Klasifikasi *Review Customer* Di *E-Commerce* Bukalapak Menggunakan Metode Support Vector Machine (SVM)

1st Ivania Nonita Chrisdiyanti
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
ivania@student.telkomuniversity.ac.id

2nd Riska Yanu Fa'rifah
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
riskayanu@telkomuniversity.ac.id

3rd Oktariani Nurul Pratiwi
Fakultas Rekayasa Industri
Universitas Telkom
Bandung, Indonesia
onurulp@telkomuniversity.ac.id

Abstrak—Bukalapak menempati urutan ketiga dalam top 10 e-commerce Indonesia, tujuan pemeringkatan tersebut yaitu agar pihak Bukalapak dapat meningkatkan kualitas dan kuantitas layanannya. Mengklasifikasi review dari customer Bukalapak yang terlalu banyak membutuhkan waktu yang lama jika dilakukan dengan cara manual. Dibutuhkan suatu metode yang dapat mengklasifikasikan customer review. Metode yang digunakan untuk mengkasifikasikan review adalah Support Vector Machine. Review akan diklasifikasi menjadi dua jenis yaitu positif dan negatif review. Tahapan untuk melakukan klasifikasi pada penelitian ini adalah preprocessing data, ekstraksi fitur dengan TF-IDF, analisis SVM, dan evaluasi.Terdapat 3 skenario yang digunakan dalam penelitian ini, yaitu perbandingan 60:40, 70:30, dan 80:20. Hasil klasifikasi dengan SVM dan fungsi kernel linier pada data training menunjukkan bahwa ketiga rasio mempunyai akurasi dari model terbaik yang dibentuk oleh SVM adalah rasio 60:40. Evaluasi dari model terbaik dari SVM didapatkan akurasi sebesar 85%, Recall sebesar 79%, Precision 89%, dan F1-Score sebesar 84%. Hasil dari K-Fold Cross Validation dengan 10 Fold menunjukkan hasil yang tidak jauh berbeda dari evaluasi yaitu rata-rata sebesar 84%. Hasil klasifikasi kategori positif dapat dijadikan acuan untuk mempertahankan kualitas layanan dan hasil klasifikasi negatif dapat digunakan sebagai bahan evaluasi dalam meningkatkan layanan di Bukalapak.

Kata kunci— customer review klasifikasi, SVM, kernel linear

I. PENDAHULUAN

A. Latar Belakang

Teknologi saat ini berkembang pesat ke arah digital. Kemajuan teknologi, komputer dan telekomunikasi telah mendukung perkembangan teknologi internet. Menurut survei yang dilakukan oleh APJJI (Asosiasi Penyelenggara Jasa Internet Indonesia) jumlah pengguna internet di Indonesia mencapai 196,7 juta dari tahun 2019 hingga kuartal kedua 2020. Jumlah ini setara dengan 73,7% dari jumlah penduduk Indonesia sebanyak 266,91 juta jiwa. Di masa pandemi, transaksi belanja online meningkat pesat. Berdasarkan data Bank Indonesia (BI), jumlah transaksi *e-commerce* meningkat hampir dua kali lipat di masa pandemi Covid-19. Dari 80 juta transaksi di 2019 menjadi 140 juta transaksi di tahun 2020 [1].

Didirikan pada tahun 2010, Bukalapak secara signifikan mentransformasi cara masyarakat Indonesia dalam menjalani aktivitas ekonomi. Bukalapak merupakan startup e-commerce kedua di Indonesia yang meraih predikat sebagai unicorn. Dilansir dari CBInsights, valuasi Bukalapak bahkan mencapai US$ 2,5 miliar atau setara Rp 35 triliun [2]. Bukalapak juga menempati urutan ketiga dalam top 10 e-commerce Indonesia yang dikeluarkan oleh Iprice Insight.

Dengan adanya pemeringkatan tersebut, pihak Bukalapak dapat meningkatkan kualitas dan kuantitas layanannya dengan cara mengetahui hasil customer review terhadap Bukalapak. Sehingga dengan adanya customer review tersebut, pihak Bukalapak dapat melakukan evaluasi terhadap layanan yang diberikan kepada customer sehingga nantinya layanan yang diberikan kepada customer akan semakin baik dan dapat meningkatkan pemeringkatan.

Permasalahan yang dihadapi adalah review dari customer Bukalapak yang terlalu banyak sehingga sulit dan membutuhkan waktu yang lama dalam mengklasifikasi dan menganalisis customer review jika klasifikasi tersebut dilakukan dengan cara manual. Oleh karena itu dibutuhkan suatu metode yang dapat mengolah data review tersebut dengan cara cepat untuk mengklasifikasikan customer review. Metode yang digunakan untuk mengkasifikasikan review tersebut adalah analisis klasifikasi. Review tersebut nantinya akan diklasifikasi menjadi dua jenis yaitu positive review dan negative review. Sehingga nantinya review dari customer tersebut dapat dimanfaatkan sebagai pertimbangan pengguna dalam menggunakan layanan di Bukalapak, ataupun pihak Bukalapak untuk evaluasi peningkatan kualitas dan kuantitas.

II. KAJIAN TEORI

A. *E-Commerce*

Perdagangan Elektronik atau e-commerce adalah hasil teknologi informasi yang saat ini sedang berkembang dengan begitu cepat terhadap pertukaran barang, jasa dan informasi melalui sistem elektronik seperti internet, televisi dan jaringan computer lainnya [3]. Munculnya e-commerce tidak terlepas dari perkembangan teknologi informasi yang begitu pesat, khususnya internet. e-commerce memungkinkan suatu perusahaan menjangkau seluruh dunia untuk memasarkan produk atau jasanya tanpa harus dibatasi oleh batas-batas geografis. e-commerce merupakan pemicu terbentuknya prinsip ekonomi baru yang kini dikenal dengan ekonomi

digital. e-commerce hadir dalam menjawab tuntutan gaya hidup modern manusia yang menuntut kemudahan dan kecepatan dalam segala bidang. Ada 5 (lima) konsep dasar yang dimiliki e-commerce yaitu:
1.Automation, otomasi bisnis proses sebagai pengganti proses manual (konsep ”enterprise resource planning”)
2.Streamlining/Integration, proses yang terintegrasi untuk mencapai hasil yang efisien dan efektif (konsep “just in time”)
3.Publishing, kemudahan berkomunikasi dan berpromosi untuk produk dan jasa yang diperdagangkan (konsep “electronic cataloging”)
4.Interaction, pertukaran informasi/data antar pelaku bisnis dengan meminimalisasikan human error (konsep “electronic data interchange”)
5.Transaction, kesepakatan dua pelaku bisnis untuk bertransaksi dengan melibatkan institusi lain sebagai fungsi pembayar (konsep “electronic payment”)

B. Text Mining

Text mining adalah seni dan ilmu untuk menemukan pengetahuan, wawasan, dan pola dari kumpulan database tekstual yang terorganisir. Penambangan tekstual dapat membantu menganalisis istilah penting dan frekuensi hubungan sematiknya. Teks merupakan bagian penting dari pertumbuhan data di dunia. Teknologi media sosial telah memungkinkan pengguna untuk menjadi produsen teks dan gambar dan jenis informasi lainnya. Penambangan teks dapat diterapkan pada data media sosial yang berskala besar untuk mengumpulkan preferensi dan mengukur sentiment emosional. Ini dapat diterapkan pada skala sosial, organisasi, dan individu [4].

Kumpulan fitur (dimensi) teks juga disebut sebagai leksikon. Kumpulan dokumen disebut sebagai korpus. Dokumen dapat dilihat sebagai urutan, atau rekaman multidimensi. Dokumen teks adalah urutan kata-kata yang berbeda, juga disebut sebagai string. Oleh karena itu, banyak metode penambangan sekuens yang secara teoritis berlaku untuk teks. Namun, metode penambangan sekuens seperti itu jarang digunakan dalam domain teks. Hal ini karena metode penambangan urutan paling efektif ketika panjang urutan dan jumlah token yang memungkinkan keduanya relatif sederhana. Di sisi lain, dokumen seringkali bisa berupa urutan panjang yang digambar pada leksikon yang terdiri dari beberapa ratus ribu kata [5]

C. Klasifikasi

Klasifikasi merupakan cara pengelompokkan benda berdasarkan ciri – ciri yang dimiliki oleh objek klasifikasi. Dalam prosesnya, klasifikasi dapat dilakukan dengan banyak cara baik secara manual ataupun dengan bantuan teknologi. Klasifikasi yang dilakukan secara manual adalah klasifikasi yang dilakukan oleh manusia tanpa adanya bantuan dari algortima cerdas komputer. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi, memiliki beberapa algoritma, diantaranya Naïve Bayes, Support Vector Machine, Decission Tree, Fuzzy dan Jaringan Saraf Tiruan [4]

D. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi. SVM berada dalam satu kelas dengan ANN dalam hal fungsi dan kondisi permasalahan yang bisa diselesaikan. Keduanya masuk dalam kelas *supervised learning*, dimana dalam implementasinya perlu adanya tahap *training* dan disusul tahap *testing*. Baik para ilmuwan maupun praktisi telah banyak menerapkan teknik ini dalam menyelesaikan masalah-masalah nyata dalam kehidupan sehari-hari. Terbukti dalam banyak implementasi, SVM memberi hasil yang lebih baik dari ANN, terutama dalam hal solusi yang dicapai. ANN menemukan solusi berupa lokal optimal, sedangkan SVM menemukan solusi yang global optimal. Tidak heran bila menjalankan ANN, solusi dari setiap *training* hampir selalu berbeda. Hal ini disebabkan solusi lokal optimal yang dicapai tidak selalu sama. SVM selalu mencapai solusi yang sama untuk setiap *running*. Dalam teknik ini, berusaha untuk menemukan fungsi pemisah (*klasifier*) yang optimal yang bisa memisahkan dua set data dari dua kelas yang berbeda Vapink (1995). Teknik ini menarik orang dalam bidang data mining maupun *machine learning* karena performansinya yang meyakinkan dalam memprediksi kelas suatu data baru.

Konsep Klasifikasi dengan Support Vector Machine (SVM) adalah mencari hyperplane terbaik yang berfungsi sebagai pemisah dua kelas data. Ide sederhana dari SVM adalah memaksimalkan margin, yang merupakan jarak pemisah antara kelas data. SVM mampu bekerja pada dataset yang berdimensi tinggi dengan menggunakan karnel trik. SVM hanya menggunakan beberapa titik data terpilih yang berkontribusi (Support Vector) untuk membentuk model yang akan digunakan dalam proses klasifikasi [6]

E. Karakteristik SVM

SVM memerlukan proses pelatihan dengan menyimpan hasil support vector yang didapatkan untuk digunakan kembali pada saat proses prediksi/testing. SVM selalu memberikan model yang sama dan solusi yang sama dengan margin maksimal.
SVM dapat memisahkan data yang distribusi kelasnya bersifat linier maupun non linier. SVM tidak dipengaruhi oleh dimensi data yang tinggi, sehingga tidak ada proses reduksi dimensi didalamnya, memori yang digunakan dalam SVM dipengaruhi oleh banyaknya data, bukan besarnya dimensi data [6].

F. Kelebihan SVM

Kelebihan SVM antara lain sebagai berikut :
1. Generalisasi
Generalisasi didefinisikan sebagai kemampuan suatu metode untuk mengklasifikasikan suatu pattern, yang tidak termasuk data yang dipakai dalam fase pembelajaran metode itu. Vapink menjelaskan bahwa generalization error dipengaruhi oleh dua faktor: error terhadap training set, dan satu faktor lagi yaitu dipengaruhi oleh dimensi VC (Vapink-Chervokinensis). Strategi pembelajaran pada neural network dan umumnya metode learning machine difokuskan pada usaha untuk meminimalkan error pada training set. Strategi ini disebut Empirical Risk Minimization (ERM). Adapun SVM selain meminimalkan error pada training set, juga

meminimalkan faktor kedua. Strategi ini disebut Structural Risk Minimization (SRM), dan dalam SVM diwujudkan dengan memilih hyperplane dengan margin terbesar. Berbagai studi empiris menunjukkan bahwa pendekatan SRM dan SVM memberikan error generalisasi yang lebih kecil daripada yang diperoleh dari strategi ERM pada neural network maupun metode lain [6].

2. Curse of Dimensionality

Curse of dimensionality didefinisikan sebagai masalah yang dihadapi suatu metode pattern recognition dalam mengestimasikan parameter (misalnya jumlah hidden neuron pada neural network, stopping criteria dalam proses pembelajaran) dikarenakan jumlah sampel data yang relatif sedikit dibandingkan dimensional ruang vektor data tersebut. Semakin tinggi dimensi dari ruang vector informasi yang diolah, membawa konsekuensi dibutuhkannya jumlah data dalam proses pembelajaran. Seringkali terjadi data yang diolah berjumlah terbatas, dan untuk mengumpulkan data yang lebih banyak tidak mungkin dilakukan karena kendala biaya dan kesulitan teknis. Dalam konsisi tersebut, jika metode itu "terpaksa" harus bekerja pada data yang berjumlah relative sedikit dibandingkan dimensinya, akan membuat proses estimasi parameter metode menjadi sangat sulit [6].

3. Feasibility

SVM dapat diimplementasikan relative mudah, karena proses penentuan support vector dapat dirumuskan dalam QP problem. Dengan demikian jika memiliki library untuk menyelesaikan QP problem, dengan sendirinya SVM dapat diimplementasikan dengan mudah. Selain itu dapat diselesaikan dengan metode sekuensial sebagaimana penjelasasn sebelumnya [6]

G. SMOTE

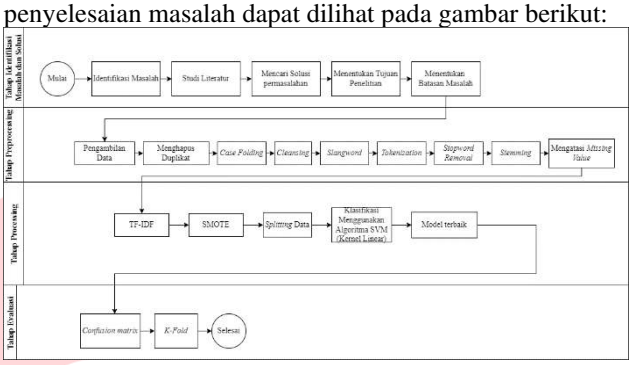
Teknik SMOTE berguna untuk menghasilkan data yang lebih baik dan efektif untuk menangani ketidakseimbangan kelas yang mengalami over-fitting pada proses teknik over-sampling untuk kelas minoritas (positif). SMOTE menciptakan sebuah contoh dari kelas minoritas sintesis yang beroperasi di ruang fitur daripada ruang data. Dengan menduplikasi contoh kelas minoritas, teknik SMOTE menghasilkan contoh sintetis baru dengan melakukan ekstrapolasi sampel minoritas yang ada dengan sampel acak yang diperoleh dari nilai k tetangga terdekat. Dengan hasil sintetis pada contoh yang lebih lebih dari kelompok minoritas, sehingga mampu memperluas area keputusan mereka untuk minoritas. Prinsip dari metode SMOTE adalah dengan menambah jumlah data kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data buatan. Data buatan atau sintesis tersebut dibuat berdasarkan k-tetangga terdekat (k-nearest neighbor). Jumlah k-tetangga terdekat ditentukan dengan mempertimbangkan kemudahan dalam melaksanakannya. Pembangkitan data buatan yang berskala numerik berbeda dengan kategorik. Data numerik diukur jarak kedekatannya dengan jarak Euclidean sedangkan data kategorik lebih sederhana yaitu dengan nilai modus [7].

H. K-Fold Cross Validation

Cross validation merupakan metode statistik mengevaluasi dengan membagi data menjadi dua segmen: satu digunakan untuk training dan yang lain digunakan untuk memvalidasi model/testing. Dalam Cross validasi, training set dan testing set diatur sedemikian rupa sehingga setiap data pernah menjadi training set dan testing set. Misalnya: 5-fold validation berarti bahwa record dataset dibagi 4 subset menjadi training set dan 1 subset sebagai testing set [8].

III. METODE

Proses penyelesaian masalalah dibagi ke dalam 4 tahap, yaitu tahap identifikasi masalah dan solusi, tahap preprocessing, tahap processing, dan tahap evaluasi. Sistematika penyelesaian masalah dapat dilihat pada gambar berikut:



GAMBAR III.1
SISTEMATIKA PENYELESAIAN MASALAH

1. Tahap identifikasi masalah dan solusi

Tahap pertama yang dilakukan adalah mengidentifikasi masalah dengan melihat customer review Bukalapak kemudian melakukan studi literatur dengan mencari jurnal maupun referensi yang berhubungan dengan studi kasus sebagai acuan untuk memberikan solusi yang tepat dari permasalahan tersebut. Setelah menemukan solusi dari permasalahan yang terjadi, tahap selanjutnya yaitu menentukan tujuan penelitian dan menentukan Batasan masalah dari penelitian untuk membatasi ruang lingkup permasalahan yang ada.

2. Tahap preprocessing

Setelah melakukan tahap identifikasi masalah dan solusi, tahap yang harus dilakukan selanjutnya yaitu tahap preprocessing. Tujuan dilakukan preprocessing adalah mempersiapkan data yang tidak terstruktur menjadi data terstruktur yang siap digunakan untuk proses selanjutnya dengan cara menghilangkan noise, menyeragamkan bentuk kata dan mengurangi volume kata [9]. Tahap preprocessing dimulai dari pengambilan data. Setelah data berhasil didapatkan, kemudian dilakukan beberapa proses sebagai berikut:

a. Menghapus Duplikat

Menghapus duplikat adalah proses dimana data yang sama akan dihapus

b. Case Folding

Case Folding adalah tahapan untuk merubah bentuk kata-kata menjadi bentuk yang sama, baik menjadi huruf kecil semua atau huruf besar semua [10]

c. Data Cleansing

Cleansing merupakan proses membersihkan istilah-istilah yang tidak diperlukan untuk mengurangi noise. Kata yang dihilangkan seperti titik(.), Koma(,), seru(!) [10]).

d. Slangword

Slangword merupakan proses mengubah kata tidak baku menjadi kata baku. Tahap ini dilakukan dengan menggunakan bantuan kamus slangword dan padanannya dalam kata-kata baku. Tahapan ini akan memeriksa kata yang

terdapat dalam kamus slangword atau tidak. Jika kata tidak baku terdapat dalam kamus slangword maka kata tidak baku akan diubah ke kata baku yang terdapat didalam kamus slangword [11].

d.Tokenization

Tahap berikutnya adalah tokenisasi yaitu memotong dokumen menjadi kata-kata yang berdiri sendiri contoh kalimat barang sesuai dengan gambar jika dilakukan tokenisasi menghasilkan 4 token yaitu [‘barang’,’sesuai’,’ dengan’,’gambar’] [12].

e. Stopword Removal

Tahap Stopword Removal adalah tahap dimana kata-kata penting diekstraksi dari hasil token. Bisa menggunakan algoritma stoplist (menghapus kata-kata yang kurang penting) atau wordlist (menyimpan kata-kata penting). Stoplist / stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang. Contoh stopword adalah “yang”,”dan”,”di” dan lain-lain [13].

f. Stemming

Stemming adalah tahapan untuk mengembalikan kata yang berimbuhan kembali ke bentuk asalnya. Contoh kata membeli setelah melewati tahap ini maka akan menjadi “beli” [10].

3. Tahap Processing

Setelah melakukan tahap preprocessing, tahap yang dilakukan selanjutnya yaitu tahap processing. Pada tahap ini, dilakukan penanganan missing value, kemudian dilakukan proses pemberian nilai terhadap setiap term menggunakan metode TF-IDF. TF-IDF (Term Frequency-Inverse Document Frequency) adalah sebuah metode pembobotan yang menggabungkan dua konsep, yaitu term frequency dan document frequency. Term frequency adalah konsep pembobotan dengan mencari seberapa sering (frekuensi) munculnya sebuah term dalam satu dokumen sedangkan Document Frequency adalah banyaknya jumlah dokumen di mana sebuah term itu muncul [14]. Selanjutnya dilakukan balancing data dengan menggunakan teknik SMOTE Pada tahap ini, data dibagi secara seimbang (balanced) di setiap kelas, hal ini karena jika data tidak seimbang (imbalanced), klasifikasi yang dihasilkan cenderung mengabaikan kelas minoritas [15]. Setelah dilakukan balancing data, kemudian dilakukan Splitting data dengan membagi data menjadi dua bagian yaitu training dan testing. Data training adalah data ysng digunakan untuk membangun sebuah model sedangkan data testing adalah data yang digunakan untuk pengujian model yang telah dibuat dengan data lainnya untuk mengetahui akurasi dari model tersebut [16] Setelah itu, dilakukan klasifikasi menggunakan algoritma SVM

4. Tahap Evaluasi

Tahap akhir adalah tahap evaluasi yaitu melakukan evaluasi pada model terbaik dengan menggunakan confusion matrix dan k-fold

IV. HASIL DAN PEMBAHASAN

A. Pengumpulan Data

Data diambil dari ulasan (review) customer Bukalapak. Data diperoleh dari kaggle dan dikumpulkan oleh Christofel dengan jumlah 96328. Data tersebut merupakan kumpulan komentar customer review Bukalapak. Data ini mempunyai dua kelas, yaitu kelas 1 (positive) dan kelas 0 (negative). Data yang diperoleh dalam format CSV kemudian akan diolah di tahap preprocessing. Berikut merupakan contoh dari data customer bukalapak yang belum diolah:

TABEL IV.1
STRUKTUR *DATASET*

id	Header_review	Review_sangat_singkat	Label
0	Barang Sesuai	Terima kasih bukalapak barang sesuai dengan keinginan dan memuaskan	1
1	Barang Sesuai Pesanan	Tks gan barang sesuai pesanan, cepat sampai, admin toko fast respon. mantep pokoknya. Sukses gan	1
2	Dvd Rusak	Pas saya play untuk pertama kali gak ada masalah. Tapi pas kedua kalinya malah gak bisa sama sekali. Kecewa saya	0
3	Kecewa	Tidak sesuai dengan pesanan, memesan dengan warna apa di kirimnya apa.	0

B. Data *Preprocessing*

Preprocessing dilakukan untuk memperoleh format data yang sesuai untuk tahap klasifikasi sebelum dilakukan pembobotan dan klasifikasi menggunakan TF-IDF. Tahap preprocessing ini terdiri dari case folding, cleaning, slangword, tokenization, stopwords removal, dan stemming. Berikut merupakan contoh *preprocessing* data:

TABEL IV.2
HASIL *PREPROCESSING*

Data	Barang sesuai dengan gambar dan kualitas oke.... respon cepat...
Case folding	barang sesuai dengan gambar dan kualitas oke.... respon cepat...
Cleaning	barang sesuai dengan gambar dan kualitas oke respon cepat
Slangword	barang sesuai dengan gambar dan kualitas oke respon cepat
Tokenization	['barang', 'sesuai', 'dengan', 'gambar', 'dan', 'kualitas', 'oke', 'respon', 'cepat']
Stopwords removal	['barang','sesuai', 'gambar', 'kualitas', 'oke', 'respon', 'cepat']
Stemming	barang sesuai gambar kualitas oke respon cepat

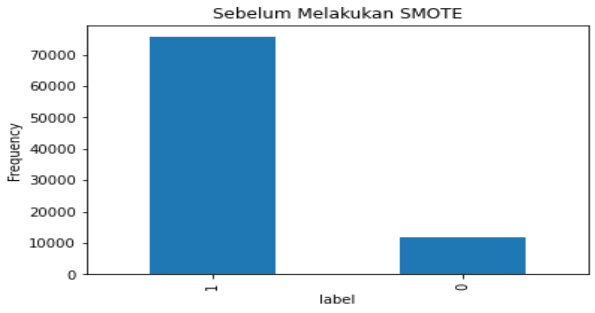
C. Mengatasi *Missing value*

Missing value adalah keadaan dimana beberapa nilai atribut dalam dataset kosong (tidak ada nilainya). Pada customer review Bukalapak terdapat missing value sebanyak 418 sehingga missing value tersebut harus diatasi terlebih dahulu sebelum ke tahap selanjutnya. Metode yang paling umum digunakan adalah mengganti missing value dengan nilai kecenderungan pusat atributnya, yaitu mengganti dengan nilai modus untuk tipe data atribut kategorikal

D. *Balancing* Data Menggunakan SMOTE

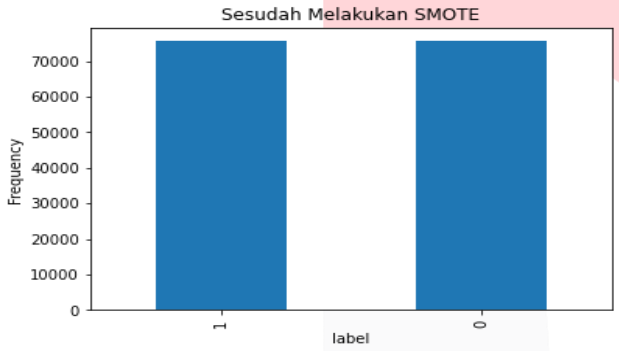
Pada penelitian ini, dataset yang digunakan merupakan dataset yang memiliki imbalance class. Sehingga pada penelitian ini menggunakan metode Syntetic Minority Over Sampling Technique (SMOTE) untuk mengatasi permasalahan imbalance class tersebut dan untuk memaksimalkan kinerja dari algroritma SVM. Metode ini

merupakan pendekatan yang bekerja dengan membuat replikasi data dari data minoritas [17].



GAMBAR IV.1
VISUALISASI DATA SEBELUM SMOTE

Dapat dilihat pada gambar di atas terdapat sebanyak 87241 data dengan 75585 data berkomentar positif dan 11656 data berkomentar negative. Dikarenakan terdapat jumlah data yang tidak seimbang maka dilakukan balancing data menggunakan metode SMOTE



GAMBAR IV.2
VISUALISASI DATA SESUDAH SMOTE

Gambar IV.2 merupakan data yang sudah dilakukan balancing dengan menggunakan metode SMOTE terdapat sebanyak 151170 data dengan 75585 data berkomentar positif dan 75585 data berkomentar negative.

E. Pembobotan TF-IDF

Pada tahap ini dilakukan pembobotan fitur kata pada customer review bukalapak. TF-IDF merupakan gabungan dari term frequency (TF) dan invers document frequency (IDF) yang digunakan dalam menghitung bobot setiap kata (term) pada setiap dokumen. Berikut merupakan contoh dari hasil TF-IDF.

TABEL IV.3
HASIL TF-IDF

Stemming	TF-IDF
'mantap' 'barang' 'sesuai'	'mantap': 0.657459 'barang': 0.528694 'sesuai': 0.391737

F. *Splitting* Data

Pada tahap ini, data dibagi kedalam dua proses yaitu proses training dan proses testing. Proses pertama yang dilakukan adalah proses training untuk pelatihan, kemudian proses testing untuk pengujian. Data tersebut akan dibagi berdasarkan tiga macam pembagian rasio dengan menggunakan metode Splitting data yaitu dengan rasio 60:40, 70:30 dan 80:20 sebelum dilakukan SMOTE dan sesudah dilakukan SMOTE. Kemudian data diuji akurasinya dengan ketiga rasio tersebut. Pembagian rasio dengan tingkat akurasi paling tinggi yang digunakan untuk penelitian ini. Adapun banyaknya data yang akan terbagi dapat dilihat pada tabel

TABEL IV.4
SPLITTING DATA SEBELUM SMOTE

Rasio	Proses		Total Data
	Training	*Testing*	
60:40	52344	34897	87241
70:30	61068	26173	
80:20	69792	17449	

Berdasarkan tabel diatas, setelah dilakukan Splitting data sebelum menggunakan metode smote dengan rasio 60:40 didapatkan data training sebanyak 52344 data dan data testing sebanyak 34897, Splitting data dengan rasio 70:30 didapatkan data training sebanyak 61068 dan data testing sebanyak 26173, dan Splitting data dengan rasio 80:20 didapatkan data training sebanyak 69792 dan data testing sebanyak 17449

TABEL IV.5
SPLITTING DATA SESUDAH SMOTE

Rasio	Proses		Total Data
	Training	*Testing*	
60:40	90702	60468	151170
70:30	105819	45351	
80:20	120936	30234	

Berdasarkan tabel di atas, setelah dilakukan Splitting data sesudah menggunakan metode SMOTE dengan rasio 60:40 didapatkan data training sebanyak 90702 data dan data testing sebanyak 60468, Splitting data dengan rasio 70:30 didapatkan data training sebanyak 105819 dan data testing sebanyak 45351, dan Splitting data dengan rasio 80:20 didapatkan data training sebanyak 120936 dan data testing sebanyak 30234.

Data yang digunakan dari penelitian adalah hasil dari data yang telah melalui proses SMOTE. Setelah dilakukan proses SMOTE, data dibagi menjadi data training dan testing. Data training digunakan untuk memprediksi klasifikasi dengan metode SVM sedangkan data testing digunakan untuk mengevaluasi hasil prediksi pada training

G. Klasifikasi dengan SVM

Implementasi algoritma yang dipilih yaitu algoritma SVM dengan menggunakan kernel linear. Terdapat 3 skenario training yang digunakan dalam penelitian ini, yaitu rasio 60:40, rasio 70:30 dan rasio 80:20

TABEL IV.6
HASIL AKURASI *TRAINING*

Skenario	Akurasi
60:40	89,12%
70:30	89,05%
80:20	89,07%

Setelah dilakukan pengujian pada data *training* dengan perbandingan rasio 60:40, 70:30, dan 80:20, hasil akurasi yang didapatkan untuk rasio 60:40 sebesar 89,12%, untuk rasio 70:30 sebesar 89,05%, dan untuk rasio 80:20 sebesar

89,07%. Berdasarkan hasil akurasi yang telah didapatkan, rasio terbaik terdapat pada rasio 60:40 dengan model hyperplane:

$$Y = 0.404918547879 * x + 1.9865196845991209$$

H. Confusion matrix

Dengan menggunakan confusion matrix, didapatkan hasil *accuracy*, *recall*, *precision*, *F1 Score*

TABEL IV.7
HASIL CONFUSION MATRIX

Confusion Matrix		Predicted		Accuracy	Recall	Precision	F1 Score
		Negatif	Positif				
Actual	Negatif	27247	3019	85%	79%	89%	84 %
	Positif	6349	23853				

Berdasarkan tabel diatas, diperoleh *confusion matrix* dengan nilai *true positive* (TP) sebanyak 23853, *true negative* (TN) sebanyak 27247, *false positive* sebanyak 3019 dan *false negative* (FN) sebanyak 6349 dengan nilai akurasi 85% yang masuk ke dalam kategori *good classification*, nilai recall dari label yang diprediksi positif dari keseluruhan label positif yang benar bernilai positif sebesar 79%, nilai *precision* dari label positif yang benar bernilai positif dari keseluruhan label positif yang diprediksi positif sebesar 89% dan nilai *f1-score* sebesar 84%. Hasil tersebut diperoleh dari data testing.

I. K-Fold Cross Validation

Validasi pada model classifier SVM menggunakan satu model yang telah dibangun, sedangkan k pada metode *k-cross fold validation* menggunakan nilai k=10. Hasil Validasi menunjukkan bahwa akurasi optimal berdasarkan metode *k-Fold Cross Validation* memiliki akurasi lebih sedikit dibandingkan dari akurasi awal, dimana model classifer SVM memiliki rata-rata akurasi sebesar 84% Nilai akurasi metode klasifikasi SVM menggunakan 10-fold cross validation adalah sebagai berikut:

TABEL IV.8
K-FOLD CROSS VALIDATION

Iterasi-ke	Nilai akurasi (dalam %)
1	83%
2	85%
3	84%
4	84%
5	84%
6	84%
7	84%
8	85%
9	84%
10	83%
Rata-Rata	84%

V. KESIMPULAN

A. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, dapat diambil kesimpulan sebagai berikut:

1. Analisis klasifikasi terhadap customer review Bukalapak menggunakan algoritma SVM dilakukan dengan beberapa tahap mulai dari preprocessing, ekstraksi fitur dengan TF-IDF, klasifikasi menggunakan SVM dengan akurasi tertinggi pada proses training yaitu pada split data dengan rasio 60:40 dengan model hyperplane yang didapatkan $0.404918547879 * x + 1.9865196845991209$ dan Evaluasi menggunakan *confusion matrix* dan *k-fold cross validation*
2. Berdasarkan hasil evaluasi yang telah dilakukan berdasarkan proses training dengan akurasi tertinggi yaitu pada rasio 60:40, didapatkan akurasi sebesar 85% yang masuk ke dalam kategori good classification, recall sebesar 79%, precision sebesar 89% dan f1-score sebesar 84%. Hasil dari k-fold cross validation dengan 10 fold menunjukkan hasil yang tidak jauh berbeda dari evaluasi yaitu rata-rata sebesar 84%.
3. Analisis klasifikasi terhadap review customer Bukalapak menghasilkan 33596 klasifikasi positif dan 26872 klasifikasi negatif. klasifikasi positif yang diberikan customer ke bukalapak dapat dipertahankan sedangkan klasifikasi negatif yang diberikan customer Bukalapak perlu diperbaiki sehingga nantinya dapat meningkatkan layanan Bukalapak

B. SARAN

Dari hasil tugas akhir ini terdapat beberapa hal yang dapat menjadi saran serta rekomendasi pada penelitian selanjutnya. Berikut merupakan sarannya:

1. Diharapkan untuk penelitian selanjutnya dapat menambahkan lebih banyak koleksi kamus untuk kata-kata Bahasa Indonesia gaul dan singkatan-singkatan tertentu sehingga mempermudah peneliti dalam melakukan proses klasifikasi karena pada review customer Bukalapak banyak menggunakan Bahasa yang kurang baku
2. Diharapkan peneliti selanjutnya dapat mengklasifikasikan review customer berdasarkan kategori barang

REFERENSI

[1] F. Ariyanti, "Survei APJII: Mayoritas Orang RI Merasa Data Pribadinya di Internet Aman," 19 November 2020.

[2] Y. Astutik, "Bukalapak Raih Penghargaan The Best E-Commerce 2019," 04 Desember 2019.

[3] R. M. D. H. Saputra, D. W. Purba, M. Iswahyudi, A. R. Banjarnahor, A. H. Perdana Kusuma, F. Effendy, O. K. Sulaiman and J. Simarmata, "E-Commerce: Implementasi, Strategi dan Inovasinya," Yayasan Kita Menulis, 2019.

[4] A. P. Wibawa, M. G. Aji Purnama, M. F. Akbar and F. A. Dwiyanto, "Metode-metode Klasifikasi," *Prosiding Seminar Ilmu Komputer dan Teknologi Informasi,* p. 1, 2018.

[5] N. Purwanti, H. Kurniawan and S. Karnila, Data Mining, Banyumas: Zahira Media Publisher, 2021.

[6] F. A. Sianturi, P. M. Hasugian, A. Simangunsong and B. Nadeak, Data Mining : Pengembangan Aplikasi WEKA, Sumatera Utara: IOCS Publisher, 2019.

[7] A. N. Rais and A. Subekti, "Integrasi SMOTE dan Ensamble AdaBoost Untuk Mengatasi Imbalance Class Pada Data Bank Direct Marketing," *Jurnal Informatika,* 2019.

[8] M. Lutfi and M. Hasyim, "PENANGANAN DATA MISSING VALUE PADA KUALITAS PRODUKSI JAGUNG DENGAN MENGGUNAKAN METODE K-NN IMPUTATION PADA ALGORITMA C4.5," *JURNAL RESISTOR,* 2019.

[9] V. I. Santoso, G. Virginia and Y. Lukito, "79JURNAL TRANSFORMATIKA, Volume 14, Nomor 2, Januari2017PENERAPAN SENTIMENT ANALYSIS PADA HASIL EVALUASI DOSENDENGAN METODE SUPPORT VECTOR MACHINE," *Jurnal Transformatika,* 2017.

[10] L. Wilianto, T. H. Pudjiantoro and F. R. Umbara, "ANALISIS SENTIMEN TERHADAP TEMPAT WISATA DARI KOMENTAR PENGUNJUNG," *Prosiding SNATIF,* 2017.

[11] I. Zulfa and E. Winarko, "Sentimen Analisis Tweet Berbahasa Indonesiadengan Deep Belief Network," *Indonesian Journal of Computing and Cybernetics System,* 2017.

[12] A. N. Kasanah, M. and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *Rekayasa Sistem dan Teknologi Informasi,* 2019.

[13] Z. U. Siregar, R. R. A. Siregar and R. Arianto, "KLASIFIKASI SENTIMENT ANALYSIS PADA KOMENTAR PESERTA DIKLAT MENGGUNAKAN METODEK-NEAREST NEIGHBOR," *JURNAL KILAT,* 2019.

[14] N. M. S. Hadna, W. Winarno and P. I. Santosa, "Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen di Twitter," in *Seminar Nasional Teknologi Informasi dan Komunikasi 2016*, Yogyakarta, 2016.

[15] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," *Integer Journal,* 2017.

[16] D. A. Nasution, H. H. Khotimah and N. Chamidah, "PERBANDINGAN NORMALISASI DATA UNTUK KLASIFIKASI WINE MENGGUNAKAN ALGORITMA K-NN," *CESS (Journal of Computer Engineering System and Science),* 2019.

[17] Y. E. Ardiningtyas and P. H. Prima Rosa, "ANALISIS BALACING DATA UNTUK MENINGKATKAN AKURASI DALAM KLASIFIKASI," *SNAST,* 2021.