Analisis Prediksi Rating Store berdasarkan Ulasan Pelanggan dengan Metode Supervised Learning (SVM, KNN, dan Naïve Bayes), da pembobotan kata TF-IDF

Ryan Ramdhani[1], Yanuar Firdaus A.W[2], Ibnu Asror[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]ryanrmd@students.telkomuniversity.ac.id,
[2]yanuar@telkomuniversity.ac.id,
[3]iasror@telkomuniversity.ac.id

Abstrak

Data ulasan pelanggan memiliki peran penting bagi seorang calon pembeli. Selain menjadi bahan pertimbangan sebelum membeli barang tersebut, calon pembeli dapat melakukan riset soal produk yang akan dibeli melalui ulasan online yang ada. Hal ini dikarenakan calon pembeli tidak dapat menyentuh, mencoba, ataupun melihat secara langsung. Sehingga pembeli hanya mengandalkan deskripsi produk dan membaca ulasan pelanggan yang sebelumnya telah membeli produk tersebut. Para penjual merasa terbantu dengan adanya ulasan pelanggan selain meningkatkan kepercayaan calon pembeli, juga dapat meninjau ulang produk yang akan dijual ke pelanggan. Dikarenakan penjual tidak hanya menjual produknya di satu merchant, terdapat data ulasan pelanggan yang belum dicatat otomatis. Oleh karena itu, penelitian ini bertujuan untuk memprediksi rating toko menggunakan supervised learning untuk menghasilkan model yang memiliki performansi yang optimal. Metode yang digunakan ialah pemrosesan teks yang merupakan bagian dari machine learning dengan metode Naïve Bayes karena hanya membutuhkan jumlah data pelatihan yang kecil untuk menentukan estimasi parameter yang diperlukan. SVM karena dapat melakukan komputasi dengan cepat dalam menentukan jarak support vector dan KNN karena sederhana dan juga tahan terhadap data yang memiliki derau. Ketiga Metode tersebut digunakan untuk memprediksi rating store dengan membandingkan data actual dan data predict yang telah di rata-rata kan sebelumnya menggunakan pembobotan kata TF-IDF. Hasil penelitian menunjukkan nilai Akurasi K-Nearest Neighbor memiliki tingkat akurasi 95.01%, *Support Vector Machines* memiliki tingkat akurasi 95.37%, dan *Naïve Bayes* memiliki tingkat akurasi sebesar 95,44%. Pada hasil penelitian ini bahwa Algoritma SVM dan Naïve Bayes dapat membantu permasalahan dalam memprediksi rating store berdasarkan ulasan pelanggan dan model yang digunakan dapat meningkatkan penjualan.

Kata kunci: Machine Learning, SVM, KNN, Naïve Bayes, Pemrosesan Teks

Abstract

Customer review data plays an important role for a potential buyer. In addition to being a consideration before buying the item, prospective buyers can do research on the products to be purchased through online reviews. This is because prospective buyers cannot touch, try, or see directly. So buyers rely solely on product descriptions and reading customer reviews that have previously purchased those products. Sellers find it helpful to have customer reviews in addition to increasing the trust of potential buyers, can also review products to be sold to customers. Because sellers don't only sell their products at one merchant, there's customer review data that hasn't been automatically recorded. Therefore, this study aims to predict store ratings using supervised learning to produce models that have optimal performance. The method used is text processing which is part of machine learning with Naïve Bayes method because it only requires a small amount of training data to determine the estimated parameters needed. SVM because it can compute quickly in determining the distance of vector support and KNN because it is simple and also resistant to data that has noise. The three methods are used to predict store ratings by comparing actual data and predict data that have been averaged before using tf-IDF word weighting. The results showed the K-Nearest Neighbor Accuracy value had an accuracy rate of **95.01%, *Support Vector Machines*** had an accuracy rate of **95.37%, and *Naïve Bayes* had an accuracy rate of 95.44%. In the results of this study,** SVM and **Naïve Bayes algorithms can help problems in predicting store ratings based on customer reviews and the model used can increase sales.**

Keywords: Machine Learning, SVM, KNN, Naïve Bayes, Text Processing

1. Pendahuluan

Latar Belakang

Kehadiran layanan penjualan digital atau *e-commerce*, semakin diminati oleh masyarakat *millennial*, khususnya di ibukota. *E-commerce* pun juga menawarkan beberapa promosi seperti hal nya *discount*, *cashback*, dan juga kredit hingga 12 bulan, menjadi alasan *e-commerce* lebih dipilih masyarakat dibandingkan *offline store*. Berdasarkan laporan iPrice, tiga besar *e-commerce* yang mendominasi pasar Indonesia adalah Tokopedia, Shopee, dan Bukalapak [1].

Marketplace yang akan dijadikan bahan penelitian penulis ialah Tokopedia. *Official Store* yang akan penulis analisis ialah Vyatta, Pinzy, Naxen dan Jete. Fokus utama dari ketiga *Official Store* tersebut yaitu menjual *gadget* dan elektronik terbaru. Dalam kegiatan pembelian barang di marketplace, pembeli mampu mengimbuhkan ulasan sehabis menerima barang yang dibeli. Ulasan pembelian produk terdiri berasal dari bintang dan mengisi komentar ulasan yang memuat tanggapan, apresiasi, komentar kekecewaan karena barang tidak cocok maupun kritik dan masukan terhadap produk yang udah dibeli tersebut [2]. Semakin banyak kuantitas bintang yang dimiliki, maka jadi baik pula reputasi yang dimiliki produk berikut [3]. Ulasan pembelian product miliki dampak yang penting terhadap minat membeli dari konsumen lain [4]. Ulasan pembelian produk dapat digunakan oleh penjual untuk mendapat informasi untuk bahan perbaikan pada produk dan fasilitas supaya tercipta kepuasan pelanggan. Kepuasan pelanggan merupakan hal perlu yang jadi tujuan perusahaan [5].

Analisis ulasan secara ringan sanggup dijalankan dengan memandang jumlah bintang yang diberikan oleh pembeli, tetapi jumlah bintang tidak sanggup mewakili isi berasal dari keseluruhan ulasan. Diperlukan memandang seluruh isi komentar ulasan untuk sanggup mengetahui keseluruhan maksud ulasan. Sangat tidak barangkali untuk menganalisis ulasan secara manual dengan memandang satu persatu, tetapi bila ulasan yang dimiliki banyak akan lebih cepat mengfungsikan proses asumsi sentimen. Pada penelitian ini peneliti melaksanakan asumsi sentimen pada ulasan product di marketplace Tokopedia (Vyatta, Pinzy, Naxen dan Jete). Bagian ulasan produk terdiri dari mengisi komentar dengan format teks bebas dan peringkat jumlah bintang dari 1 sampai 5. Isi komentar ulasan digunakan untuk memahami Info yang jadi fokus konsumen di dalam mengimbuhkan ulasan. Isi komentar ulasan bisa memuat lebih dari satu penilaian atribut produk, namun tiap-tiap ulasan cuma miliki satu penilaian jumlah bintang, supaya jumlah bintang tidak bisa mewakili tiap-tiap fitur produk yang dinilai oleh pembeli. Informasi yang disampaikan konsumen bisa merujuk terhadap fitur produk layaknya harga, kualitas, bahan, warna, bentuk, ukuran, rasa, jumlah, maupun terhadap pelayanan yang diberikan layaknya pengemasan, lama pengiriman, dan respon penjual.

Dikarenakan penjual tidak hanya menjual produknya di satu merchant, alias di berbagai merchant *online* maupun sosial media, terdapat data ulasan pelanggan yang belum dicatat otomatis. Oleh karena itu, penelitian ini ditujukan untuk penjual dalam memprediksi rating toko mereka menggunakan *supervised learning*. Karena data ulasan merupakan bahan pertimbangan bagi seorang calon pembeli dalam melakukan riset mengenai produk yang akan dibeli melalui *online*. Dapat juga meningkatkan kepercayaan calon pembeli terhadap barang yang akan dibeli di *store* tersebut.

Konsep yang ingin penulis lakukan adalah membuat analisis prediksi *rating* dari masing-masing *store* dari *marketplace* Tokopedia (Vyatta, Pinzy, Naxen dan Jete). Dengan bermodalkan ulasan dari para pengguna platform yang telah membeli produk tersebut, yang selanjutnya akan dijadikan *dataset* oleh penulis. Jika data ulasan tersebut sudah memiliki label dan memiliki skor tiap-tiap terhadap tiap-tiap ulasannya. Maka, pemanfaatan metode supervised learning merupakan metode yang sesuai dalam persoalan ini. Pada permasalahan diatas peneliti ingin membagi suatu opini kedalam opini positif dan opini negatif pada website Tokopedia (Vyatta, Pinzy, Naxen dan Jete). Metode *Supervised Learning* yang digunakan adalah *Support Vector Machine* (SVM) alasan pemilihan algoritma ini karena Kemampuan generalisasi SVM untuk mengklasifikasikan suatu pattern, yang tidak termasuk data yang dipakai dalam fase pembelajaran metode[6], *K-Nearest Neighbo*rs (KNN) alasan pemilihan algoritma ini karena KNN memiliki beberapa kelebihan yaitu bahwa dia tangguh terhadap training data yang noisy dan efektif apabila training data-nya besar[7], dan *Naïve Bayes* alasan pemilihan algoritma ini karena *Naïve Bayes* ini memanfaatkan teori probabilitas yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya[8]. Ide untuk menggunakan metode *Supervised Learning* ini berasal dari studi sebelumnya dan untuk membandingkan performansi dari setiap algoritma.

Topik dan Batasannya

Berdasarkan latar belakang tersebut, perumusan kasus dalam penulisan penelitian ini sebagai berikut.

a. Bagaimana melakukan prediksi rating store pada Tokopedia menggunakan SVM, KNN, dan *Naïve Bayes*?
b. Bagaimana tingkat akurasi dari SVM, KNN, dan *Naïve Bayes* dalam kasus prediksi rating store?

Tugas Akhir ini memiliki batasan dan ruang lingkup penelitian yang mencangkup:

a. Dataset ulasan yang digunakan diambil dari *store* Vyatta, Pinzy, Naxen dan Jete di *marketplace* Tokopedia.
b. Metode yang digunakan dalam proses *learning* ialah *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Naïve Bayes*.
c. Bahasa yang digunakan dalam dataset ulasan pengguna ialah Bahasa Indonesia dan *English.*
d. Pengumpulan dataset dilakukan dengan menggunakan cara *scraping* (Octoparse)
e. Jumlah dataset yang dilakukan untuk penelitian ini berjumlah 4678 baris terdiri dari 3 kolom yaitu ulasan, bintang dan nama.

Tujuan

Berdasarkan masalah yang ada, tujuan yang ingin dicapai dalam tugas akhir ini dapat dijabarkan Sebagai berukut :

a. Menganalisis performa model Supervised learning dalam prediksi rating store
b. Membandingkan performa dari model *Supervised* Learning yang digunakan menggunakan TF-IDF

Organisasi Tulisan

Tugas Akhir ini disusun bersama struktur Sebagai berikut. Setelah dijelaskan pendahuluan terhadap anggota pertama, terhadap anggota kedua mencantumkan studi terkait, setelah itu dijelaskan pemodelan sistem terhadap anggota ketiga, sehabis itu, dijelaskan evaluasi terhadap sistem yang dibangun terhadap anggota ketiga. Oada anggota kelima dijelaskan pemikiran dan anjuran untuk penelitian setelah itu.

2. Studi Terkait

2.1 *Opinion Mining*

Opinion Mining merupakan studi yang melakukan analisis terhadap suatu opini setiap orang dengan melihat *sentiment*, evaluasi, perilaku, ataupun emosi yang terkandung dalam suatu produk, pelayanan, organisasi, individu, isu, kejadian, topik, dan setiap atributnya. Opini tersebut kemudian dikelompokkan berdasarkan polaritasnya, polaritas dapat berbentuk opini positif, dan opini negatif [9].

Selain dari segi polaritas, *Opinion Mining* juga dibagi menjadi tiga berdasarkan level klarifikasinya, yaitu dokumen level, kaluat level, dan fitur level atau aspek. Yang akan dibahas dalam tugas akhir ini termasuk analisis *sentiment* level dokumen, yaitu opini di ekstrak dari review dan dilakukan klasifikasi terhadap *review* tersebut berdasarkan polaritasnya. Klasifikasi level dokumen cocok untuk data yang ditulis oleh seseorang dan memuat opini atau sentimen dari orang tersebut [9].

2.2 *Supervised Learning Methods*

Clustering atau klasterisasi adalah suatu Teknik atau metode untuk mengelompokkan data. Menurut Tan, 2006 clustering adalah sebuah proses untuk mengelompokkan data ke dalam sebagian cluster atau grup sehingga data dalam satu cluster miliki tingkat kemiripan yang maksimum dan data antar cluster miliki kemiripan yang minimum [10].

Clustering merupakan sistem partisi satu set objek data ke dalam himpunan anggota yang disebut bersama dengan cluster. Objek yang di dalam cluster punyai kemiripan karakteristik antar satu sama lainnya dan berlainan bersama dengan cluster yang lain. Partisi tidak dijalankan secara manual melainkan bersama dengan suatu algoritma clustering. Oleh karena itu, clustering benar-benar berguna dan mampu menemukan grup atau grup yang tidak dikenal dalam data. Clustering banyak digunakan dalam beraneka aplikasi seperti seumpama terhadap business intelligence, pengenalan pola citra, web site search, bidang ilmu biologi, dan untuk keamanan (security). Di dalam business intelligence, clustering mampu mengatur banyak pembeli ke dalam banyaknya grup Contohnya mengelompokan pembeli ke dalam lebih dari satu cluster bersama dengan kesamaan karakteristik yang kuat. Clustering juga dikenal sebagai data segmentasi karena clustering mempartisi banyak data set ke dalam banyak grup berdasarkan kesamaannya. Selain itu clustering juga bias sebagai outlier detection [10].

2.3 *E-Commerce*

E-commerce didefinisikan sebagai sistem pembelian, penjualan, mentransfer atau bertukar produk, jasa atau informasi melalui jaringan computer melalui Internet. Dengan mengambil bentuk-bentuk tradisional dari sistem bisnis dan mengfungsikan jejaring sosial melalui internet, langkah bisnis sanggup berhasil kecuali dijalankan bersama benar, yang pada akhirnya menghasilkan peningkatan pelanggan, kesadaran brand dan pendapatan. Keputusan pembelian pelanggan dipengaruhi oleh persepsi, motivasi, pembelajaran, sikap dan keyakinan. Persepsi dipantulkan ke terhadap bagaimana pelanggan memilih, mengatur, dan menginterpretasikan informasi untuk membentuk pengetahuan. Motivasi tercermin keinginan pelanggan untuk mencukupi keperluan mereka sendiri [11].

2.4 *Marketplace*

Marketplace merupakan model E-Business yang terkait dengan penjual dan customer (seller & buyer). Marketplace di Indonesia merupakan keliru satu tempat penggerak ekonomi nasional di dalam rangka menghadapi jaman globalisasi. Untuk itu, harus dikembangkan Marketplace yang teratur, wajar dan efisien. Pada kebanyakan Marketplace yang efektif dapat menaikkan iklim investasi di perusahaan dan memudahkan arus input dan output barang [12].

2.5 Tokopedia

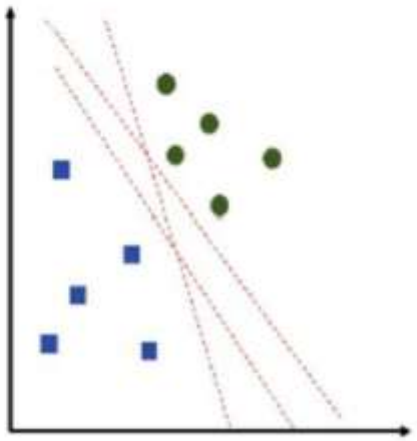
Tokopedia adalah salah satu perusahaan jual membeli berbasis digital terbesar di Indonesia. Tokopedia resmi diluncurkan ke publik pada 17 Agustus 2009 di bawah naungan PT Tokopedia yang didirikan oleh William Tanuwijaya dan Leontinus Alpha Edison pada 6 Februari 2009. Sejak resmi diluncurkan, PT Tokopedia berhasil jadi salah satu perusahaan e-commerce di Indonesia dengan perkembangan yang benar-benar pesat. Dengan mengusung model usaha marketplace dan mall online, Tokopedia terlalu mungkin tiap-tiap individu, toko kecil, dan brand untuk membuka dan mengelola toko online. Tokopedia punyai visi untuk "Membangun Indonesia yang Lebih Baik Lewat Internet", Tokopedia membawa program untuk menunjang para pelaku Usaha Mikro Kecil Menengah (UMKM) dan perorangan untuk mengembangkan usaha mereka dengan memasarkan produk secara online [13].

2.6 *Support Vector Machine* (SVM)

Support Vector Machine adalah metode klasifikasi untuk membagi dua kelas berdasarkan pembagian hyperplane[16]. Penggunaan kernel pada metode SVM bisa menyelesaikan problem information yang non-linear dengan memetakan information ke dimensi yang lebih besar. Proses untuk mencari hyperplane pada Support Vector Machine menurut Campbell sebagai berikut [14]:

1. Terdapat data $xi \vec{} \in (x1 \vec{}, x2 \vec{}, \ldots x\vec{}n)$ yang mana xi adalah data yang terdiri dari n atribut dan dua kelas $yi \in +1,-1$.
2. asumsi data linear dan kelas antara +1 dan -1 dapat dipisah oleh hyperplane. Kondisi ini dapat didefinisikan pada (1) :

$$w\vec{}.x + b = 0 \ (1)$$



Gambar 1Visualisasi Hyperplane

Dari persamaan di atas akan diperoleh (2) dan (3):

$$w\vec{}.x + b \geq 1, untuk\ kelas + 1 \ (2)$$

$$w^{\rightarrow}.x + b \geq -1, untuk\ kelas - 1\ (3)$$

Di mana $w^{\rightarrow}$ sebagai weight, x sebagai data input, b adalah posisi bidang relatif.

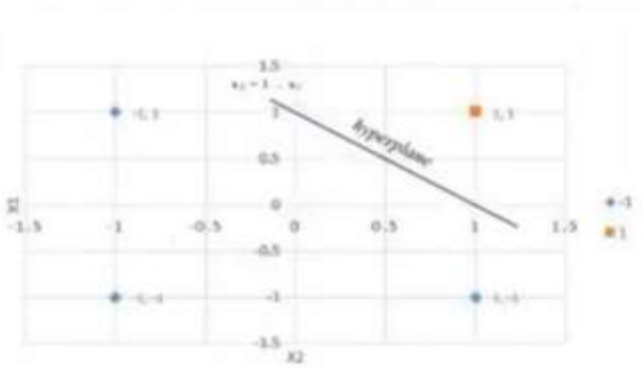
3. Dalam menemukan hyperplane yang optimal, maka diperlukan mencari hyperplane pemisah yang memaksimalkan jarak antara dua kelas. Untuk itu pencarian titik minimal diperlukan. Adapun mencari titik minimal dapat dilakukan dengan (4):

$$minw1\ 2\ (\|w^{\rightarrow}\|)2\ (4)$$

dengan kendala pada (5) berikut:

$$yi(xi^{\rightarrow}, w^{\rightarrow} + b) - 1 \geq 0\ (5)$$

Di mana $xi^{\rightarrow}$ adalah data input ke-i, $w^{\rightarrow}$ adalah weight, b adalah posisi bidang relatif, dan yi adalah kelas target ke-i.



Gambar 2 Hasil Pencarian Hyperplane

4. Penentuan Kelas

Penentuan kelas ini dilakukan berdasarkan perbandingan jumlah nilai dari atribut yang dikandung menggunakan persamaan berikut.

$$sign(x1 + x2 - 1)\ (6)$$

2.7 *K-Nearest Neighbors* (KNN)

Metode K-NN merupakan suatu metode untuk klasifikasi, metode ini lakukan klasifikasi terhadap suatu object yang berdasar kepada information training yang membawa jarak yang paling dekat berasal dari object tersebut. Semua ruangan yang digunakan adalah sebagai information klasifikasi untuk sample training. Ketika sebuah sample baru digunakan untuk test, sesudah itu dapat dihitung jarak antara sample test dan setiap sample training. Sample tesnya adalah diberikan terhadap klas yang memiliki kumpulan berasal dari lebih dari satu jarak K yang paling sedikit [15]. Rumus euclidean distance digunakan untuk mengitung jaraknya, nilai k yang terbaik pada metode ini bergantung pada data. Diketahui bahwa nilai k yang tinggi mampu mengurangi pengaruh noise di dalam klasifikasi, tetapi mampu menyebabkan batas antar tiap-tiap klasifikasi menjadi kabur. Sehingga di dalam penentuan nilai k yang pas mampu pengaruhi hasil klasifikasi information terbaik.

K-Nearest Neighbour (KNN), dapat mengklafisifikasikan citra uji ke dalam kelas bersama dengan jumlah bagian terbanyak. Prinsip kerja KNN adalah melacak jarak terdekat antara knowledge yang dapat dievaluasi bersama dengan k tetangga (neighbor) terdekatnya dalam data pelatihan [16].

KNN adalah salah satu metode klasifikasi yang paling digemari banyak orang dikarenakan kesederhanaan dan keefektifannya yang masuk akal itu tidak memerlukan pemasangan model dan sudah terbukti baik kinerja untuk mengklasifikasikan beraneka model data.

$$k = \sqrt{\sum_{i=1}^{n}(a-b)^2}\ (7)$$

2.8 *Naïve Bayes*

Naive Bayes merupakan metode pengklasifikasian probabilistik sederhana. Metode ini dapat menghitung sekumpulan probabilitas bersama menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Metode naive bayes menganggap semua atribut pada setiap kategori tidak memiliki ketergantungan satu sama lain (independen) [17].

Keuntungan penggunaan Naive Bayes yaitu cuma butuh sejumlah kecil information latih untuk memilih parameter mean dan variansi berasal dari variabel yang diperlukan untuk klasifikasi [18]. Naive Bayes merupakan metode supervised document classification yang bermakna memerlukan knowledge training sebelum meakukan sistem klasifikasi.

$$P(R|S) = \frac{P(R)P(S|R)}{P(s)} \quad (8)$$

2.9 Pembobotan TF-IDF

Term frequence (TF) dan Inverse document Frequency (IDF) adalah pembobotan yang paling sering digunakan [19]. Metode TF-IDF merupakan langkah untuk mencari bobot suatu kata (term) pada sebuah dokumen. Metode TF-IDF mencampurkan dua langkah untuk perhitungan bobotnya, yaitu bersama menghitung frekuensi kemunculan kata di sebuah dokumen khusus (TF) dan laksanakan perhitungan invers pada frekuensi dokumen yang mengandung kata berikut (IDF) [19].

Perhitungan invers document frequency (IDF) digunakan untuk mengkalkulasi jumlah term yang bermanfaat sebagai ukuran tingkat signifikansi suatu term dalam sebuah dokumen. Perhitungan TF dan IDF sanggup dilihat terhadap Persamaan 9 dan 10.

$$tf(i) = \frac{freq(t_1)}{\sum freq(t)} \quad (9)$$

Keterangan:

tf(i) : nilai *Term Frequency* sebuah kata dalam sebuah dokumen.

Freq (t_i) : frekuensi kemunculan sebuah kata dalam sebuah dokumen.

$\sum freq\ (t)$: jumlah keseluruhan kata dalam dokumen.

Sementara IDF (Inverse Document Frequency) mengkalkulasi logaritma berasal dari kuantitas semua dokumen dan dibandingkan bersama kuantitas dokumen di mana dalam dokumen berikut kata (t) yang dimaksud muncul. Berikut persamaan 2.10 yang digunakan untuk mengkalkulasi IDF.

$$tdf(i) = log \frac{|D|}{|\{d:t_i \in d|} \quad (10)$$

Keterangan:

idf(i) : nilai *Inverse Document Frequency* sebuah kata di seluruh isi dokumen.

|D| : jumlah seluruh dokumen.

|(d: ti∈ d}| : jumlah dokumen yang mengandung kata (t).

2.10 Confusion Matrix

Untuk menguji hasil klasifikasi terhadap sistem yang sudah dibangun, maka diperlukan suatu metode perhitungan evaluasi performasi yakni bersama dengan mengkalkulasi nilai precission, recall dan f1-measure. Dalam evaluasi performasi ini dapat dihitung nilai akurasi dan F1-Measure. Akurasi adalah bagaimana suatu sistem sanggup lakukan klasifikasi true terhadap knowledge true dan false, tetapi f1-measure untuk menilai performasi dari keseluruhan sistem bersama dengan mengkalkulasi nilai precision dan recall. Adapun perhitungan sanggup dilihat terhadap confussion matrix [20]:

Tabel 1 *Confusion Matrix*

		Nilai Prediksi	
		Positive	Negative
Nilai Aktual	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

TP (True Positive) merupakan prediksi positif dan nilai sesungguhnya positif, TN (True Negative) merupakan prediksi negatif dan nilai sesungguhnya negatif, FP (False Positive) merupakan prediksi positif dan nilai sesungguhnya negatif dan FN (False Negative) merupakan prediksi negatif dan nilai sesungguhnya positif.

$$Precission = \frac{TP}{TP+FP} \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (13)$$

$$F1 - Measure = 2 * \frac{precission*recall}{precission+recall} \quad (14)$$

2.10 *Pre-processing*

Metode preprocessing sangat penting dan berpengaruh pada model klasifikasi, karena setiap metode yang dilakukan akan menghasilkan performansi yang variatif saat dilakukannya penelitian [21]. Proses ini memungkinkan data mentah diproses menjadi lebih cepat dalam penelitian karena sudah dinormalisasikan. Sebelum data diolah, data akan dibersihkan dengan beberapa proses. Pada penelitian ini pertama kali yang dilakukan ialah Case Folding yang berfungsi untuk menghapus simbol atau angka pada teks dan mengubah seluruh kalimat menjadi lowercase agar memudahkan computer melakukan komputasi. Setelah itu dilakukan proses Stop Removal dengan tujuan untuk menghapus kata yang tidak berpengaruh pada suatu kalimat, karena terdapat kata yang tidak perlu atau tidak memiliki bobot seperti contoh "yang", "ia", "ya", "nggak" dan sebagainya. Dan terakhir dilakukan proses Stemming yang bertujuan untuk merubah kata menjadi bentuk dasar dari suatu kata. Untuk setiap proses yang digunakan penelitian ini menggunakan Python Library bernama Sastrawi.

3. Rancangan Sistem Prediksi Rating Store berdasarkan ulasan pelanggan

Sub-bab ini menjelaskan langkah-langkah sistem dalam memproses data sampai mendapatkan analisis prediksi rating dengan menggunakan tiga algoritma yang berbeda.

Gambar 3 Gambaran umum sistem.

Berdasarkan Gambar 3.1, untuk menghasilkan prediksi dari *rating store* maka akan melalui beberapa tahap. Tahap awal ialah pengumpulan data ulasan pelanggan dari pengguna di *store* dari *website marketplace* tokopedia 'https://www.tokopedia.co.id' [22]. Data yang sudah terkumpul akan di proses pada tahap *pre-processing*, untuk membuat data lebih terstruktur dan rapi. Data selajutnya dapat langsung dilakukan *split data* tersebut. Penulis juga sudah menyiapkan data train dan juga data test. Selanjutnya data di proses menggunakan metode supervised (Naïve Bayes, Support Vector Machine (SVM), dan K-Nearest Neighbors (KNN)). Performa dari setiap algoritma tersebut akan dilakukan perbandingan dan ditinjau mana yang lebih baik performa nya dengan menggunakan pembobotan dengan TF-IDF sehingga penulis dapat menyimpulkan hasil dari penelitian ini.

4. Evaluasi

4.1 Hasil Pengujian

A. Dataset

Data yang digunakan pada penelitian ini diambil dengan cara *Scraping* dari *website marketplace* tokopedia. Ulasan diambil dari empat *store* yaitu Vyatta, Pinzy, Naxen dan Jete. Contoh data hasil *crawling* ada pada tabel 2.

Tabel 2 Contoh Hasil *Crawling Data*

No	Ulasan	Bintang	Nama
1	Barang sdh sampai dan sesuai pesanan, puas saya, trims gam	5	Vyatta
2	Barang bagus, packing tidak safety.	5	Vyatta
3	Barang sesuai.	5	Vyatta
4	Katanya waterproof. Tapi kena air dikit langsung error.	4	Vyatta
5	Good product and delivery	5	Vyatta

B. Data Pre-Processing

Data hasil crawling dijadikan dataset dan dikelompokan berdasarkan *feature* untuk diberi bobot. Pada tahap ini data tersebut dilakukan *pre-processing* untuk memperbaiki struktur dan menghindari data yang tidak sempurna. Pada data ulasan tersebut dengan kata yang kurang dari 20.

a) *Case Folding*

Setiap kata pada ulasan yang ada di dataset akan diubah menjadi *lowercase*. Contoh proses *case folding* pada kasus set data ini ada pada tabel 2.

Tabel 3 Contoh Hasil *Case Folding*

Ulasan (*Before*)	Ulasan (*After*)
Barang sdh sampai dan sesuai pesanan, puas saya, saya trims gam	Barang sdh sampai dan sesuai pesanan, puas saya, saya trims gam

b) *Stop-word Removal*

Kata yang tidak memiliki ketergantungan suatu topik atau tanda baca akan dihapus karena tidak relevan pada perhitungan bobot pada model *Clustering*. Hasil kalimat yang telah dilakukan proses *stop-word removal* terdapat pada tabel 3.

Tabel 4 Contoh Tabel *Stop-word Removal*

Ulasan (*Before*)	Ulasan (*After*)
Barang sdh sampai dan sesuai pesanan, puas saya, saya trims gam	Barang sdh sampai dan sesuai pesanan puas saya saya trims gam

c) *Tokenization*

Proses *tokenization* dilakukan untuk memisahkan sebuah kalimat menjadi kata-kata atau disebut token. Tahap ini dilakukan untuk mempermudah proses menyamakan kata pada kalimat ulasan yang nantinya akan digunakan kata nya saja, bukan kalimat. Hasil dari *tokenization* pada suatu ulasan terdapat pada tabel 4.

Tabel 5 Contoh Tabel Tokenization

Ulasan (*Before*)	Ulasan (*After*)
Barang sdh sampai dan sesuai pesanan puas saya saya trims gam	Barang, sdh, sampai, dan, sesuai, pesanan, puas, saya, saya , trims, gam

d) *Stemming*

Tahap terakhir yang dilakukan pada *pre-processing* yaitu *stemming*. Untuk proses *clustering* dengan model berdasarkan *feature* yang telah dibuat, setiap kata harus diubah ke akar kata tersebut menjadi kata dasar. Imbuhan yang ada pada kata dihapus sehingga kata tersebut berubah menjadi kata dasar. Pada ulasan, kata yang digunakan adalah kata dasar. Contoh hasil dari proses *stemming* terdapat pada tabel 5.

Tabel 6 Contoh Tabel *Stemming*

Ulasan (*Before*)	**Ulasan (*After*)**
Barang, sdh, sampai, dan, sesuai, pesanan, puas, saya, saya, trims, gam	Barang, sampai, sesuai, pesan, puas, saya, saya

C. Pembobotan TF-IDF

TF-IDF merupakan salah satu metode yang digunakan untuk melakukan pembobotan kata. Kata yang akan digunakan pada tahap pembobotan kata adalah kata hasil proses stemming. Adapun tahap pembobotan kata atau TF-IDF adalah menentukan banyaknya kemunculan term *t* pada setiap dokumen, kemudian menghitung $df_i, idf_i, W_{tf_{td}}, W_{td}$ hingga normalisasi. Berikut adalah proses penentuan banyaknya term *t* pada setiap dokumen ditunjukan pada Tabel .

Tabel 7 Hasil Perhitungan Term Untuk Setiap Dokumen

TF		Doc ke-i
Nomor	Term	Doc1
1	barang	1
2	sampai	1
3	sesuai	1
4	pesan	1
5	puas	1
6	saya	2

Pada perhitungan TF-IDF data yang akan digunakan adalah data training. Sedangkan perhitungan TF-IDF untuk data testing akan menggunakan df_t dan idf_t yang dihasilkan pada data training. df_t merupakan jumlah dokumen yang mengandung term *t* dan idf_t merupakan kebalikan dari df_t. Berikut adalah perhitungan df_t dan idf_t menggunakan data training ditunjukan pada Tabel .

$$idf_t = log_1\left(\frac{N}{df_i}\right)$$

$$idf_{saya} = log_1\left(\frac{N}{df_{box}}\right)$$

$$idf_{saya} = log_1\left(\frac{2}{1}\right)$$

$$idf_{saya} = log_1(2)$$

$$idf_{saya} = 0{,}301$$

Nilai TF-IDF dari sebuah term *t* merupakan perkalian antara $w_{tf_{t,d}}$ dan idf_t. TF-IDF dapat dihitung sebagai berikut.

$w_{t,d} = w_{tf_{t,d}} \, x \, idf_t$
$w_{1,1} = 1 \, x \, 0{,}301$
$w_{1,1} = 0{,}301$

D. Klasifikasi Naïve Bayes

Model ini dibuat untuk memperoleh hasil prediksi menggunakan algoritma Naïve Bayes berdasarkan ulasan pengguna pada *store* tersebut, dan perolehan skor bintang yang telah disediakan. Pada proses ini menggunakan *supervised* yaitu *Naïve Bayes*. Berikut adalah proses menggunakan algoritma *naive bayes*.

Tabel 8 Hasil Dataset

Kategori Ulasan	Jumlah	Perkiraan Rating
Positif	3	≥ 3 bintang
Negatif	3	≤ 2 bintang

Diketahui :

- P Positif (klasifikasi positif)

  = Jumlah positif : (jumlah postif + jumlah negatif)
  = 3 : (6)
  = 0,5
- P Negatif (klasifikasi negatif)

  = Jumlah negatif : (jumlah postif + jumlah negatif)
  = 3 : (6)
  = 0,5

Tabel 9 Jumlah Kelas Data Ulasan

Kategori Ulasan	Jumlah
Positif	3
Negatif	3

- Data ulasan baru :

Tabel 10 Data Ulasan

mustopa produk sampe di tujuan dengan baik.

Kata kunci = produk, baik

Tabel 11 Jumlah Kata Kunci

nama	produk	baik
mustopa	1	1

- Proses :

$$P(Positif) = \frac{((Sum\ C) + 1)}{(3 + 2)} = 0{,}6$$

Karena jumlah kata kunci ada 2, maka hasil dari kelas positif dikalikan dengan banyaknya jumlah kata kunci yaitu 2, sehingga kelas positif memiliki nilai 0,18.
Berikut adalah hasil dari perbandingan kedua klasifikasi pada data ulasan baru.

Tabel 12 Prediksi Kelas Positif

nama	produk	baik	klasifikasi
mustopa	1	1	
positif	0.6	0.6	0.18

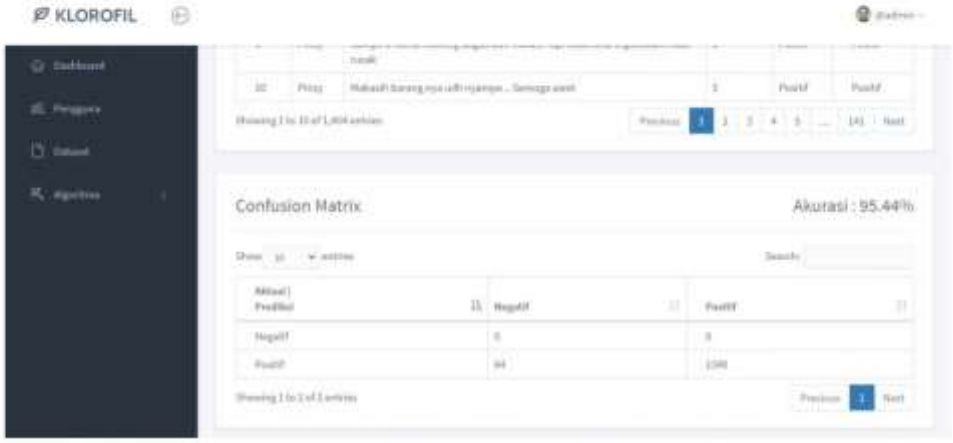
Tabel 13 Prediksi Kelas Negatif

nama	rusak	kecewa	klasifikasi
mustopa	0	0	
negatif	0.2	0.2	0.02

Berdasarkan data perbandingan kedua kelas diatas, maka dapat disimpulkan bahwa data ulasan terbaru termasuk dalam kategori positif yang kemungkinan akan memberikan rating (bintang) ≥ 3 bintang.

c. Hasil Klasifikasi Naïve Bayes Program

Hasil Algoritma Naïve Bayes menampilkan hasil klasifikasi dari 1.404 data testing dan menggunakan metode confusion matrix Sebagai metode evaluasi hasil algoritma maka didapatkan data prediksi negative sebanyak 64, prediksi positif sebanyak 0, aktual negative sebanyak 0 dan aktual positif sebanyak 1340 maka hasil akurasi dari algoritma Naïve Bayes sebesar 95,44%



Gambar 4 Halaman Hasil Algoritma Naive Bayes

Sistem ini digunakan untuk mengukur ketepatan data dalam menganggap presisi (p) dan recall (r) untuk menghitung skor yang didapat. Dalam menghitung prediksi pengklasifikasian digunakan Confusion Matrix yang berguna mendapatkan hasil data mana saja yang sesuai dengan data aktual.

Tabel 14 Confusion Matrix Naïve Bayes

x	Prediksi Negatif	Prediksi Positif
Aktual Negatif	TN = 0	FP = 0
Aktual Positif	FN = 64	TP = 1340

Akurasi – merupakan proporsi jumlah total prediksi yang benar dengan persamaan :

$$Akurasi = \frac{1340 + 0}{(0 + 64 + 0 + 1340} * 100\% = 95{,}44\%$$

Precision – Proporsi kumppulan data relevan yang diprediksi benar dengan persamaan:

$$Precision = \frac{1340}{(0 + 1340)} * 100\% = 100\%$$

Recall – Proporsi kumpulan data relevan yang diidentifikasi dengan benar, melalui persamaan :

$$Recall = \frac{1340}{(64 + 1340)} * 100\% = 95\%$$

Setelah mendapatkan nilai dari Precision dan Recall maka didapat F – Measure yang digunakan untuk mengoptimalkan nilai nilai keduanya yang terdapat distribusi kelas dengan nilai tidak rata[11], dapat diketahu melalui persamaan :

$$F1 - Measure = 2 * \frac{0.95 * 1}{0.95 + 1} * 100\% = 97\%$$

$F1{-}Measure{=}2{*}0.95 * 10.95{+}1$

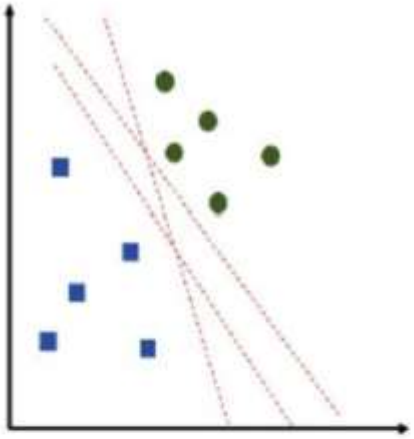
d. Klasifikasi SVM

Support vector machine adalah model yang memiliki garis pemisah dari data yang dikumpulkan. Untuk memahami proses *support vector machine* dalam klasifikasi prediksi *rating*, berikut adalah prosesnya.

1.) Hyperline

Hyperline digunakan untuk menentukan garis pemisah dan visualisasi data awal.

Data awal dapat dianalogikan seperti gambar dibawah dimana hyperline belum dibentuk (diketahui), maka berikut adalah proses pencarian hyperline dalam SVM :



Gambar 5 Visualisasi Hyperline

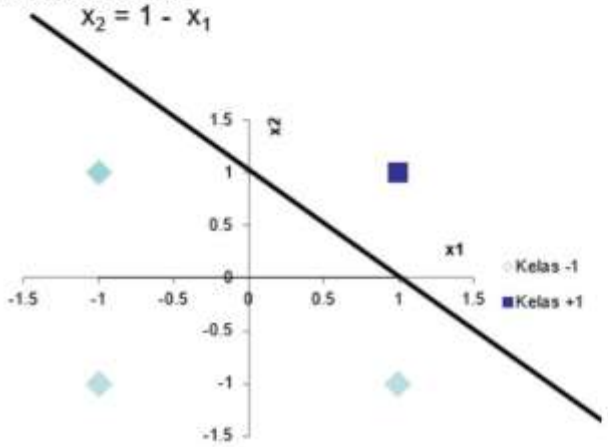
Keterangan :

- Titik hijau = data kelas positif
- Titik biru = data kelas negatif

Tabel 15 Variabel Kelas

(variabel)		Kelas
x1	x2	y
1	1	1
1	-1	-1
-1	1	-1
-1	-1	-1

Dari tabel diatas maka dapat dibuat hyperline dimana x1 = 1 dan x2 = 1 maka y = 1 dapat terbentuk hyperline seperti berikut :



Gambar 6 *New Hyperline*

Keterangan :

- xi = variabel
- y = kelas

yi(X1.W1 + X2.W2 + b) ≥1
y1(1.W1 + 1.W2 + b) ≥1
y1(W1 + W2 + b) ≥1

Maka dilakukan penjumlahan silang seperti berikut :

(W1 + W2 + b) ≥1 (1)
(-W1 + W2 - b) ≥1 (2)
(W1 - W2 - b) ≥1 (3)
(W1 + W2 - b) ≥1 (4)

Maka persamaan (1) dan (2) :

(W1 + W2 + b) ≥1
(-W1 + W2 - b) ≥1
2W2 = 2
W2 = 1

Maka persamaan (1) dan (3) :

(W1 + W2 + b) ≥1
(W1 - W2 - b) ≥1
2W1 = 2
W1 = 1

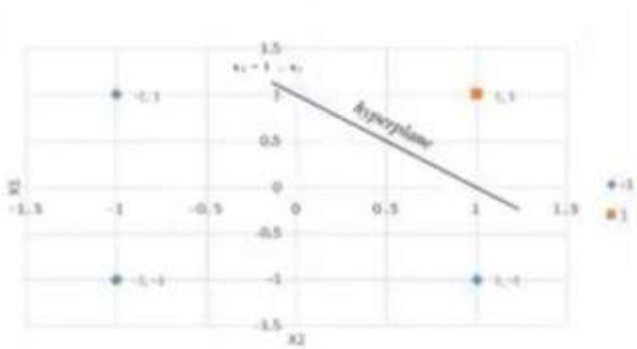
Maka persamaan (2) dan (3) :

(-W1 + W2 - b) ≥1
(W1 - W2 - b) ≥1
-2b = 2
B = -1

(w1.x1 + w2.x2 + b = 0)
1.x1 + 1.x2 – 1 = 0
x1 + x2 – 1 = 0
x2 = 1 – x1 = 3

Tabel 16 Pencarian Hyperline

x1	x2 = 1 – x1
-2	3
-1	2
0	1
1	0
2	-1



Gambar 7 Visualisasi Hasil Pecarian Hyperline

Hyperlane digunakan untuk menentukan garis jarak antar kelas yang telah ditentukan.

2.) Hasilkan Kelas
Diketahui :
f(x) = x1 + x2 – 1
kelas = sign(f(x))
Data :
X1 = jumlah kriteria pada X1 masuk ke kelas (-)
X2 = jumlah kriteria pada X2 masuk ke kelas (+)

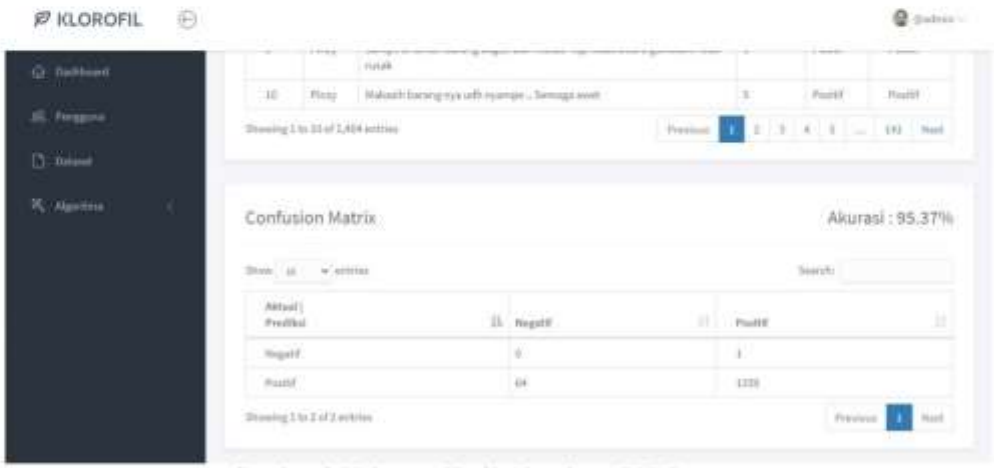
Tabel 17 Data Kelas Bentuk Hyperline

No.	Data Ujian		Hasil Klasifikasi	Kelas
	X1	X2	Kelas = sign (x1 + x2 -1)	
1	-1	5	Sign (1 + 5 – 1) = +1	Positif
2	-1	4	Sign (-1 + 4 – 1) = +1	Positif
3	0	7	Sign (0 + 7 – 1) = +1	Positif
4	-9	0	Sign (-9 + 0 – 1) = -1	negatif
5	-2	-2	Sign (2 – 2 -1) = -1	negatif

- Jika kelas (+) maka dikategorikan dalam klasifikasi rating A (inisialisasi sebuah kelas/klasifikasi) dimana kemungkinan bintang yang diberikan adalah ≥ 3 bintang.
- Jika kelas (-) maka dikategorikan dalam klasifikasi rating B (inisialisasi sebuah kelas/klasifikasi) dimana kemungkinan bintang yang diberikan adalah ≤ 2 bintang.

e. Hasil Klasifikasi SVM Program

Hasil Algoritma SVM menampilkan hasil klasifikasi dari 1.404 data testing dan menggunakan metode confusion matrix Sebagai metode evaluasi hasil algoritma maka didapatkan data prediksi negative sebanyak 64, prediksi positif sebanyak 1, aktual negative sebanyak 0 dan aktual positif sebanyak 1339 maka hasil akurasi dari algoritma SVM sebesar 95,37%



Gambar 8 Halaman Hasil Algoritma SVM

Tabel 18 Confusion Matrix SVM

x	Prediksi Negatif	Prediksi Positif
Aktual Negatif	TN = 0	FP = 1
Aktual Positif	FN = 64	TP = 1339

Akurasi – merupakan proporsi jumlah total prediksi yang benar dengan persamaan :

$$Akurasi = \frac{(1339 + 0)}{(0 + 64 + 1 + 1339)} * 100\% = 95.37\%$$

Precision – Proporsi kumppulan data relevan yang diprediksi benar dengan persamaan:

$$Precision = \frac{1339}{(1+1339)} * 100\% = 99\%$$

Recall – Proporsi kumpulan data relevan yang diidentifikasi dengan benar, melalui persamaan :

$$Recall = \frac{1339}{(64+1339)} * 100\% = 95\%$$

Setelah mendapatkan nilai dari Precision dan Recall maka didapat F – Measure yang digunakan untuk mengoptimalkan nilai nilai keduanya yang terdapat distribusi kelas dengan nilai tidak rata[11], dapat diketahu melalui persamaan :

$$F1 - Measure = 2 * \frac{0.99 * 0.95}{0.99 + 0.95} * 100\% = 96\%$$

f. Klasifikasi KNN

K-Nearest Neighbor adalah algoritma dengan prinsip ketetanggaan terdekat, berikut adalah proses dari algoritma *k-nearest neighbor* dalam penelitian ini.

K-Nearest Neighbor (KNN)

Tabel 19 Data Training

No	Kelas	Kriteria			
		proses	produk	rusak	baik
1	Positif	3	4	0	5
2	Negatif	2	1	3	1

Tabel 20 Data Testing

No.	Nama	proses	produk	fungsi	baik
1	X	2	3	0	2

Proses hitung kelas positif :

$$\sqrt{(3-2)^2 + (4-3)^2 + (0-0)^2 + (5-2)^2} = 3{,}31$$

Proses hitung kelas negatif :

$$\sqrt{(2-2)^2 + (1-3)^2 + (3-0)^2 + (1-2)^2} = 3{,}74$$

Hasil KNN :

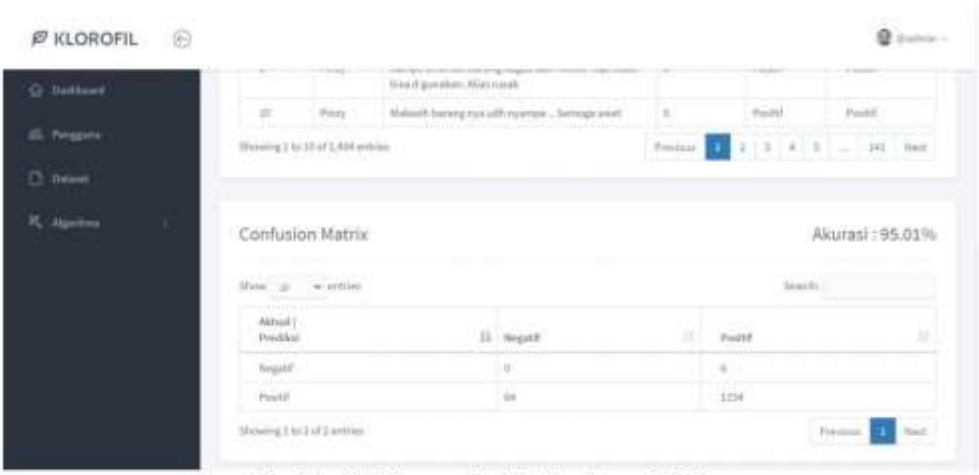
Tabel 21 Hasil KNN

No	Kelas	Kriteria				Square Distance of Query DIstance	Rangking
		Proses	Produk	Rusak	Baik		
1	Positif	3	4	0	5	3,31	1
2	Negatif	2	1	3	1	3,74	2

Dari hasil tersebut, jika diambil nilai K = 2 terdekat, maka tetangga terdekat dari data yang dimasukkan adalah kelas positif, sehingga kemungkinan rating yang diberikan adalah ≥ 3 bintang.

g. Hasil Algoritma KNN Program

Hasil Algoritma KNN menampilkan hasil klasifikasi dari 1.404 data testing dan menggunakan metode confusion matrix Sebagai metode evaluasi hasil algoritma maka didapatkan data prediksi negative sebanyak 64, prediksi positif sebganyak 6, aktual negative sebanyak 0 dan aktual positif sebanyak 1334 maka hasil analisis dari algoritma KNN dengan menggunakan jumlah tetangga terdekat sebanyak 3 didapat hasil akurasi sebesar 95,01%. Dengan menerapkan jumlah data uji lebih banyak sehingga hasil yang diperoleh untuk prediksi klasifikasi lebih akurat, sehingga tingkat akurasinya dapat meningkat.



Gambar 9 Halaman Hasil Algoritma KNN

Tabel 22 Confusion Matrix KNN

x	Prediksi Negatif	Prediksi Positif
Aktual Negatif	TN = 0	FP = 6
Aktual Positif	FN = 64	TP = 1334

Akurasi – merupakan proporsi jumlah total prediksi yang benar dengan persamaan :

$$Akurasi = \frac{(1334 + 0)}{(0 + 64 + 6 + 1334)} * 100\% = 95.01\%$$

Precision – Proporsi kumppulan data relevan yang diprediksi benar dengan persamaan:

$$Precision = \frac{1334}{(6 + 1334)} * 100\% = 99\%$$

Recall – Proporsi kumpulan data relevan yang diidentifikasi dengan benar, melalui persamaan :

$$Recall = \frac{1334}{(64 + 1334)} * 100\% = 95\%$$

Setelah mendapatkan nilai dari Precision dan Recall maka didapat F – Measure yang digunakan untuk mengoptimalkan nilai nilai keduanya yang terdapat distribusi kelas dengan nilai tidak rata[11], dapat diketahu melalui persamaan :

$$F1 - Measure = 2 * \frac{0.99 * 0.95}{0.99 + 0.95} * 100\% = 96\%$$

4.2 Analisis Hasil Pengujian

Berdasarkan hasil pengujian menggunakan 4.677 dataset yang terdiri dari 3.273 data training, 1.404 data testing dan menggunakan metode confusion matrix Sebagai metode evaluasi hasil algoritma maka didapatkan data prediksi negative sebanyak 64, prediksi positif sebganyak 0, aktual negative sebanyak 0 dan aktual positif sebanyak 1340 maka hasil akurasi dari algoritma Naïve Bayes sebesar 95,44%. Sedangkan hasil algoritma SVM didapatkan data prediksi negative sebanyak 64, prediksi positif sebganyak 1, aktual negative sebanyak 0 dan aktual positif sebanyak 1339 maka hasil akurasi dari algoritma SVM sebesar 95,37%. Sedangkan hasil algoritma KNN didapatkan data prediksi negative sebanyak 64, prediksi positif sebganyak 6, aktual negative sebanyak 0 dan aktual positif sebanyak 1334 maka hasil analisis dari algoritma KNN dengan menggunakan jumlah tetangga terdekat sebanyak 3 didapat hasil akurasi sebesar 95,01%.

Tabel 23 Confusion Matrix Hasil Pengujian Keseluruhan

Model	Akurasi	Presisi	Recall	F1-Score
Naïve Bayes	95,44%	1%	0,95%	0,97%
SVM	95,37%	0,99%	0,95%	0,96%
KNN	95,01%	0,99%	0,95%	0,96%

5. Kesimpulan

Kesimpulan dari penelitain ini menghasilkan pengujian yang dilakukan untuk mengetahui hasil dari pengujian yang telah dilakukan dimana algoritma Naïve Bayes memiliki hasil akurasi yang sedikit lebih unggul dibandingkan dengan algoritma lain walaupun beda tipis, karena memproses data ulasan yang terstruktur (setelah *pre-processing*). Hasil penelitian menunjukkan nilai Akurasi K-Nearest Neighbor memiliki tingkat akurasi 95.01%, *Support Vector Machines* memiliki tingkat akurasi 95.37%, dan *Naïve Bayes* memiliki tingkat akurasi sebesar 95,44%. Berdasarkan penelitian metode *Naïve Bayes* memiliki hasil akurasi yang paling tinggi dibandingkan K-*Nearest Neighbor* dan *Support Vector Machines*, sehingga metode *Naïve Bayes* menjadi metode terbaik dalam prediksi rating store.

Daftar Pustaka

[1] S. R. D. Setiawan, "E-commerce Apa yang Pimpin Pasar Indonesia?," 2019. [Online]. Available: https://money.kompas.com/read/2019/08/26/122218226/e-commerce-apa-yang-pimpin-pasar-indonesia.

[2] and T. S. A. Spink, B. J. Jansen, D. Wolfram, "From E-sex to e-commerce: Web search changes," *Computer (Long. Beach. Calif).*, 2002.

[3] J. S. and H. Sidgwick, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Trans*, 2005.

[4] and N. A. A. I. R. S. Servanda, R. K. S. Putri, ""Peran Ulasan Produk dan Fot Produk yang Ditampilkan Penjual pada Marketplace Shopee terhadap Minat Beli Pria dan Wanita," *J. Manaj. dan Bisnis*, p. J. Manaj. dan Bisnis, 2019.

[5] and I. S. M. Firdaus, F. Rizki, F. Gaus, "Analisis Sentimen Dan Topic Modelling Dalam Aplikasi Ruangguru," *J-SAKTI (Jurnal Sains Komput. dan Inform)*, 2020.

[6] E. Indrayuni, "Analisa Sentimen Review Hotel Menggunakan Algoritma Support Vector Machine Berbasis Particle Swarm Optimization," *EVOLUSI J. Sains dan Manaj.*, vol. 4, no. 2, 2016.

[7] M. Youllia, I. N., Hermana , A. N., & Kharisma, "Penerapan Algoritma K-Nearest Neighbor pada Game Pesawat untuk Pembelajaran Matematika Dasar," 2019.

[8] Y. I. Kurniawan, "PERBANDINGAN ALGORITMA NAIVE BAYES DAN C.45 DALAM KLASIFIKASI DATA MINING," 2018.

[9] G. I. L. P. WANGI, "Opinion Mining Dengan Menggunakan Kombinasi Metode Lexicon-Based dan Multinomial Naive Bayes," 2017.

[10] M. F. Irwansyah, "Advanced Clustering Teori dan Aplikasi," 2015.

[11] M. Pradana, "KLASIFIKASI BISNIS E-COMMERCE DI INDONESIA," *MODUS*, pp. 163–174, 2015.

[12] N. M. S. W. T. N. K. Y. Utami, "KAJIAN USABILITY E-MARKETPLACE BLUPRIN SEBAGAI DIREKTORI BIDANG ARSITEKTUR DAN DESAIN INTERIOR DALAM DUNIA DIGITAL," *Pros. Semin. Nas. Desain dan Arsit.*, 2019.

[13] Y. Paragian, "Berusia lima tahun, Tokopedia kirimkan dua juta produk tiap bulannya," 2014. [Online]. Available: https://id.techinasia.com/toko-online-tokopedia-kirim-dua-juta-barang-per-bulan.

[14] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, pp. 1–5, 2019, doi: 10.1109/ICIC47613.2019.8985884.

[15] C. A. Rahardja, T. Juardi, and H. Agung, "Implementasi Algoritma K-Nearest Neighbor Pada Website Rekomendasi Laptop," *J. Buana Inform.*, vol. 10, no. 1, p. 75, 2019, doi: 10.24002/jbi.v10i1.1847.

[16] I. N. Atthalla, A. Jovandy, and H. Habibie, "Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K-Nearest Neighbor," *Pros. Annu. Res. Semin.*, vol. 4, no. 1, pp. 978–979, 2018.

[17] F. Liantoni and H. Nugroho, "Klasifikasi Daun Herbal Menggunakan Metode Naïve Bayes Classifier Dan Knearest Neighbor," *J. Simantec*, vol. 5, no. 1, pp. 9–16, 2015.

[18] R. Samsir, S., Ambiyar, A., Verawardina, U., Edi, F., & Watrianthos, "Analisis Sentimen Pembelajaran Daring Pada Twitter di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes," *J. MEDIA Inform. BUDIDARMA*, vol. 5(1), pp. 157–163, 2021.

[19] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 113, 2018, doi: 10.26418/jp.v4i2.27526.

[20] J. Han, *Data Mining Concept and Techniques*. Elsivier, 2013.

[21] A. Salam, J. Zeniarja, and R. S. U. Khasanah, "Analisis Sentimen Data Komentar Sosial Media Facebook Dengan K-Nearest Neighbor (Studi Kasus Pada Akun Jasa Ekspedisi Barang J&T Ekpress Indonesia)," *Pros. SINTAK*, pp. 480–486, 2018.

[22] W. T. and L. A. Edison, "Tokopedia," *PT. Tokopedia*, 2009. [Online]. Available: https://www.tokopedia.co.id.