanakan tidak sesuai dengan standar pelayanan minimal yang harus diberikan kepada mahasiswa sehingga mahasiswa cenderung mengeluh terhadap layanan yang diberikan. Penyampaian Kritikan, keluhan, masukan, atau saran terhadap ketidakpuasan dan permasalahan yang ada di lingkungan universitas masih sangat terbatas. Keluhan dapat bersifat membangun apabila disampaikan kepada pihak dan tempat yang tepat. Dalam Penelitian ini pengolahan data komplain email dari mahasiswa yang dilakukan pada sisfo akademik mahasiswa *(students.bsi.ac.id).* Data komplain mahasiswa yang akan diolah merupakan data berupa file komplain format **.xls,* Sebelum suatu data teks dianalisis menggunakan metode dalam *text mining* perlu dilakukan *pre processing text* diantaranya adalah *tokenizing*, *case folding*, *stopwords*, dan *stemming* (Asiyah & Fithriasari, 2016). Setelah dilakukan *pre processing* maka selanjutnya dilakukan metode klasifikasi dalam mengelompokkan dalam masing-masing kategori komplain dan membagi statusnya menjadi dua bagian yaitu *complaint* dan *not complaint* agar status menjadi kondisi normal dalam penelitian *text mining* . Tujuan penelitian ini adalah untuk mendapatkan algoritma yang paling akurat dalam klasifikasi komplain mahasiswa dan dapat mengetahui hasil klasifikasi dari metode algoritma Naïve Bayes dan *Support vector Machine* yang digunakan dan dibandingkan. Dalam penelitian ini, hasil pengujian dengan pengukuran kinerja kedua algoritma ini *menggunakan Cross Validation,*

Confusion Matrix dan *Kurva ROC*. Diperoleh algoritma *Support vector Machine* memiliki nilai akurasi tertinggi dibanding Naïve Bayes. Nilai AUC = 0,922. untuk metode *Support vector Machine* dengan menggunakan dataset sisfo akademik mahasiswa *(students.bsi.ac.id)* memiliki 84,45%, dari algoritma *Naïve Bayes* memiliki tingkat akurasi sekitar 69.75**%** dan nilai AUC=0.679**.**

Kata Kunci*: Text Mining,Layanan Komplain, mahasiswa, algoritma Support Vector Machine (SVM), Naive Bayes.*

PENDAHULUAN

Pelayanan dalam dunia pendidikan merupakan unsur penting untuk terciptanya suasana akademik yang kondusif untuk terlaksananya proses belajar mengajar yang sukses. Pada institusi pendidikan pelayanan prima kepada mahasiswa merupakan salah satu faktor yang perlu diperhatikan dengan baik untuk menjaga kelancaran studi mahasiswa . Proses pelayanan terhadap mahasiswa ada kecenderungan dilaksanakan tidak sesuai dengan standar pelayanan minimal yang harus diberikan kepada mahasiswa sehingga mahasiswa cenderung mengeluh terhadap layanan yang diberikan.

Pelayanan yang diberikan kepada pelanggan akan menjadi efektif apabila sesuai dengan keinginan dan harapan pelanggan. Untuk itu Perusahaan harus dapat menyesuaikan diri dengan kebutuhan pelanggan agar keluhan dari pelanggan yang berujung pada konflik yang berkepanjangan. Terjadinya kesenjangan emosi antara Perusahaan dengan pelanggan disebabkan oleh ketidak mampuan Perusahaan dalam menciptakan iklim yang kondusif dengan pelanggan(Irfani, 2014). Kualitas pelayanan dalam bidang pendidikan bukanlah hal yang dapat diperoleh dengan mudah dan tanpa usaha. Suatu jasa disebut berkualitas jika jasa tersebut mampu memenuhi kebutuhan dan memberikan kepuasan (Indriyani, Susi, 2016).

Penyampaian Kritikan, keluhan, masukan, atau saran terhadap ketidakpuasan dan permasalahan yang ada di lingkungan universitas masih sangat terbatas. Bahkan, tidak sedikit mahasiswa yang masih bingung harus ke mana untuk mengadukan keluhannya. Hal ini menyebabkan permasalahan yang ada hanya akan menjadi buah bibir di lingkungan universitas dan tak kunjung diproses. Sebagai contoh, kekurangan pembayaran kuliah dan hasil nilai ujian tidak sesuai dengan pengerjaanya. Kebanyakan mahasiswa, masih bingung harus melaporkan permasalahan ini ke mana. Selain itu, permasalahan lain muncul bila keluhan yang telah diutarakan baik lisan maupun tertulis tidak sampai kepada pengelola layanan universitas.

Keluhan yang diberikan oleh pelanggan dalam pemakaian produk atau jasa merupakan suatu umpan balik dari kualitas produk atau jasa yang digunakan oleh pelanggan. Semakin banyak keluhan yang diberikan oleh pelanggan membutuhkan adanya perhatian ekstra bagi setiap perusahaan yang menerima keluhan untuk memperbaiki dalam pembuatan produk atau jasa. Untuk itu setiap perusahaan harus mampu memberikan kepuasan kepada para pelanggannya dengan cara menyediakan produk yang mutunya lebih baik dan harga yang relatif terjangkau (Indriyani, Susi, 2016)

Jumlah keluhan yang tercatat yang berjumlah begitu besar tersebut dapat didefinisikan sebagai *Big Data. Big Data* merupakan data yang mempunyai jumlah dan variasi besar, serta bergerak cepat, sehingga melampaui kapasitas pengolahan database konvensional (Dumbill, 2014). Dalam mengolah *Big Data, Data Mining* merupakan metode yang dapat mengotomatisasi proses pengolahan data untuk mengekstraksi pengetahuan dari informasi yang tidak bisa diamati hanya dengan melihat data karena terlalu rumit atau multidimensi. Pada kasus data keluhan mahasiswa yang merupakan data teks, jenis metode Data Mining yang dapat digunakan adalah *Text Mining*. *Text Mining* memegang peran penting dalam analisis *Big Data* yang bersifat tidak terstruktur seperti data teks dan dalam jumlah yang sangat besar (Xiang, Schwartz, Gerdes, & Uysal, 2015).

Text mining sebenarnya merupakan bagian dari data *mining* dimana proses yang dilakukan utamanya adalah melakukan ekstraksi pengetahuan dan informasi dari pola-pola yang terdapat dalam sekumpulan dokumen teks menggunakan alat analisis tertentu (Monarizqa, Nugroho, & Hantono, 2014)

Pada umumnya keluhan tercipta sebagai akibat dari kejadian yang tidak diinginkan atau hal yang terjadi tidak sesuai harapan. Keluhan dapat bersifat membangun apabila disampaikan kepada pihak dan tempat yang tepat. Namun dapat menjadi isu negatif dan penebar kebencian apabila tidak disampaikan dengan tepat dan dikonsumsi oleh mahasiswa yang tidak memiliki dasar pengetahuan akan hal terkait.

Terdapat beberapa penelitian sebelumnya terkait komplain layanan yang dilakukan oleh beberapa peneliti seperti, Analisa Sentiment Untuk Opini Alumni Pada Perguruan Tinggi (Dharmendra, Saputra, & Pramaita, 2019). Klasifikasi Topik Keluhan Pelanggan Berdasarkan *Tweet* dengan Menggunakan Penggabungan *Feature* Hasil Ekstraksi pada Metode *Support Vector Machine* (Pratama & Trilaksono, 2015).

Klasifikasi Keluhan Menggunakan Metode *Support Vector Machine (SVM)* (Studi Kasus : Akun *Facebook Group iRaise Helpdesk)* (Basari, Hussin, Ananta, & Zeniarja, 2013). Penerapan *principal component analysis* dan *genetic algorithm* pada analisis sentimen *review* pengiriman barang menggunakan algoritma *support vector machine* (Rachmi, 2017).

Berdasarkan penelitian sebelumnya, algoritma *Naive Bayes* dan Support Vector Machine merupakan algoritma klasifikasi yang banyak digunakan oleh para peneliti dibidang *text mining*. Kedua algoritma tersebut digunakan dalam klasifikasi komplain mahasiswa dengan tujuan agar algoritma terpilih merupakan algoritma yang paling akurat sehingga dapat melakukam komplain mahasiswa. Dalam (Nurajijah & Riana, 2019), ditunjukkan bahwa SVM memiliki ketahanan dan kemampuan generalisasiyang lebih tinggi serta akurasi klasifikasi yang lebihstabil dibandingkan dengan algoritme yang lain. Dalam penelitian ini juga menggunakan validasi standar yaitu 10 *fold cross-validation* dimana proses ini membagi data secara acak ke dalam 10 bagian (Kusmira, 2019). Dari Proses pengujian data pada rapidminer dimulai dengan pembentukan model dengan data pada bagian pertama pembagian data training dan data testing. Setelah melakukan pengujian hasil Akurasi dapat diukur dengan menggunakan *confusion matrix*, dan *performance* diukur menggunakan *accuracy* dan *AUC* serta akan ditampilkan dalam bentuk *kurva ROC* untuk mengklasifikasikan teks pada komplain mahasiswa.

BAHAN DAN METODE

Teknik klasifikasi adalah sebuah model dalam *data mining* dimana *classifier* dikontruksi untuk memprediksi *categorical* label seperti "aman" atau "beresiko" untuk data aplikasi peminjaman uang, " ya" atau "tidak" untuk data marketing atau *"treatment A", "treatment B", "treatment C"* untuk data medis. Kategori tersebut dapat direpresentasikan dengan nilai yang sesuai dengan kebutuhannya (Vulandari, 2017). Klasifikasi merupakan tugas yang sama dengan *data mining*, dimana tujuan utama dari klasifikasi adalah prediksi label kelas.

Tiap teknik klasifikasi menggunakan suatu algoritma pembelajaran untuk mendapatkan suatu model yang paling memenuhi hubungan antara himpunan atribut dan label kelas dalam data masukan. Biasanya masukan dari model klasifikasi merupakan sekumpulan *record (training set).* Tiap *record* meliputi himpunan *attributes* yang salah satu atributnya merupakan *class.* Model untuk atribut kelas merupakan suatu fungsi dari nilai-nilai atribut lainnya. Suatu *test set* digunakan untuk menentukan keakuratan model tersebut. Biasanya dataset yang diberikan dibagi menjadi *training* dan *test sets,* dimana *training set* digunakan untuk membangun model dan *test set* digunakan untuk memvalidasi (Herawati, Fajar, 2013). Berikut ini algoritma klasifikasi yang akan digunakan dalam penelitian ini antara lain:

1. Support Vector Machine (SVM)

SVM merupakan metode klasifikasi untuk data linear dan nonlinier. Singkatnya, sebuah SVM adalah algoritma yang bekerja menggunakan pemetaan nonlinier untuk mengubah data pelatihan asli menjadi dimensi yang lebih tinggi. Dalam dimensi baru ini, ia mencari *hyperplane* yang memisahkan optik linear (yaitu, "batas keputusan" memisahkan *tuple* dari satu kelas dari kelas yang lain). Dengan pemetaan nonlinier yang tepat ke dimensi yang cukup tinggi, data dari dua kelas selalu dapat dipisahkan oleh *hyperplane*. SVM berusaha menemukan *hyperplan* menggunakan vektor dukungan ("esensial" pelatihan tuples) dan margin (ditentukan oleh vektor dukungan) (North, 2012).

Teknik ini termasuk dalam metode klasifikasi jenis terpandu (*supervised*) karena memiliki target pembelajaran tertentu. Klasifikasi dilakukan dengan mencari hyperplane atau garis pembatas (*decision boundary*) yang memisahkan antara satu kelas dengan kelas lainnya. Dalam konsep ini, SVM berusaha untuk mencari hyperplane terbaik diantara fungsi yang tidak terbatas jumlahnya. Fungsi yang tidak terbatas dalam pencarian hyperplane di metode *Support Vector Machine* merupakan sebuah keuntungan, dimana pemrosesan pasti akan selalu bisa dilakukan bagaimanapun data yang dimilikinya (North, 2012).

Berikut ini merupakan kekuatan dari *Support Vector Machine* (SVM) antara lain (Suyanto, 2017):

1) Mempunyai kemampuan generalisasi yang tinggi.
2) Mampu menghasilkan model klasifikasi yang baik meskipun dilatih dengan himpunan data yang relatif sedikit hanya dengan pengaturan parameter yang sederhana. SVM memiliki konsep dan formulasi yang jelas dengan sedikit parameter yang harus diatur.
3) Relatif mudah diimplementasikan karena penentuan SVM dapat dirumuskan dalam masalah QP (*Quadratic Programming*).

Sementara itu, kelemahan yang terdapat dalam *Support Vector Machine* (SVM) sebagai berikut:

1) Sulit diaplikasikan untuk himpunan data dengan jumlah sampel dan dimensi yang sangat besar.
2) Umumnya hanya diformulasikan untuk meyelesaikan masalah klasifikasi dua kelas. Walupun dapat dikembangkan untuk menyelesaikan masalah klasifikasi multi kelas, namun masing-masing strategi multi kelas SVM juga memiliki kelemahan.

2. *Naive bayes*

Algoritma s*Naive bayes* adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. Menurut Wu dan Kumar bahwa *Naive bayes* merupakan metode klasifikasi populer dan masuk dalam sepuluh algoritma terbaik dalam data mining. *Naive bayes* menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training (Mukminin & Riana, 2017).

Metode *Naive bayes* menempuh dua tahap dalam proses klasifikasi teks, yaitu tahap pelatihan dan tahap klasifikasi. Pada tahap pelatihan dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary*,yaitu kata yang mungkin muncul dalam koleksi dokumen sampel yang sedapat mungkin dapat menjadi representasi dokumen. Selanjutnya adalah penentuan probabilitas bagi tiap kategori berdasarkan sampel dokumen. Pada tahap klasifikasi ditentukan nilai kategori dari suatu dokumen berdasarkan term yang muncul dalam dokumen yang diklasifikasi (Hamzah, 2012).

HASIL DAN PEMBAHASAN

Business Understanding

Pada tahapan *business understanding*, dilakukan pemahaman terhadap objek penelitian. Dalam penelitian ini penulis menggunakan data komplain mahasiswa sebagai objek penelitian ini yang diambil dari database sisfo akademik *(students.bsi.ac.id)*. Motivasi pada tahap ini data komplain yang disajikan dalam bentuk teks pada database sisfo akademik mahasiswa *(students.bsi.ac.id)* yang akan dikelompokkan berdasarkan isi pembahasan dari masing-masing kategori komplain.

Pada tahap ini juga dilakukan pemahaman untuk mencari metode klasifikasi yang terbaik agar dapat membantu pada saat proses pengolahan data yang akan dilakukan dengan cara membandingkan hasil dari algoritma.

Data Understanding

Tahap ini adalah proses memahami data yang akan digunakan sebagai bahan yang akan diteliti untuk bisa dilakukan ke tahap setelahnya yaitu *Preprocessing*. Dibawah ini adalah langkah-langkah yang akan dikerjakan.

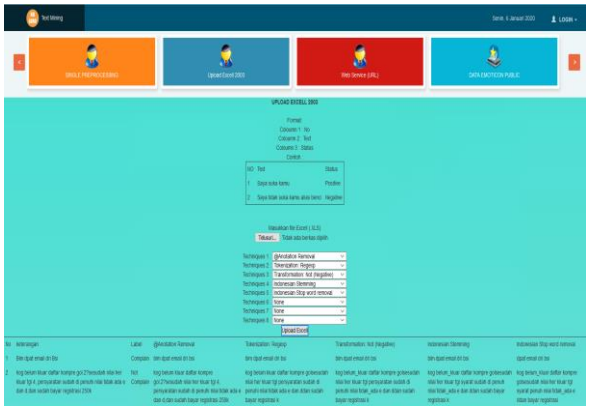
Menyiapkan data komplain email mahasiswa yang diambil dari database sisfo akademik *(students.bsi.ac.id)*. Data yang diambil sebanyak 8699 data, data yang didapat dari database sisfo akademik *(students.bsi.ac.id)*, maka dilakukan proses yang dinamakan *cleaning* data. Hal ini dilakukan guna menghapus *duplicate data* maupun yang tidak bisa digunakan dalam penelitian ini. Setelah dilakukan *cleaning* didapatkan data sebanyak 5954 data yang kemudian dilabeli statusnya berdasarkan kategori komplain menggunakan metode *Crowdsourced labelling* adalah metode pelabelan data yang melibatkan partisipasi khalayak umum, tentunya untuk *dataset* yang tidak membutuhkan keahlian khusus untuk melabelinya atau melakukan pembelajaran kepada partisipan dalam melakukan pelabelan (Rachmat & Lukito, 2016). Dengan pemberi label yang berjumlah banyak diharapkan proses pelabelan akan membutuhkan waktu yang lebih singkat (Rachmat & Lukito, 2016).

Biaya yang dikeluarkan untuk proses pemberian label pada model ini juga tidak sebanyak jika menggunakan bantuan ahli untuk melakukannya dan membagi statusnya menjadi dua bagian yaitu 3115 komplai e-mail dan 2941 bukan komplain agar status menjadi kondisi normal dalam penelitian *text mining*. Semua data komplain email tersebut dikelompokan menjadi satu baik itu *complain* atau *not complain* dan disimpan dalam bentuk ekstensi *.xls*

Data Preparation

Tahap selanjutnya adalah melakukan persiapan data sebelum data akan dilakukan modelling atau disebut dengan *Data Preparation.* Untuk tahap yang ke-2 ini yaitu mempersiapkan data untuk melakukan langkah-langkah yang disebut dengan *text preprocesing*, dengan menggunakan dua aplikasi *preprocessing*, pertama menggunakan *Gata Framework* yang diakses melalui *link http://gataframework.com/textmining* yang dapat digunakan secara gratis juga mudah dalam penggunaan dikarenakan tidak harus membuat *account* untuk memakai servicenya dan dilanjutkan *preprocessing* dari *rapidminer*, berikut adalah tahapannya:

Terakreditasi Sinta Peringkat 3 (Sinta 3) berdasarkan SK Dirjen Penguatan RisBang Kemenristekdikti
Nomor 28/E/KPT/2019, Tanggal 26 September 2019



Sumber:(Hermanto, Mustopa, & Kuntoro, 2019)

Gambar 2. Tampilan tools Gata *Framework Textmining*

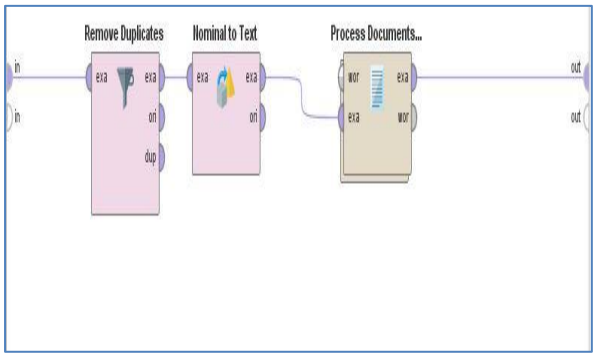
Gata *Framework* merupakan alternatif dalam *pre-processing* teks berbahasa Indonesia yang dikombinasikan dengan aplikasi *RapidMiner* untuk memproses kata-kata dalam bahasa Indonesia, hal ini dikarenakan dalam aplikasi RapidMiner sudah ada fasilitas kamus untuk mengubah akronim, dan *stopword*, tetapi masih terbatas pada bahasa *Inggris*, Cina, dan Arab, sedangkan untuk bahasa Indonesia masih belum tersedia. Dari hasil *pre-processing* dengan menggunakan Gata *Framework*, maka data set akan dilakukan *pre-processin*g lagi dengan menggunaan *tools RapidMiner* untuk membersihkan data agar lebih baik lagi hasilnya.

Remove Duplicates

Ini merupakan tahapan *data preparation* selanjutnya yang digunakan pada *software rapidminer*. *Remove duplicates* digunakan untuk menghilangkan *text* yang sama atau duplikat. Hal ini dilakukan agar data tidak dipenuhi oleh *text* yang sama sehingga memperlambat proses *running* software untuk menganalisa model.

Nominal to Text

Ini merupakn operator yang ada dalam rapidminer yang berfungsi untuk mengubah semua angka yang ada dalam *text* menjadi sebuah *text*. Sehingga angka yang ada akan dianggap jenis data text bukan *numeric* atau *nominal*. Gambar 3. memperlihatkan bagaimana penggunaan operator ini digunakan pada proses yang ada pada rapidminer.



Sumber:(Hermanto et al., 2019)

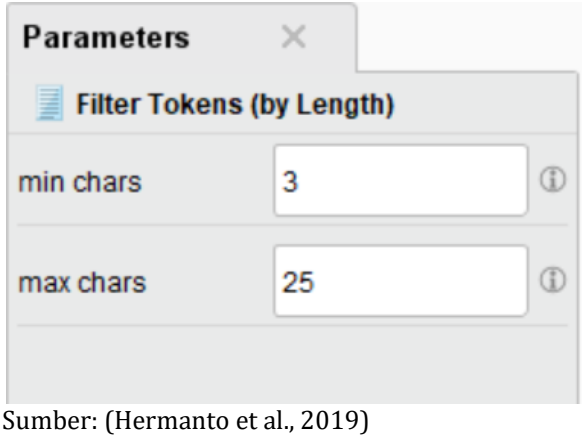
Gambar 3. Desain Model *Preprocessing Data Local* menggunakanoperator *Remove Duplicates* dan *Nominal to Text*

Transform Case.

Operator yang digunakan pada tahapan ini adalah untuk mengubah huruf kapital yang masih ada pada *text* akan diubah menjadi huruf kecil semua. Hal ini dilakukan agar ketikan dilakukan proses ke dalam model klasifikasi terdapat keseragaman huruf dan tidak terjadi kesalahan dalam proses *tokenize*.

Filter Token (by Length)

Ini adalah proses yang ada pada data preparation untuk menghilangkan sejumlah kata (setelah proses *tokenize*) dengan panjang karakter tertentu. Pada penelitian ini panjang minimum karakter yang digunakan adalah 3 karakter dan panjang maksimum 25 karakter. Artinya kata yang panjangnya kurang dari 3 karakter dan lebih dari 25 karakter akan dihilangkan. Untuk mendapatkan hasil seperti ini maka dilakukan setting pada Parameters dari operator ini (Gambar 4).



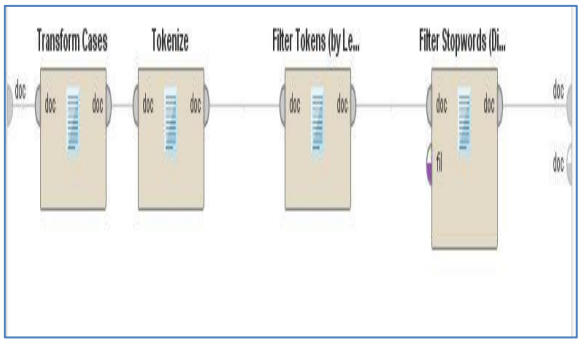
Sumber: (Hermanto et al., 2019)

Gambar 4. Parameters dari *Filter Tokens (by Length)*

Filter Stopword (Dictionary)

Selanjutnya adalah penggunaan operator *Stopword Removal (by Directory)* yang berfungsi untuk menghilangkan kata-kata yang tidak

hubungan dengan isi *text*. Pada tahapan sebelumnya dengan menggunakan *service text mining Gataframework* telah dilakukan namun ada beberapa kata yang belum dapat bisa dihilangkan oleh *service* sebelumnya karena belum dimasukkan sebagai kata yang harus dihapus. Maka dengan operator *Stopword Removal (by Directory)* peneliti dapat mendaftarkan kata yang harusnya dihapus dari *text*. Gambar 5. merupakan penjelasan dari penggunaan operator pada proses *rapidminer*.



Sumber:(Hermanto et al., 2019)

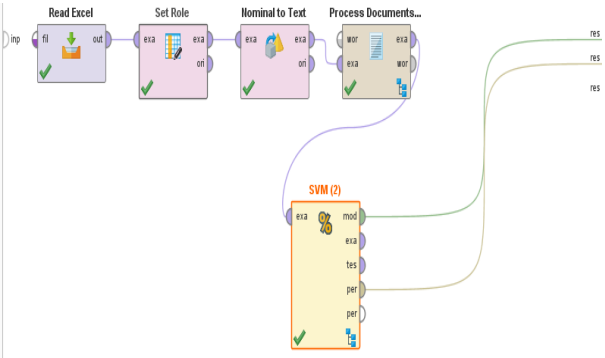
Gambar 5. Desain dari Penggunaan operator untuk *Data Preparation.*

Tahapan Pemodelan

Merupakan tahap pemilihan teknik mining dengan menentukan algoritma yang akan digunakan. *Tool* yang digunakan adalah *RapidMiner* versi 9.1. Hasil pengujian model yang dilakukan adalah mengklasifikasikan benar complain email dan tidak complain email menggunakan algoritma *Naive bayes* dan *Support Vector Machine* untuk mendapatkan nilai akurasi terbaik. Berikut adalah desain model *Rapidminer* yang digunakan yaitu :

Pengujian Model dengan Algoritma SVM

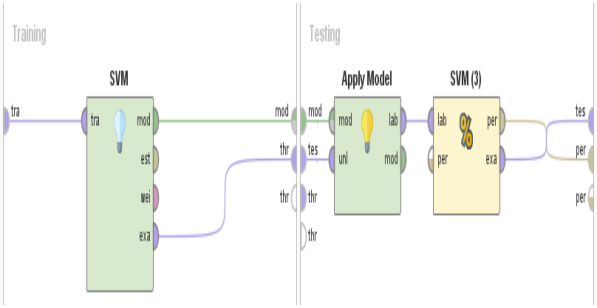
Pengaturan dan penggunaan operator serta parameter dalam *frameworks Rapid Miner* sangat berpengaruh terhadap akurasi dan model yang terbentuk, sebagai contoh dalam penggunaan model SVM Gambar 6.



Sumber:(Hermanto et al., 2019)

Gambar 6. Desain Model Algoritma SVM

Gambar diatas adalah model pengujian dari algoritma *support vector machine* (svm) menggunakan *rapidminer*, diawali dari memasukan data kemudian mengatur set *role* yang nantinya menentukan label disana dan nominal *text* lalu keproses dokumen yang berisikan seperti gambar 6. setelah itu barulah masuk kemodel perhitungan *support vector machine*nya seperti gambar 7.



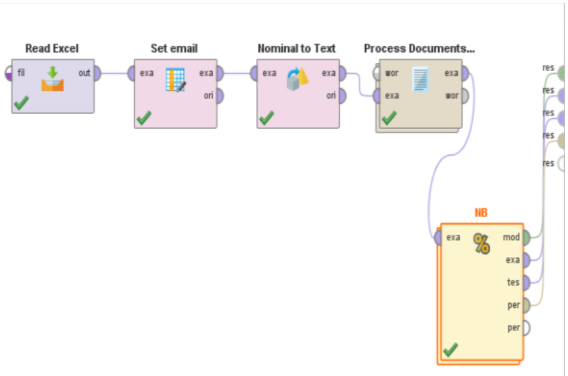
Sumber:(Hermanto et al., 2019)

Gambar 7. Desain Proses *10-Fold Cross Validation* untuk SVM

Gambar 7 Menjelaskan desain proses di dalam operator *cross validation SVM* pada gambar 6. Pada pengujian ini, data digunakan adalah data bersih yang telah melalui *preprocessing.* Data tersebut diambil dari operator *Read Excel*, hal ini dilakukan karena dataset disimpan dalam bentuk Excel (.xlsx). *Process documents from files* untuk mengkonversi *files* menjadi dokumen. Process validasi terdiri dari *data training* dan *data testing.* Kemudian masuk kemodel algoritmanya *support vector machine* didalamnya ada perhitungan algoritmanya kemudian modelnya *diapply* setelah itu masuk kepenilaian *performancenya* barulah muncul hasil nilai *accuracy* dan *aucnya.*

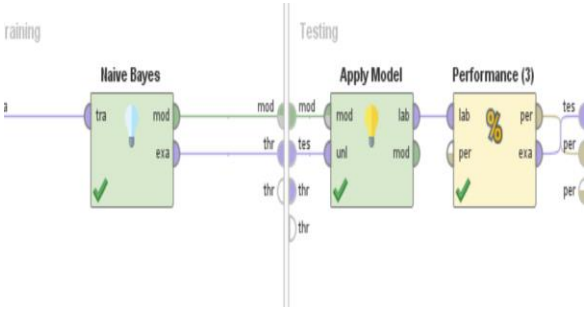
Pengujian Model dengan Algoritma *Naive bayes*

Pengaturan dan penggunaan operator serta parameter dalam *frameworks Rapid Miner* sangat berpengaruh terhadap akurasi dan model yang terbentuk, sebagai contoh dalam penggunaan model *Naive Bayes* seperti gambar 8.



Sumber: (Hermanto et al., 2019)
Gambar 8. Desain Model Algoritma NB

Gambar diatas adalah model pengujian dari algoritma *naive bayes* (NB) menggunakan *rapidminer*, diawali dari memasukan data kemudian mengatur set *role* yang nantinya menentukan label disana dan nominal *text* lalu keproses dokumen yang berisikan seperti gambar 8. setelah itu barulah masuk kemodel perhitungan *naive bayes* nya seperti gambar 9.



Sumber: (Hermanto et al., 2019)
Gambar 9. Proses 10-*Fold Cross Validation NB*

Gambar 9 Menjelaskan desain proses di dalam operator *cross validation naive bayes* pada gambar 8 Pada pengujian ini, data digunakan adalah data bersih yang telah melalui *preprocessing.* Data tersebut diambil dari operator *Read Excel*, hal ini dilakukan karena dataset disimpan dalam bentuk Excel (.xlsx). *Process documents from files* untuk mengkonversi *files* menjadi dokumen. Process validasi terdiri dari *data training* dan *data testing.* Kemudian masuk kemodel algoritmanya *naive bayes* didalamnya ada perhitungan algoritmanya kemudian modelnya di*apply* setelah itu masuk kepenilaian *performancenya* barulah muncul hasil nilai *accuracy* dan *aucnya.*

Model Evaluasi

Tahapan evaluasi bertujuan untuk menentukan nilai kegunaan dari model yang telah berhasil dibuat pada langkah sebelumnya. Untuk evaluasi digunakan *10-fold cross validation*. Dari hasil pengujian model dari dua algoritma yang dipakai adalah untuk menghasilkan sebuah nilai *Accuracy (Confusion Matrix)* dan *AUC (Area Under Curve)*. Maka mendapatkan hasil grafik ROC dengan nilai *AUC (Area Under Curve).*

Nilai Accuracy dari Algoritma SVM

Dari hasil pengujian model diatas dengan menggunakan algoritma SVM maka dapat menghasilkan sebuah nilai *Accuracy (Confusion Matrix)* yang dapat dilihat pada Tabel 1.

Tabel 1. Nilai Accuracy Algoritma SVM

Accuracy : 84.45% +/- 2.89% (*micro average*: 84.45%)

	true Complain	*true NotComplain*	*class precision*
pred. Complain	527	117	81.83%
pred. NotComplain	68	478	87.55%
class recall	88.57%	80.34%	

Sumber: (Hermanto et al., 2019)

$$\text{Acc}\,(Accuracy) = \frac{TP + TN}{TP + TN + FP + FN} + \frac{527 + 478}{527 + 117 + 68 + 478} = \frac{1005}{1190} = 0{,}844$$

Jumlah *True Complain* (TP) adalah 527 *record* diklasifikasikan sebagai *Complain* dan *False Not Complain* (FN) adalah 68 *record* diklasifikasikan sebagai *Not Complain*. Berikutnya 117 *False Complain* diklasifikasikan sebagai *Complain* dan 478 *record True Not Complain* diklasifikasikan sebagai *Not Complain*. Berdasarkan tabel 4.1 diatas menunjukan bahwa, tingkat akurasi dengan menggunakan algoritma SVM adalah sebesar 84,45%.

Nilai Accuracy dari Algoritma *Naive bayes*

Dari hasil pengujian model diatas dengan menggunakan algoritma *naive bayes* maka dapat menghasilkan sebuah nilai *Accuracy (Confusion Matrix)* yang dapat dilihat pada tabel 2.

Tabel 2. Nilai Accuracy Algoritma *Naive bayes*

Accuracy : 69.75% +/- 2.89% (*micro average*: 69.75%)

	true Complain	*true NotComplain*	*class precision*
pred. Complain	524	289	64.45%
pred. NotComplain	71	306	81.17%
class recall	88.07%	51.43%	

Sumber: (Hermanto et al., 2019)

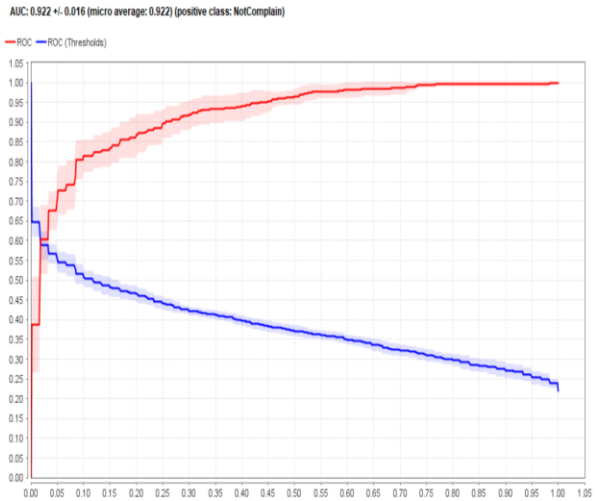
$$\text{Acc}\,(Accuracy) = \frac{TP + TN}{TP + TN + FP + FN} + \frac{524 + 289}{591 + 8 + 4 + 587} = \frac{813}{1190} = 0{,}69{,}75$$

Jumlah *True Complain* (TP) adalah 524 *record* diklasifikasikan sebagai *Complain* dan *False Not Complain* (FN) adalah 71 *record* diklasifikasikan sebagai *Not Complain*. Berikutnya 289 *Complain False* diklasifikasikan sebagai *Complain* dan 306 *record True Not Complain* diklasifikasikan sebagai *Not Complain*. Berdasarkan tabel 2. diatas menunjukan bahwa, tingkat akurasi dengan

menggunakan algoritma *Naive bayes* adalah sebesar 69.75 %.

Nilai *AUC* dari Algoritma SVM

Berikut ini akan dijelaskan Kurva ROC dan Confusion Matrix dari algoritma *Support Vector Machine*:



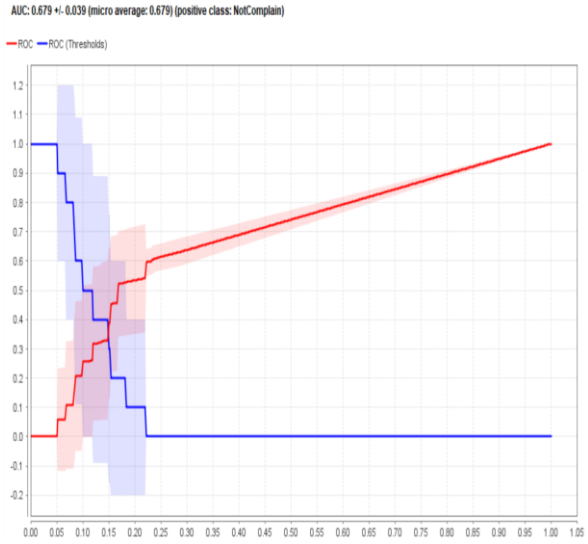
Sumber: (Hermanto et al., 2019)

Gambar 10. Nilai AUC dalam Algoritma SVM

Dari hasil pengujian model yang telah dilakukan adalah untuk mendapatkan hasil akurasi dan *Area Under Curve (AUC)*. Maka mendapatkan hasil grafik ROC dengan nilai *Area Under Curve (AUC)* sebesar 0,922 dengan *performance* akurasi yaitu *excellent*.

Nilai *AUC* dari Algoritma *Naive bayes*

Berikut ini akan dijelaskan *Kurva ROC* dan *Confusion Matrix* dari algoritma *Naive bayes*:



Sumber: (Hermanto et al., 2019)

Gambar 11. Nilai AUC dalam Algoritma *Naive Bayes*

Dari hasil pengujian model yang telah dilakukan adalah untuk mendapatkan hasil akurasi dan *Area Under Curve (AUC)*. Maka mendapatkan hasil grafik ROC dengan nilai *Area Under Curve (AUC)* sebesar 0.679 dengan performance akurasi yaitu *failure*.

Perbandingan *Accuracy*

Berdasarkan hasil analisis dari masing-masing algoritma diatas, maka dapat dirangkum hasilnya seperti tabel 3.

Tabel 3. Perbandingan Performance Algoritma

	Support Vector Machine	***Naive Bayes***
Akurasi	**84,45%**	69.75%
AUC	**0,922**	0.679

Sumber: (Hermanto et al., 2019)

KESIMPULAN

Dalam penelitian ini setelah dilakukan *preprocessing* dan dilakukan pengujian model dengan membandingkan dua metode data mining yaitu *support vector machine (*SVM) dan *Naive Bayes*, hasil evaluasi dan validasi, diketahui bahwa nilai akurasi untuk menentukan bahwa komplain email mahasiswa tersebut ya komplain email dan tidak komplain email, dapat dibuktikan dengan nilai akurasi dan nilai *AUC* dari masing-masing algoritma yaitu untuk SVM nilai akurasi = 84,45% dan nilai *AUC* = 0,922, sedangkan untuk algoritma *Naive Bayes* nilai akurasi = 69.75%. dan nilai *AUC* = 0.679. Dalam penelitian dapat diketahui bahwa tingkat akurasi yang didapatkan algoritma Support Vector Machine lebih unggul dibanding dengan Naive Bayes. Pada penelitan (Dharmendra et al., 2019) dengan menggunakan SVM dengan data lain dihasilkan akurasi 75,76%. Untuk itu, penerapan Support Vector Machine pada peneltian ini memiliki akurasi yang lebih tinggi sehingga dapat digunakan untuk memberikan solusi terhadap permasalahan analisis sentimen pada komplain mahasiswa.

REFERENSI

Asiyah, S. N., & Fithriasari, K. (2016). Klasifikasi Berita Online Menggunakan Metode Support Vector Machine Dan K-Nearest Neighbor. *Jurnal Sains Dan Seni ITS*, *5*(2), 317–322. https://doi.org/10.12962/j23373520.v5i2.16643

Basari, A. S. H., Hussin, B., Ananta, I. G. P., &

Zeniarja, J. (2013). OpinionMining of Movie Review Using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. *Procedia Engineering*, *53*, 453–462. https://doi.org/10.1016/j.proeng.2013.02.059

Dharmendra, I. K., Saputra, K. O., & Pramaita, N. (2019). Analisa Sentiment Untuk Opini Alumni Pada Perguruan Tinggi. *Majalah Ilmiah Teknologi Elektro*, *18*(2), 227–234. Retrieved from https://ocs.unud.ac.id/index.php/JTE/article/view/48059

Dumbill, E. (2014). Volume, Velocity, Variety: What You Need to Know About Big Data. Retrieved from https://www.forbes.com/sites/oreillymedia/2012/01/19/volume-velocity-variety-what-you-need-to-know-about-big-data/

Hamzah, A. (2012). Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis. *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III*, 269– 277.

Herawati, Fajar, A. (2013). *Data Mining*. Yogyakarta, Indonesia: Andi Offset.

Hermanto, Mustopa, A., & Kuntoro, A. Y. (2019). *Hasil Akhir Penelitian Mandiri: Algoritma Klasifikasi Naive Bayes Dan Support Vector Machine Dalam Layanan Komplain Mahasiswa*. Jakarta, Indonesia.

Indriyani, Susi, S. M. (2016). Pengaruh Penanganan Keluhan (Complaint Handling) Terhadap Kepercayaan Dan Komitmen Mahasiswa Pada Perguruan Tinggi Swasta Di Bandar Lampung. *Jurnal Bisnis Darmajaya*, *2*(1), 1–13. Retrieved from https://jurnal.darmajaya.ac.id/index.php/JurnalBisnis/article/view/615/

Irfani, E. (2014). Prediksi Keluhan Pelanggan Pada Apartemen Menggunakan Algoritmac4.5. *Jurnal Paradigma*, *16*(2), 13–20. Retrieved from https://ejournal.bsi.ac.id/ejurnal/index.php/paradigma/article/view/773

Kusmira, M. (2019). ANALISIS SENTIMEN REGISTRASI ULANG KARTU SIM PADA TWITTER MENGGUNAKAN ALGORITMA SVM DAN K-NN | INTI Nusa Mandiri. *INTI Nusa Mandiri*, *14*(1), 105–110. Retrieved from http://ejournal.nusamandiri.ac.id/index.php/inti/article/view/541/

Monarizqa, N., Nugroho, L. E., & Hantono, B. S. (2014). Penerapan Analisis Sentimen Pada Twitter Berbahasa Indonesia Sebagai Pemberi Rating. *Jurnal Penelitian Teknik Elektro Dan Teknologi Informasi*, *1*, 151–155.

Mukminin, A., & Riana, D. (2017). Komparasi Algoritma C4 . 5 , Naïve Bayes Dan Neural Network untuk Klasifikasi Tanah. *Jurnal Informatika*, *4*(1), 21–31. Retrieved from https://pdfs.semanticscholar.org/fa81/c97fcc8eb80c32922b710dd20f7c3fad70d4.pdf

Nurajijah, & Riana, D. (2019). Algoritma Naïve Bayes, Decision Tree, dan SVM untuk KlasifikasiPersetujuan Pembiayaan Nasabah Koperasi Syariah. *Jurnal Teknologi Dan Sistem Komputer*, *7*, *no*(10.14710/jtsiskom.7.2.2019), 77–82.

Pratama, E. E., & Trilaksono, B. R. (2015). Klasifikasi Topik Keluhan Pelanggan Berdasarkan Tweet dengan Menggunakan Penggabungan Feature Hasil Ekstraksi pada Metode Support Vector Machine (SVM). *JEPIN*, *1*(2), 53-59. Retrieved from https://www.researchgate.net/profile/Riyanto_Bambang/publication/318962570_Klasifikasi_Topik_Keluhan_Pelanggan_Berdasarkan_Tweet_dengan_Menggunakan_Penggabungan_Feature_Hasil_Ekstraksi_pada_Metode_Support_Vector_Machine_SVM/links/59949897458515c0ce653243/

Rachmat, A., & Lukito, Y. (2016). Implementasi Sistem Crowdsourced Labelling Berbasis Web dengan Metode Weighted Majority Voting. *Jurnal ULTIMA InfoSys*, *6*(2), 76–82. https://doi.org/10.31937/si.v6i2.223

Rachmi, H. (2017). Penerapan principal component analysis dan genetic algorithm pada analisis sentimen review pengiriman barang menggunakan algoritma support vector machine. *Jurnal Evolusi*, *5*(2). Retrieved from https://ejournal.bsi.ac.id/ejurnal/index.php/evolusi/article/view/3130

Suyanto. (2017). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung, Indonesia: Informatika.

Vulandari, R. (2017). *Data Mining Teori dan*

Aplikasi Rapidminer. Surakarta, Indonesia: Penerbit Gava Media.

Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, *44*, 120–130. https://doi.org/10.1016/j.ijhm.2014.10.013

Terakreditasi Sinta Peringkat 3 (Sinta 3) berdasarkan SK Dirjen Penguatan RisBang Kemenristekdikti
Nomor 28/E/KPT/2019, Tanggal 26 September 2019