

## PENERAPAN *TEXT MINING* PADA APLIKASI TOKOPEDIA MENGUNAKAN ALGORITMA *K-NEAREST NEIGHBOR*

<sup>1</sup>M. Afdal, <sup>2</sup>Lufti Rahma Elita

<sup>1,2</sup>Program Studi Sistem Informasi, Fakultas Sains dan Teknologi UIN Suska Riau  
Jl. HR Soebrantas KM.18 Panam Pekanbaru - Riau

Email: <sup>1</sup>m.afdal@uin-suska.ac.id, <sup>2</sup>11753201367@students.uin-suska.ac.id

### ABSTRAK

*Google play* menyediakan fitur kepada pengguna untuk dapat memberikan ulasan terhadap aplikasi yang digunakan, salah satunya yaitu ulasan terhadap aplikasi Tokopedia. Tokopedia merupakan salah satu aplikasi yang menggunakan model bisnis *marketplace* dan *mall online*. Peningkatan jumlah pengunjung dan pengguna baru yang akan mengunduh aplikasi Tokopedia berkaitan dengan komentar terhadap aplikasi Tokopedia. Ulasan aplikasi Tokopedia dibagi menjadi ulasan positif, netral, dan negatif. Guna meningkatkan kepercayaan, pengguna bisa melihat komentar terhadap aplikasi Tokopedia. Oleh karena itu, diperlukan suatu teknik pengolahan data dan analisis terhadap komentar. Sentimen atau komentar tersebut dapat dianalisis dengan menggunakan *text mining*. *Text mining* merupakan teknik dalam pengambilan informasi dari sejumlah data tak terstruktur dari sebuah topik tertentu yang memiliki kualitas tinggi serta dapat diperoleh data-data permasalahan dalam teks. Penelitian ini bertujuan untuk mengklasifikasikan ulasan aplikasi tokopedia menggunakan algoritma *K-Nearest Neighbor*. Algoritma *K-Nearest Neighbor* merupakan salah satu algoritma yang terdapat dalam teknik klasifikasi. Pada penelitian ini, ulasan yang digunakan sebanyak 5.254 data yang dibagi menggunakan *K-Fold Cross Validation* yang kemudian dihitung akurasi. Parameter yang digunakan yaitu K=5, K=10, K=25, K=50, K=100, dan K=200. Pada parameter K=5 diperoleh akurasi sebesar 69,02%, parameter K=10 sebesar 68,48%, parameter K=25 sebesar 71,39%, parameter K=50 sebesar 70,84%, parameter K=100 sebesar 70,03%, dan parameter K=200 diperoleh akurasi sebesar 71,12%. Hasil kemunculan kata yang sering muncul untuk kelas positif yaitu kata “tokopedia”, untuk kelas netral kemunculan kata paling banyak yaitu kata “saya”, sementara untuk kelas negatif, kemunculan kata yang paling banyak yaitu pada kata “saya”.

**Kata Kunci:** Akurasi, *K-Fold Cross Validation*, *K-Nearest Neighbor*, Klasifikasi, *Text Mining*

### A. PENDAHULUAN

Dengan perkembangan teknologi yang semakin canggih membuat masyarakat Indonesia dapat melakukan komunikasi dengan pihak lain tanpa batasan waktu dan jarak [9]. Indonesia mengalami pertumbuhan tahunan dalam *e-commerce*. Pertumbuhan ini tidak terlepas dari besarnya jumlah pengguna internet di Indonesia [8]. Selain itu, Indonesia juga merupakan negara besar yang memiliki jumlah penduduk yang banyak sehingga membuat perkembangan *e-commerce* menjadi semakin meningkat. Berdasarkan survey yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) pada tahun 2020, pengguna internet di Indonesia mencapai angka 196,7 juta orang yang sebelumnya pada tahun 2018 hanya 171,17 juta jiwa.

Salah satu perusahaan yang paling berdampak pada kehadiran internet bagi masyarakat adalah *e-commerce* [19]. Pada saat ini, *layanan e-commerce*

sudah ditawarkan diberbagai macam bentuk aplikasi. Sudah banyak bisnis *e-commerce* yang hadir di Indonesia dan Tokopedia merupakan salah satu diantaranya. Pada tahun 2021, berdasarkan peta *e-commerce* Indonesia, Tokopedia menempati posisi pertama dengan jumlah pengunjung web bulanan terbanyak yaitu sebanyak 147,8 juta pengguna internet, urutan kedua dengan Shopee sebanyak 126,9 juta pengunjung dan Lazada diurutan ketiga sebanyak 29,5 juta pengunjung [11]. Sedangkan untuk pengikut di Twitter, Tokopedia memiliki pengikut sebanyak 657 ribu, sedangkan untuk pengikut Shopee dan BliBli yaitu 514 ribu pengikut dan kemudian diikuti oleh Lazada dan Bukalapak yang masing-masing jumlah pengikutnya 411 ribu dan 199 ribu pengikut. Jika dibandingkan dengan jumlah karyawan, Tokopedia juga memiliki jumlah karyawan terbanyak dengan 4.700 orang yang lebih unggul dibandingkan dengan Bukalapak yaitu 2.651 orang [11].

Peningkatan jumlah pengunjung dan pengguna baru yang akan mengunduh aplikasi Tokopedia berkaitan dengan komentar terhadap aplikasi Tokopedia. Guna meningkatkan kepercayaan, pengguna bisa melihat komentar terhadap aplikasi Tokopedia. Komentar terhadap aplikasi Tokopedia juga mampu meningkatkan kualitas terhadap layanan [5]. Pengguna memberikan pendapat mereka tentang aplikasi Tokopedia yang disediakan *Google Play*. Ulasan yang dikemas menjadi bentuk komentar. Komentar yang ada meliputi komentar positif, netral, dan negatif. Oleh karena itu, diperlukan suatu teknik pengolahan data dan analisis terhadap komentar [21]. Sentimen atau komentar tersebut dapat dianalisis dengan menggunakan *text mining*.

*Text mining* diartikan sebagai suatu teknik untuk pengambilan informasi dari sejumlah data unsupervised dari sebuah topik tertentu yang memiliki kualitas tinggi agar dapat diperoleh data-data permasalahan dalam teks [18]. *Text mining* mampu mengidentifikasi emosional terhadap suatu pernyataan [24].

Penelitian sebelumnya yang dilakukan oleh Assad Hidayat (2019), Klasifikasi Opini Masyarakat terhadap Toko *Online* menggunakan Algoritma *K-Nearest Neighbor* dan *Probabilistic Neural Network* pada *Twitter* dengan hasil, algoritma KNN memiliki nilai akurasi terbaik pada Lazada dan Blibli jika dibandingkan dengan menggunakan algoritma PNN, yaitu dengan hasil Lazada dengan KNN 71,57% dan PNN 66,71 sementara Blibli di KNN 68,29% dan PNN 65,29% [10]. Selanjutnya penelitian yang dilakukan oleh Astia Weni Syaputri (2020), mengenai Analisis Sentimen pada Ulasan Hotel *Grand Elite* di *website* Traveloka menggunakan Algoritma *K-Nearest Neighbor*, diperoleh hasil bahwa akurasi tertinggi pada  $k=50$  pada split  $k-1$  dengan nilai 90,0% [21]. Kemudian penelitian yang dilakukan oleh Putri Anglenia (2020), mengenai Penerapan Algoritma Klasifikasi terhadap Sentimen Masyarakat menggunakan *Text Mining* pada Toko *Online* lazada diperoleh hasil akurasi KNN lebih tinggi dibandingkan dengan NBC dan PNN dengan akurasi KNN sebesar 72,85%, sedangkan NBC sebesar 66,66%, dan PNN sebesar 67,14% [1].

Penelitian ini menggunakan algoritma yang terdapat dalam teknik klasifikasi yaitu *k-Nearest Neighbor* (KNN). KNN merupakan algoritma yang digunakan sebagai pengklasifikasian suatu objek

terhadap nilai  $k$  data latih dengan syarat jaraknya yang terdekat dengan objek, nilai  $k$  harus kecil dari jumlah *training set* dan nilai  $k$  tidak boleh genap dan lebih dari satu [10].

Berdasarkan uraian di atas dan didukung oleh beberapa penelitian sebelumnya, maka pada penelitian ini akan dilakukan penerapan *text mining* terhadap ulasan aplikasi Tokopedia menggunakan Algoritma *K-Nearest Neighbor*.

## B. LANDASAN TEORI

### B.1 E-Commerce

*E-Commerce* atau perdagangan secara elektronik merupakan suatu kegiatan yang meliputi berbagai aktifitas seperti penyebaran, penjualan, pembelian, pemasaran produk yang berupa barang ataupun jasa yang memanfaatkan jaringan telekomunikasi. Atau dengan kata lain, *e-commerce* dapat diartikan sebagai transaksi jual beli secara elektronik [4]. *E-business* mengarah pada definisi *e-commerce* yang lebih luas, tidak hanya ada proses jual beli saja, namun ada layanan pelanggan yang berkolaborasi dengan mitra bisnis lain. *E-Commerce* terbagi menjadi beberapa jenis, diantaranya *Consumer to Consumer (C2C)*, *Business to Business (B2B)*, *Consumer to Business (C2B)*, dan *Business to Consumer (B2C)* [8].

### B.2 Tokopedia

Tokopedia bisa dikatakan sebagai salah satu *mall online* di Indonesia yang menggunakan model bisnis *marketplace* dan *mall online*. Tokopedia menjadi salah satu *online marketplace* dengan tingkat pertumbuhan yang sangat pesat di Indonesia [23]. Tokopedia merupakan *marketplace* karya anak bangsa yang berdiri pada tahun 2009 oleh William Tanuwijaya dan Leontinus Alpha Edison. Awal mulai dirintisnya Tokopedia yaitu pada tanggal 6 Februari 2009 dan resmi dirilis untuk diumumkan pada 17 Agustus 2009 [19].

Tokopedia ialah salah satu *e-commerce* yang menawarkan berbagai produk seperti pakaian pria dan wanita, berbagai peralatan rumah tangga, elektronik hingga ke perlengkapan bayi. Aplikasi Tokopedia dapat memberi wadah ke penjual untuk berjualan dengan aman serta dapat mempermudah dalam proses pembayaran. Saat ini aplikasi Tokopedia tidak hanya tersedia di *website* resminya saja, tetapi juga bisa didownload melalui perangkat mobile dengan sistem operasi *iOS* ataupun *Android*.

### B.3 Data Mining

*Data mining* merupakan suatu proses untuk mendapatkan informasi yang berharga dengan cara mengumpulkan data yang diambil dari data dengan ukuran yang cukup besar [10]. *Data mining* dapat digunakan diberbagai bidang yang mempunyai sejumlah data. *Data mining* berada dalam tahapan proses *Knowledge Discovery in Database* (KDD). *Data mining* ialah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* [13].

### B.4 Text Mining

*Text mining* ialah metode yang dapat digunakan untuk menangani masalah klasifikasi, *clustering*, *information extraction*, dan *information retrieval* [3]. *Text mining* merupakan tahapan dalam menemukan informasi agar pengguna dapat berinteraksi dengan sekumpulan dokumen dari waktu ke waktu dengan menggunakan seperangkat alat analisis [7]. Proses utama dari *text mining* ini yaitu untuk mendapatkan kata yang dapat menggantikan isi dari dokumen dan kemudian akan dilakukan proses pengkajian untuk melihat keterkaitan tiap dokumen.

### B.5 Text Preprocessing

*Preprocessing* merupakan proses pertama dari *text mining* dan mempersiapkan teks pemrosesan data lebih lanjut. Tahapan *preprocessing* dalam penelitian meliputi:

1. *Tokenizing*  
Operasi pemisahan *string* input berdasarkan tiap kata yang menyusunnya [17].
2. *Filtering*  
Mengumpulkan kata penting yang telah diurai tiap kata dari sebuah kali- mat. Pada tahap ini dilakukan pembersihan kata-kata yang dianggap tidak penting seperti tanda baca dan lainnya [17].
3. *Stemming*  
Cara untuk penataan dan pemaparan berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya (*stem*) [22].

### B.6 K-Nearest Neighbor

*K-Nearest Neighbor* (KNN) ialah teknik klasifikasi data untuk mengelompokkan objek berdasarkan *training set* yang jaraknya paling dekat dengan objek tertentu. KNN merupakan suatu metode yang menggunakan algoritma *supervised* dengan hasil *query instance* yang baru diklasifikasikan berdasarkan mayoritas dari kategori

pada KNN [20].

Algoritma KNN ini sangat sederhana, dan menetukannya berdasarkan jarak terdekat dari *query instance* ke *training sample*. Pendek atau panjangnya jarak bisa dihitung berdasarkan *Euclidean Distance* yang direpresentasikan pada Persamaan 2.1.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2} \quad (2.1)$$

Dimana  $d(x_i, x_j)$  dihasilkan dari pengurangan *record* ( $x_i$ ) dari tiap atribut kemudian dikuadratkan, lalu hasil kuadrat dijumlahkan dan diurutkan menurut nilai paling kecil untuk melihat jarak terdekat dengan data uji.

### B.7 K-Fold Cross Validation

*K-Fold Cross Validation* digunakan untuk membagi data menjadi dua bagian, yaitu *testing set* dan *training set*. Data dibagi secara *random* ke himpunan bagian yang dituju. *K-Fold Cross Validation* digunakan untuk menemukan parameter yang unggul dari satu model [10]. Ini dilakukan dengan cara menguji jumlah *error* pada *testing set*. Dalam *cross validation*, data dipecah kedalam  $k$  sampel *cross* dengan aturan yang layak. Dari  $k$  subset data yang digunakan akan dipakai  $k-1$  sampel sebagai data *training* dan 1 sampel sisanya untuk data *testing* [2].

### B.8 Python

*Python* adalah suatu bahasa pemrograman open source yang memakai contoh skrip (*scripting language*) berorientasi objek [15]. *Python* bersifat generik dengan juru bahasa dan dapat digunakan di domain aplikasi luas dan merupakan bahasa pemrograman tingkat tinggi yang fleksibel, sederhana, dan dinamis [14]. Bahasa pemrograman ini dioptimalkan untuk kualitas perangkat lunak, produktivitas pengembang, portabilitas program, dan integritas komponen. *Python* telah digunakan untuk mengembangkan berbagai jenis *software* seperti *internet scripting*, *systems programming*, *user interfaces*, *product costumization*, *numeric programming*. *Python* saat ini menempati urutan ke-4 atau ke-5 bahasa pemrograman yang paling banyak digunakan di dunia. Bahasa pemrograman ini memiliki beberapa fitur yang dimanfaatkan oleh para pengembang perangkat lunak seperti *Multi Paradigm Design*, *Open Source*, *Library Support*, *Portability*, *Extendable*, dan *Scalability* [12].

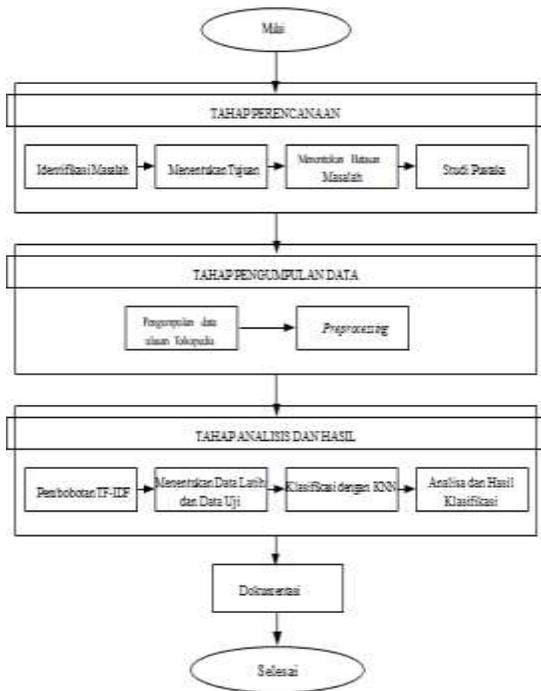
### B.9 Penelitian Terdahulu

Penelitian sebelumnya yang dilakukan oleh Assad Hidayat (2019), Klasifikasi Opini Masyarakat terhadap Toko *Online* menggunakan Algoritma *K-Nearest Neighbor* dan *Probabilistic Neural Network* pada *Twitter* dengan hasil, algoritma KNN memiliki nilai akurasi terbaik pada Lazada dan Blibli jika dibandingkan dengan menggunakan algoritma PNN, yaitu dengan hasil Lazada dengan KNN 71,57% dan PNN 66,71 sementara Blibli di KNN 68,29% dan PNN 65,29% [10].

Selanjutnya penelitian yang dilakukan oleh Astia Weni Syaputri (2020), mengenai Analisis Sentimen pada Ulasan Hotel *Grand Elite* di *website* Traveloka menggunakan Algoritma *K-Nearest Neighbor*, diperoleh hasil bahwa akurasi tertinggi pada k=50 pada split k-1 dengan nilai 90,0% [21].

Kemudian penelitian yang dilakukan oleh Putri Anglenia (2020), mengenai Penerapan Algoritma Klasifikasi terhadap Sentimen Masyarakat menggunakan *Text Mining* pada Toko *Online* lazada diperoleh hasil akurasi KNN lebih tinggi dibandingkan dengan NBC dan PNN dengan akurasi KNN sebesar 72,85%, sedangkan NBC sebesar 66,66%, dan PNN sebesar 67,14% [1].

### C. METODOLOGI PENELITIAN



Gambar 1. Metodologi Penelitian

### D. ANALISA DAN HASIL

#### D.1 Analisa Pendahuluan

Analisa ialah proses pengkajian terhadap suatu penelitian dengan maksud untuk membahas permasalahan secara mendalam. Analisa pendahuluan merupakan tahap awal yang dilakukan untuk mengidentifikasi permasalahan agar mendapat hasil dari penelitian.

##### D.1.1 Pengumpulan Data

Pengumpulan data ialah suatu fase agar memperoleh data yang digunakan pada penelitian. Data yang didapat dilakukan dengan cara *scraping* pada ulasan aplikasi Tokopedia di *Google Play* dengan menggunakan teknik *web scraping* di *Google Collab*.

##### D.1.2 Data Awal

Data awal yang diimplementasikan pada penelitian merupakan ulasan aplikasi Tokopedia di *Google Play* dari Januari 2020 hingga September 2021. Jumlah data yang akan diolah sejumlah 5.254 data. Data awal penelitian terlihat pada Tabel 1.

Tabel 1. Data Awal

No	Ulasan
1	Belum terlalu paham tapi saya akan belajar karna saya kepingin belanja ditokopedia gimana caranya Tolong kal buat iklan yg bener iklan' promo dari Mayora' hadiah handpone unlimited, giliran di klik ujung ujungnya sruh download Tokopedia, hadeeeh
2	Mna pake selamat anda memenangkan hadiah handpone, uang , giliran di cek zonk Tanpa cara begini dh pada tahu koq Tokopedia tu apa, awalnya pengguna sekarang males, auto uninstall. Saya melakukan pesanan barang dan pihak toko mengirimkan tidak sesuai jumlah pesanan, dengan bantuan tokopedia, saya dapat mengclaim kekurangan barang tersebut, terimakasih bantuannya semoga tokopedia bisa menjdi online shop yg nyaman dan terbaik,
...	...
5.254	harga nya mahal sama harga toko sebelah rumah ku.. enakan apk aku laku shopie.. terimakasih sudah mendengar aspirasi dari rakyat standar mohon harga yang pro rakyat dan mudah memahami ekonomi rakyat standar kebawah .. terimakasih .

#### D.2 Cleaning

Data ulasan yang digunakan pada penelitian ini terlebih dahulu akan melalui proses *cleaning*. Proses *cleaning* digunakan untuk mengganti kalimat tidak baku menjadi baku, menghilangkan simbol dan angka serta penggunaan URL, dan juga mengubah *text* menjadi huruf kecil dan menghapus *duplicate*. Hasil *cleaning* terlihat pada Tabel 2.

Tabel 2. Data *Cleaning*

No	Ulasan
1	belum terlalu paham tapi saya akan belajar karna saya kepingin belanja ditokopedia gimana caranya tolong kal buat iklan yg bener iklan' promo dari mayora' hadiah handphone unlimited giliran di klik ujung ujungnya sruh download tokopedia
2	hadееeh mna pake selamat anda memenangkan hadiah handphone uang giliran di cek zonk tanpa cara begini dh pada tahu koq tokopedia tu apa awalnya pengguna sekarang males auto uninstall
3	saya melakukan pesanan barang dan pihak toko mengirimkan tidak sesuai jumlah pesanan dengan bantuan tokopedia saya dapat mengklaim kekurangan barang tersebut terimakasih bantuannya semoga tokopedia bisa menjdi online shop yg nyaman dan terbaik
...	...
5.254	harga nya mahal sama harga toko sebelah rumah ku enakan apk aku laku shapie terimakasih sudah mendengar aspirasi dari rakyat standar mohon harga yang pro rakyat dan mudah memahami ekonomi rakyat standar kebawah terimakasih

### D.3 Text Preprocessing

*Preprocessing text* merupakan salah satu tahapan dari *text mining*. *Preprocessing* sangat berpengaruh terhadap analisis sentimen, terutama pada ulasan yang memiliki banyak *noise*. Pada penelitian ini dilakukan dengan tahapan *preprocessing* yaitu *filtering*, *tokenizing*, dan *stemming*.

Tahapan *filtering* dilakukan untuk menghilangkan kata tidak penting, seperti tanda baca dan lainnya, serta penghilangan terhadap tanda baca dan *stopword* [17]. *Stopword* dilakukan jika pada sebuah kalimat mengandung kata yang sering keluar dan dianggap tidak penting seperti waktu, penghubung, dan sebagainya. Adapun daftar *stopword* yang dipakai yaitu *stoplist* [22]. Berikut merupakan *stopword* dapat dilihat pada Tabel 3.

Tabel 3. Daftar *Stopwords*

No	Stopwords	No	Stopwords	No	Stopwords
1	ada	9	akankah	17	anda
2	adalah	10	akhir	18	andalah
3	adanya	11	akhiri	19	antar
4	adapun	12	akhirnya	20	antara
5	agak	13	aku	21	antaranya
6	agaknya	14	akulah	22	apa
7	agar	15	amat	...	...
8	akan	16	amatlah	758	nan

Tahapan *stemming* merupakan proses penguraian kata yang memiliki imbuhan menjadi kata dasarnya (*stem*). Tahapan *stemming* dilakukan dengan menggunakan modul sastrawi. Untuk tahapan *tokenizing*, pemisahan *string input* berdasarkan kata yang menyusunnya. Tabel 4 merupakan hasil *Text preprocessing*.

Tabel 4. *Text Preprocessing*

No	Ulasan
1	paham ajar karna kepingin belanja ditokopedia gimana tolong kal iklan bener iklan promo mayora hadiah handphone unlimited gilir klik ujung ujung sruh download tokopedia hadееeh mna pake selamat menang hadiah handphone uang gilir cek zonk dh koq tokopedia tu guna males auto uninstall
2	pesan barang toko kirim sesuai pesan bantu tokopedia mengklaim kurang barang terimakasih bantu moga tokopedia menjdi online shop nyaman baik
3	...
...	...
5.254	harga nya mahal harga toko belah rumah ku enak apk laku shapie terimakasih dengar aspirasi rakyat standar mohon harga pro rakyat mudah paham ekonomi rakyat standar bawah terimakasih

### D.4 Pelabelan

Data ulasan yang ada pada aplikasi Tokopedia di *Google Play* merupakan data teks yang belum berlabel (*unsupervised*). Agar menjadi data *supervised*, maka harus dilakukan proses pelabelan. Pada penelitian ini proses pelabelan data komentar aplikasi Tokopedia di *Google Play* dilakukan secara manual oleh seorang pakar bahasa kedalam suatu kelas positif, netral, dan negatif..

Data ulasan yang digunakan sebanyak 5.254 data. Adapun hasil pelabelan untuk ulasan aplikasi Tokopedia dapat dilihat pada Tabel 5.

Tabel 5. Hasil Pelabelan

No	Ulasan	Kelas
1	belum terlalu paham tapi saya akan belajar karna saya kepingin belanja ditokopedia gimana caranya tolong kal buat iklan yg bener iklan' promo dari mayora' hadiah handphone unlimited giliran di klik ujung ujungnya sruh download tokopedia hadееeh mna pake selamat anda memenangkan hadiah handphone uang giliran di cek zonk tanpa cara begini dh pada tahu koq tokopedia tu apa awalnya pengguna sekarang	Netral
2	...	Netral

	males auto uninstal saya melakukan pesanan barang dan pihak toko mengirimkan tidak sesuai jumlah pesanan dengan bantuan tokopedia saya dapat mengklaim kekurangan barang tersebut terimakasih bantuannya semoga tokopedia bisa menjadi online shop yg nyaman dan terbaik	Positif
...	...	...
5.254	harga nya mahal sama harga toko sebelah rumah ku enakan apk aku laku shopie terimakasih sudah mendengar aspirasi dari rakyat standar mohon harga yang pro rakyat dan mudah memahami ekonomi rakyat standar kebawah terimakasih	Negatif

### D.5 Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) dilakukan terhadap kata yang sudah di ekstrak guna mendapatkan bobot (Baharsyah, 2014). Metode TF-IDF banyak diimplementasikan untuk pencarian teks (*text retrieval*) serta pemrosesan teks (*text preprocessing*) [16]. Term Frequency-Inverse Document Frequency (TF-IDF) dipecah berdasarkan pembagian *training set* dan *testing set*. Hasil TF-IDF dapat dilihat pada Tabel 6.

Tabel 6. Hasil Pembobotan TF-IDF

No	Term					
	academy	aja	ajar	akun	...	ya
1	0,00	0,00	0.49	0,00	...	0,00
2	0,00	0,00	0,00	0,00	...	0,00
3	0,00	0,00	0,00	0,00	...	0,00
...	...	...	...	...	...	...
5.254	0,00	0,00	0,00	0,25	...	0,00

### D.6 Pembagian Data Latih dan Data Uji

Berdasarkan metode penelitian yang sudah dijelaskan, pembagian data latih dan data uji pada penelitian ini menggunakan *K-Fold Cross Validation* yang sudah melalui tahap TF-IDF.

#### D.6.1 K-Nearest Neighbor (KNN)

*K-Nearest Neighbor* (KNN) merupakan salah satu algoritma digunakan untuk mengklasifikasikan objek berdasarkan *training set* yang jaraknya paling dekat dengan objek. KNN pada penelitian ini diterapkan agar mendapatkan hasil akurasi klasifikasi pada teks ulasan Tokopedia. Parameter K yang

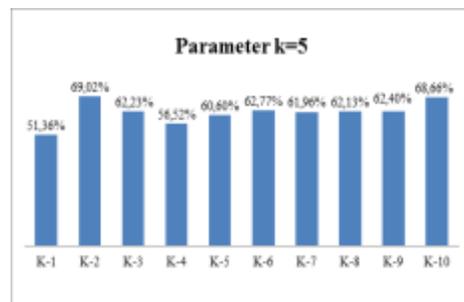
digunakan pada penelitian ini adalah k=5, k=10, k=25, k=50, k=100, dan k=200.

#### D.6.2 Pembagian Data Latih dengan K-Fold Cross Validation

Pada penelitian ini menggunakan 10 *split* yang akan dibentuk untuk menghasilkan sebuah akurasi. Pada algoritma *K-Nearest Neighbor* (KNN) menggunakan enam parameter K yang berbeda untuk klasifikasi, parameter yang digunakan yaitu k=5, k=10, k=25, k=50, k=100, dan k=200.

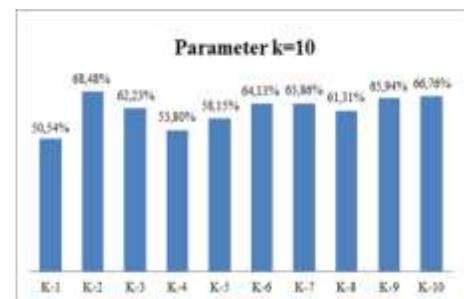
#### D.6.3 K-Fold Cross Validation KNN

*K-Fold Cross Validation* digunakan untuk menguji performa dari metode KNN untuk mendapatkan hasil yang lebih valid. *K-Fold Cross Validation* ialah metode penilaian untuk meningkatkan flaksibilitas suatu model. Hasil pembagian *training set* serta *testing set* menggunakan *K-Fold Cross Validation* parameter k=5 dapat dilihat pada Gambar 2.



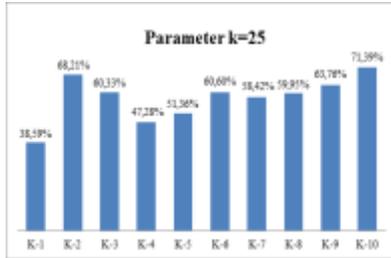
Gambar 2. Hasil Percobaan Parameter k=5

Dari Gambar 2 didapatkan hasil K-2 memiliki akurasi yang paling tinggi dengan 69,02%, dan pembobotan dengan akurasi terendah pada K-1 dengan nilai 51,36%.



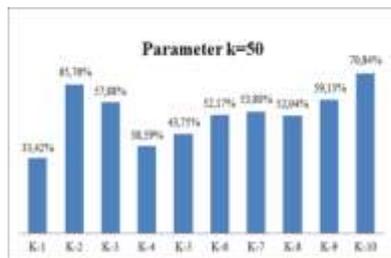
Gambar 3. Hasil Percobaan Parameter k=10

Dari Gambar 3 didapatkan hasil K-2 memiliki akurasi yang paling tinggi dengan 68,48%, dan pembobotan dengan akurasi terendah pada K-1 dengan nilai sebesar 50,54%.



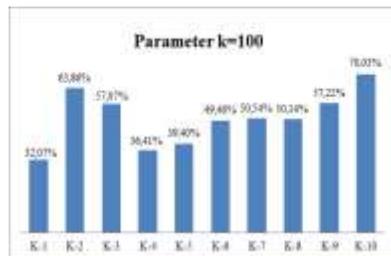
Gambar 4. Hasil Percobaan Parameter k=25

Dari Gambar 4 didapatkan hasil K-10 memiliki akurasi yang paling tinggi dengan 71,39%, dan pembobotan dengan akurasi terendah pada K-1 dengan nilai sebesar 38,59%.



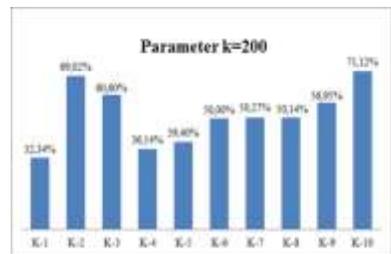
Gambar 5. Hasil Percobaan Parameter k=50

Dari Gambar 5 didapatkan hasil K-10 memiliki akurasi yang paling tinggi dengan 70,84%, dan pembobotan dengan akurasi terendah pada K-1 dengan nilai sebesar 33,42%.



Gambar 6. Hasil Percobaan Parameter k=100

Dari Gambar 6 didapatkan hasil K-10 memiliki akurasi yang paling tinggi dengan 70,03%, dan pembobotan dengan akurasi terendah pada K-1 dengan nilai sebesar 32,07%.



Gambar 7. Hasil Percobaan Parameter k=200

Dari Gambar 7 didapatkan hasil K-10 memiliki akurasi yang paling tinggi dengan 71,12%, dan pembobotan dengan akurasi terendah pada K-1 dengan nilai sebesar 32,34%.

Berdasarkan hasil percobaan ke-enam parameter yang digunakan, didapatkan hasil percobaan nilai paling tinggi yaitu pada parameter k=25 pada *split* ke K-10, dan nilai paling rendah pada parameter k=10 di *split* K-2.

#### D.6.4 Menghitung Jarak Data pada KNN

Pada penelitian ini, perhitungan jarak data pada KNN menggunakan rumus *Eculidean Distance* dengan nilai tertinggi yaitu pada Parameter k=25 pada *split* K-10 dengan nilai 71,39%. Hasil nilai jarak dapat dilihat pada Tabel 7.

Tabel 7. Hasil Jarak Data pada KNN

No	Negatif	Netral	Positif	Kelas
1	0,05	0,19	0,76	Positif
2	0,04	0,09	0,88	Positif
3	0,00	0,80	0,20	Netral
4	0,03	0,03	0,93	Positif
5	0,09	0,19	0,72	Positif
...	...	...	...	...
1573	0,08	0,32	0,60	Positif
1574	0,12	0,08	0,80	Positif
1575	0,04	0,24	0,72	Positif
1576	0,04	0,24	0,72	Positif
1577	0,00	0,00	1,00	Positif

#### D.7 Visualisasi Data

Visualisasi data bertujuan untuk melihat jumlah frekuensi kata yang sering muncul pada setiap ulasan. Visualisasi data dilakukan dengan memanfaatkan situs yaitu <https://wordart.com/>. Hasil dari visualisasi data berdasarkan opini positif, negatif, dan netral.

##### D.7.1 Visualisasi Data Ulasan Positif

Hasil visualisasi pada ulasan positif terlihat pada Gambar 8.



Gambar 8. Visualisasi Data Ulasan Positif

### D.7.2 Visualisasi Data Ulasan Netral

Hasil visualisasi pada ulasan netral terlihat pada Gambar 9.



Gambar 9. Visualisasi Data Ulasan Netral

### D.7.3 Visualisasi Data Ulasan Negatif

Hasil visualisasi pada ulasan negatif terlihat pada Gambar 10.



Gambar 10. Visualisasi Data Ulasan Negatif

### D.8 Kontribusi Penelitian

Berdasarkan penelitian yang telah dilakukan sebelumnya, terdapat 2 unsur kebaruan atau kontribusi penelitian:

1. Penelitian ini menyelesaikan pelabelan data yang dilakukan secara manual terhadap 5.254 data teks komentar aplikasi Tokopedia di *Google Play* dengan tiga kelas yaitu positif, netral, dan negatif.
2. Hasil akurasi penelitian menggunakan pembagian data *K-Fold Cross Validation* dengan membagi training set dan testing set secara acak kedalam  $k$ . Dimana proses pengujian menggunakan 10  $k$ -fold, dengan percobaan klasifikasi algoritma *K-Nearest Neighbor* menggunakan enam parameter  $K$  yaitu  $k=5$ ,  $k=10$ ,  $k=25$ ,  $k=50$ ,  $k=100$ , dan  $k=200$  terhadap komentar aplikasi Tokopedia di *google play*. Didapat nilai akurasi tertinggi yaitu pada parameter  $K=25$  di *split* 10 dengan nilai akurasi 71,39%. Berdasarkan hasil tersebut disimpulkan bahwa klasifikasi ulasan aplikasi Tokopedia menggunakan algoritma *K-Nearest Neighbor* dengan nilai akurasi terdapat pada parameter  $k=25$  di *split* 10.

### D.9 Analisis Hasil Penelitian

Hasil analisis Penelitian berguna untuk menyelaraskan hasil dengan tujuan dan manfaat pada

penelitian serta berkaitan juga dengan orang, organisasi, dan proses bisnis.

#### 1. Orang

Hasil penelitian ini nantinya dapat memberikan pengetahuan kepada masyarakat atau publik mengenai opini yang biasa disampaikan oleh masyarakat sehingga masyarakat dapat mengetahui informasi mengenai komentar Tokopedia.

#### 2. Organisasi

Hasil penelitian pada penerapan klasifikasi pada komentar Tokopedia nantinya dapat dijadikan rekomendasi bagi organisasi terkait.

#### 3. Proses Bisnis

Hasil penelitian ini memiliki tujuan dan manfaat yang nantinya dapat dijadikan sebagai rekomendasi bagi Tokopedia dalam mengambil keputusan guna meningkatkan proses bisnis.

### E. KESIMPULAN

Adapun beberapa kesimpulan yang didapatkan dari penelitian ini adalah:

1. Penelitian ini menggunakan metode pembagian data dengan *K-Fold Cross Validation* dengan membagi data menjadi *training set* dan *testing set* secara acak dengan 10  $k$ -fold. Percobaan KNN dengan  $k$ -fold menggunakan 6 parameter  $k$ , yaitu  $k=5$ ,  $k=10$ ,  $k=25$ ,  $k=50$ ,  $k=100$ , dan  $k=200$ . Dari 6 parameter tersebut didapatkan akurasi tertinggi yaitu pada  $k=25$  yaitu pada *fold*-10 dengan nilai 71,39%.
2. Frekuensi kemunculan kata paling banyak untuk ulasan tokopedia yaitu pada kelas positif yaitu kata “tokopedia”, “saya”, dan “dan”. Untuk kelas netral kemunculan kata paling banyak yaitu kata “saya”, “di”, dan “bisa”. Sementara untuk kelas negatif, kemunculan kata yang paling banyak yaitu pada kata “saya”, “di”, dan “tokopedia”.

### REFERENSI

- [1] Anglenia, Putri. 2020. “Penerapan Algoritma Klasifikasi terhadap Sentimen Masyarakat menggunakan *Text Mining* pada Toko *Online Lazada*”. Skripsi. Universitas Islam negeri Sultan Syarif Kasim Riau
- [2] Aszani, A., dan Mustakim, M. (2016). Estimasi pola cuaca wilayah pekan- baru menggunakan probabilistic neural network.

- Dalam Seminar nasional teknologi informasi komunikasi dan industri
- [3] Berry, M., dan J, K. 2010. *Systems Analysis And Design In A Changing World*. United kingdom: WILEY
- [4] CloudHost. (2020). Pengertian e-commerce dan contohnya, komponen, jenis, dan manfaat e-commerce. Retrieved from <https://idcloudhost.com/pengertian-e-commerce-dan-contohnya-komponen-jenis-dan-manfaat-e-commerce/>
- [5] Faadilah, A. (2020). Analisis sentimen pada ulasan aplikasi tokopedia di google play store menggunakan metode long short term memory (B.S. thesis). Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah.
- [6] Fadillah, Riszki. (2021). Penerapan Text Mining Terhadap Badan Penyelenggara Jasa Jaminan Sosoal Kesehatan Untuk Klasifikasi Sentimen Analisis. Skripsi. Universitas Islam Negeri Sultan Syarif Kasim Riau
- [7] Feldman, R., dan Sanger, J. 2007. *The Text Mining Handbook: Advanced Approaches In Analyzing Unstructured Data*. New York: Cambridge University Press
- [8] Gumilang, Z. A. N., dkk. (2018). Implementasi naive bayes classifier dan asosiasi untuk analisis sentimen data ulasan aplikasi e-commerce shopee pada situs google play.
- [9] Hakim, A. (2018). Klasifikasi sentimen terhadap bukalapak dengan menggunakan metode naive bayes classifier (Unpublished doctoral dissertation). Universitas Islam Negeri Sultan Syarif Kasim Riau.
- [10] Hidayat, Assad. 2019. "Klasifikasi opini masyarakat terhadap toko online menggunakan algoritma KNN dan PNN pada twitter". Skripsi. Universitas Islam Negeri Sultan Syarif Kasim Riau
- [11] Iprice. 2021. "Peta E-Commerce Indonesia". <<https://iprice.co.id/insights/map-ofecommerce/>> [diakses 24 November 2020]
- [12] Lutz, M. 2010. *Programming Python*. Fourth Edition ed. Sebastopo: O'Reilly Media, Inc.
- [13] Mardi, Y. (2017). Data mining: Klasifikasi menggunakan algoritma c4. 5. Edik Informatika, 2(2), 213–219.
- [14] Nosrati, M. (2011). Python: An appropriate language for real world programming. World Applied Programming, 1(2), 110–117
- [15] Perkasa, T. 2014. Rancang Bangun Pendeteksi Gerak Menggunakan Metode Image Subtraction Pada Single Board Computer (sbc) (unpublish doctoral dissertation). Skripsi. Institut Bisnis dan Informatika Stikom Surabaya
- [16] Pristiyani, R, I., Fauzi, M, A., dan Muflikhah, L. 2018. "Sentimen Analisis Peringkasan Review Film Menggunakan Metode Information Gain dan K-Nearest Neighbor". *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Computer (J-PTIHK)* Universitas Brawijaya
- [17] Purnomo, W. G., dan Purnomo, P. P. (2017). Akurasi text mining menggunakan algoritma k-nearest neighbour pada data content berita sms. Format, 6(1), 1–13.
- [18] Ratnawati, F. 2018. "Implementasi Algoritma Naïve Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter". *Jurnal Invotek Polbeng*, 3(1)
- [19] Rohandi, Mochamad Malik Akbar. 2017. "Effectiveness C2C e-commerce Media In Bandung (Case Study at Tokopedia.com and Bukalapak.com)". *Jurnal Manajemen dan Bisnis (Performa)* Universitas Islam Bandung
- [20] Simanjuntak, T. (2014). Implementasi modified k-nearest neighbor dengan otomatisasi nilai k pada pengklasifikasian penyakit tanaman kedelai (Unpublished doctoral dissertation). Universitas Brawijaya
- [21] Syaputri, Astia Weni. 2020. "Analisa Sentimen pada Ulasan Hotel Grand Elite di website Traveloka menggunakan Algoritma K-Nearest Neighbor". Skripsi. Universitas Islam negeri Sultan Syarif Kasim Riau
- [22] Tala, F. Z. "A Study Of Stemming Effects On Information Retrieval in Bahasa Indonesia". The Netherlands: Universiteit van Amsterdam. 2003.
- [23] Valevi, N. W. (2016). Strategi pt.tokopedia dalam promosi "rayakan kebebasan-mu" melalui media offline untuk memperbaharui kesadaran masyarakat pada tahun 2016

(Unpublished doctoral dissertation). Universitas  
Muhammadiyah Yogyakarta

- [24] Zhong, N., Li, Y., dan Wu, S-T. 2010.  
“Effective Pattern Discovery For Text Mining.  
IEEE Transactions On Knowledge And Data