p-ISSN : 2443-2210
e-ISSN : 2443-2229

K-Nearest Neighbor Berbasis *Particle Swarm Optimization* untuk Analisis Sentimen Terhadap Tokopedia

http://dx.doi.org/10.28932/jutisi.v6i2.2658

Dicki Pajri ✉ #1, Yuyun Umaidah*2, Tesa Nur Padilah#3

#*Program Studi Teknik Informatika,Universitas Singaperbangsa Karawang*
Jl. HS. Ronggo Waluyo, Telukjambe Timur, Karawang
[1]dicki.16070@student.unsika.ac.id
[3]tesa.nurpadilah@staff.unsika.ac.id

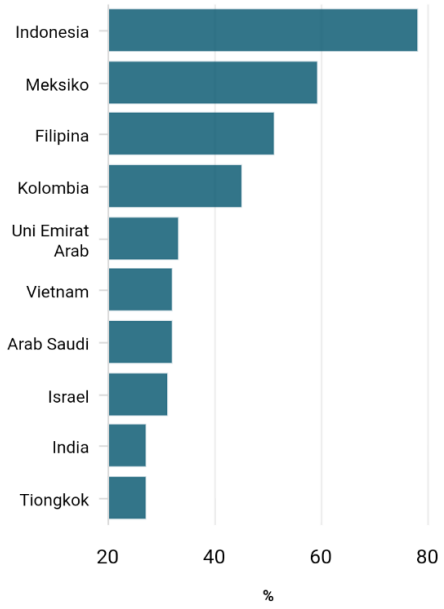
**Program Studi Teknik Informatika,Universitas Singaperbangsa Karawang*
Jl. HS. Ronggo Waluyo, Telukjambe Timur, Karawang
[2]yuyun.umaidah@staff.unsika.ac.id

Abstract — **Tokopedia is a popular marketplace used by e-commerce in Indonesia. Customers' perception of Twitter towards Tokopedia can be used as an important source of information and can be processed into useful insights. Sentiment analysis is a solution that can be used to process the customers' perception using K-Nearest Neighbor based on Particle Swarm Optimization. The purpose of this study is to classify customers' perception based on positive, neutral, and negative classes. The test is carried out with four different scenarios and k values which are evaluated using a confusion matrix. Evaluation results showed the distribution of the *dataset* is 90:10 and the value of k = 1 is the best evaluation result, which is 88.11%. The feature selection was used for results by using Particle Swarm Optimization. The Particle Swarm Optimization used 20 iterations and 10 particles. It produced 97.9% the best evaluation accuracy, 96.17% precision, 96.62% recall, and 96.39% f-measure.**

Keywords—***K-Nearest Neighbor; Particle Swarm Optimization; Sentiment Analysis;***

I. PENDAHULUAN

Internet kini menjadi salah satu kebutuhan sehari-hari, salah satunya bagi masyarakat Indonesia. Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) menyebutkan jumlah pengguna internet di Indonesia pada tahun 2019 sebanyak 171,17 juta pengguna. Angka ini mengalami pertumbuhan sebesar 10,12% dari tahun sebelumnya [1]**.** Jumlah pengguna internet yang besar tersebut mengakibatkan berkembangnya berbagai bisnis online, salah satunya yaitu *e-commerce*. Berikut ini adalah diagram sepuluh negara dengan pertumbuhan *e-commerce* tercepat [2].



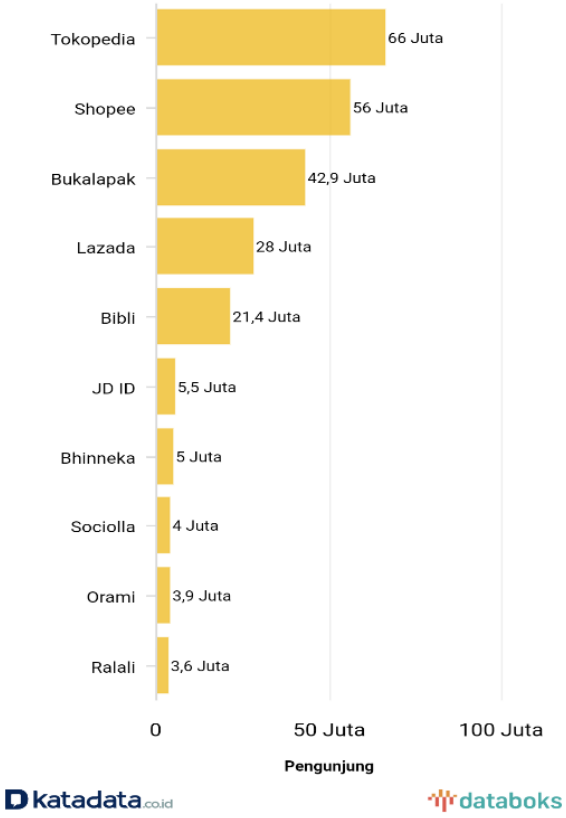
Gambar 1. Sepuluh negara dengan pertumbuhan *e-commerce* tercepat

Gambar 1 merupakan grafik sepuluh negara dengan pertumbuhan *e-commerce* tercepat di dunia tahun 2018 seperti yang dilaporkan oleh *Merchant Machine* dari situs katadata.co.id, tingkat pertumbuhan *e-commerce* di Indonesia menempati urutan pertama dibandingkan dengan negara lain yaitu mencapai 78%. Selain itu, *We Are Social* menyebutkan bahwa 96% pengguna internet di Indonesia pernah mengakses *e-commerce* [3]**.**

Jumlah pengguna *e-commerce* di Indonesia pun mengalami pertumbuhan. *Statista* dalam katadata.co.id mencatat jumlah pengguna *e-commerce* di Indonesia semakin meningkat tiap tahunnya. Pada tahun 2017 jumlah

pengguna *e-commerce* di Indonesia mencapai 139 juta pengguna dan naik menjadi 154,1 juta pengguna di tahun 2018. Angka ini diprediksi akan terus tumbuh pada tahun-tahun berikutnya [4].

Tokopedia merupakan salah satu *marketplace* populer yang digunakan oleh *e-commerce* yang ada di Indonesia. Berdasarkan data yang dirilis oleh *iPrice Group* pada tahun 2019, Tokopedia merupakan *marketplace* dengan jumlah pengunjung terbesar pada kuartal III 2019 di Indonesia. Laporan *E-Warungs*: *Indonesia New Digital Battleground* menyebutkan Tokopedia merupakan *e-commerce* dengan nilai transaksi terbesar yang ada di Indonesia. Berikut adalah daftar 10 *marketplace* dengan pengunjung terbesar kuartal III tahun 2019 [5].



Gambar 2. *Marketplace* dengan pengunjung terbesar kuartal III 2019

Gambar 2 merupakan grafik dari *marketplace* dengan pengunjung terbesar kuartal III tahun 2019. Peringkat pertama yaitu Tokopedia dengan 66 juta, disusul oleh Shopee dengan 56 juta dan diperingkat terakhir yaitu Ralali dengan 3.6 juta pengunjung.

Untuk selalu terhubung dengan pelanggannya, biasanya sebuah *brand* mempunyai akun sosial media. Banyak sosial media yang digunakan, salah satunya Twitter. Twitter digunakan karena kemudahan dalam berinteraksi antar penggunanya. Tokopedia sebagai salah satu *brand* di Indonesia pun mempunyai akun Twitter. Pada tahun 2019, akun Twitter Tokopedia (@Tokopedia) menempati urutan teratas daftar 10 *brand* di Indonesia. *Top brand* ini diukur dari interaksi yang terjadi antara *brand* dengan pengguna [6].

Dikutip dari Techinasia, Bhayu Rareli Arsyad selaku *Senior Customer Experience* Tokopedia mengatakan bahwa suara pelanggan di media sosial menjadi salah satu sumber informasi yang penting dan dapat diolah menjadi wawasan yang bermanfaat, selain itu suara pelanggan dapat memberikan rekomendasi kepada Tokopedia dan dapat ditindaklanjuti oleh Tokopedia sehingga layanannya menjadi memuaskan. Salah satu poin yang diukur adalah *social sentiment* yang diberikan oleh pengguna media sosial yang berupa testimoni yang dapat meyakinkan pelanggan baru untuk beralih ke Tokopedia [7].

Suara pelanggan yang berupa teks pada Twitter dapat diolah dengan menggunakan *text mining*. *Text mining* adalah sebuah proses ekstraksi informasi berkualitas tinggi dari teks terstruktur, semi terstruktur dan tidak terstruktur [8]. *Text mining* dapat diterapkan untuk analisis sentimen. Analisis sentimen dapat digunakan untuk mengelompokkan polaritas dari teks untuk mengetahui apakah opini atau pendapat dari sebuah teks tersebut bersifat positif atau negatif [9]. Hasil akhir pada analisis sentimen ini menjadi suatu alat untuk memberikan informasi berupa sekumpulan teks untuk digunakan oleh orang yang mengambil atau memberi keputusan.

Berdasarkan latar belakang tersebut peneitian ini bertujuan untuk menerapkan algoritma *K-Nearest Neighbor* berbasis *Particle Swarm Optimization* untuk menganalisis data teks dari Twitter mengenai Tokopedia. Penelitian ini diharapkan dapat membantu pihak Tokopedia untuk mengembangkan bisnisnya.

II. Penelitian Sebelumnya

Penelitian dilakukan oleh Abdul Malik Zuhdi, Ema Utami dan Suswanto Raharjo mengenai Analisis Sentimen Twitter Terhadap Capres Indonesia 2019 dengan Metode K-NN, *dataset* yang digunakan sebanyak 1.000 data, kelas sentimen dibagi menjadi 3 kelas yaitu positif, negatif, dan netral. Hasil pengujian dengan pembagian 70% data latih dan 30% data uji didapat bahwa nilai *k* terbaik adalah 3, dengan tingkat akurasi mencapai 81,83% [10].

Lila Dini Utami, Hilda Rachmi, dan Dini Nurlaela melakukan Komparasi Algoritma Klasifikasi Pada Analisis *Review* Hotel dengan membandingkan *Naïve Bayes*, *Support Vector Machine* dan *K-Nearest Neighbor*. Hasil pengujian didapatkan bahwa algoritma *K-Nearest Neighbor* menghasilkan nilai akurasi yang paling tinggi dibandingkan dengan *Naïve Bayes* dan *Support Vector Machine* [11].

Melisa Winda Pertiwi melakukan perbandingan algoritma untuk analisis sentimen opini publik mengenai sarana transportasi mudik tahun 2019 dengan *dataset* berjumlah 347 data, Hasil penelitian didapatkan bahwa akurasi algoritma *K-Nearest Neighbor* lebih tinggi jika dibandingkan dengan algoritma *Support Vector Machine*, *Neural Network*, dan *Naïve Bayes* [12].

Dinar Ajeng Kristianti melakukan Analisis Sentimen Review Produk Kosmetik Melalui Komparasi *Feature Selection* antara *Genetic Algorithm* dan *Particle Swarm Optimization*. Hasilnya adalah dibandingkan dengan *Genetic Algorithm*, *Particle Swarm Optimization* lebih dapat meningkatkan akurasi [13].

Riswanti, Irwan Budiman, dan Ahmad Rusadi Arrahimi melakukan penelitian *Sentiment Analysis* SVM dan SVM-PSO Pada Kolom Komentar Evaluasi Dosen dengan *dataset* berjumlah 8.326 data. Penelitian dilakukan dengan algoritma *Support Vector Machine* menghasilkan parameter *cost* terbaik adalah 1. Kemudian dioptimasi dengan menggunakan algoritma *Particle Swarm Optimization* menghasilkan akurasi tertinggi 82,59% dengan jumlah partikel 30 dan iterasi 10 [14].

III. LANDASAN TEORI

A. *Text Mining*

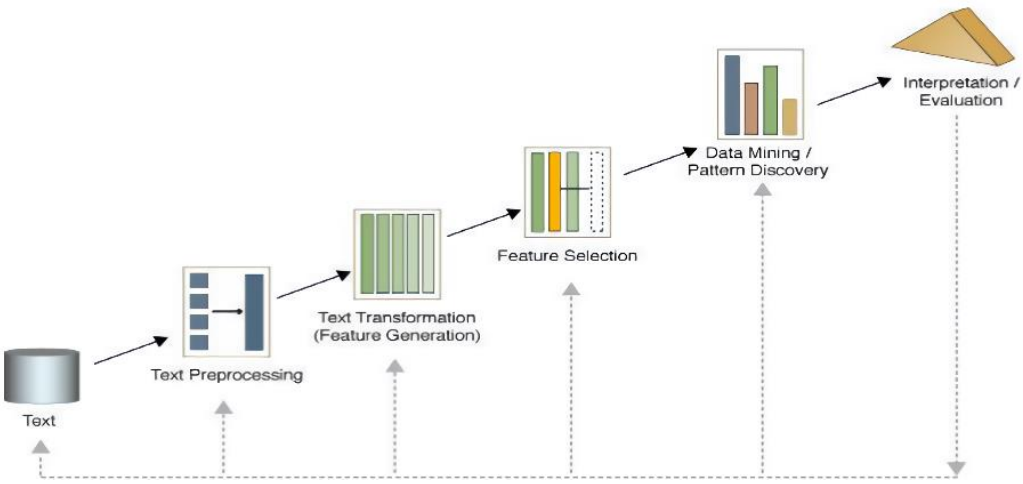
Text Mining atau *Knowledge Discovery from Text* (KDT) adalah sebuah proses ekstraksi informasi berkualitas tinggi dari teks terstruktur (data RDBMS), semi terstruktur (XML dan JSON) dan tidak terstruktur seperti dokumen, video, dan gambar [8].

B. *Analisis Sentimen*

Analisis sentimen adalah bidang studi yang menganalisis opini, sentimen, evaluasi, penilaian, sikap, dan emosi orang-orang terhadap entitas seperti produk, layanan, organisasi, individu, isu, peristiwa, topik, dan atribut mereka [15]. Analisis sentimen akan mengelompokkan polaritas dari teks yang ada dalam dokumen untuk mengetahui pendapat yang dikemukakan dalam dokumen apakah bersifat positif, negatif, atau netral [12]. Hasil akhir pada analisis sentimen ini menjadi suatu alat untuk memberikan informasi berupa sekumpulan teks untuk digunakan oleh orang yang mengambil atau memberi keputusan.

C. *Knowledge Discovery in Database (KDD)*

Knowledge Discovery in Database (KDD) adalah proses menganalisis terstruktur untuk memperoleh informasi yang benar, baru, dan menemukan pola dari data yang besar dan kompleks. *Data mining* menjadi inti dari proses *Knowledge Discovery in Database* (KDD) yaitu dengan menggunakan algoritma tertentu untuk mengeksplorasi data, membangun model, dan menemukan pola yang belum diketahui [16]. Proses *text mining* dengan *Knowledge Discovery in Database* (KDD) sama dengan pada *data mining* [17]. Gambar 3 merupakan tahapan dalam *Knowledge Discovery in Database* [18].



Gambar 3.Tahapan *Knowledge Discovery in Database* (KDD)

1. *Text*: Data berupa teks merupakan sebuah fragmen yang dianggap sebagai unit. Data dapat berupa buku, paragraf, abstrak, maupun judul.
2. *Text Prepocessing*: Tahap ini bertujuan untuk mengurangi atribut yang kurang berpengaruh terhadap proses klasifikasi Secara umum *text processing* terdiri dari *case folding*, *cleaning*, *convert emoticon*, *tokenizing*, *stopword removal*, dan *stemming*.
3. *Text Transformation:* Sebuah dokumen diwakili oleh fitur (kata-kata) yang dikandungnya. Dokumen harus ditransformasikan dari versi teks lengkap menjadi bentuk vektor dokumen.
4. *Feature Selection*: *Feature selection* merupakan tahap pengurangan dimensi dari *noise* yang mengganggu hasil penelitian. Pemilihan fitur penting digunakan dalam pembuatan model karena data mengandung banyak fitur yang berlebihan dan tidak relevan.
5. *Data mining/Pattern Discovery*: Tahap yang digunakan pada *text mining* sama seperti pada *data mining*. Misalnya klasifikasi, asosiasi dan lain-lain.

6. *Evaluation*: Evaluasi dilakukan terhadap hasil dari *pattern discovery*, umumnya menggunakan suatu nilai performansi.

D. K-Nearest Neighbor

K-Nearest Neighbor merupakan suatu pendekatan klasifikasi yang mencari semua data latih yang relatif mirip dengan data uji. *K-Nearest Neighbor* merupakan teknik klasifikasi *lazy learning* karena teknik ini tidak membangun model klasifikasi terlebih dahulu [19]. Dalam menentukan hasil klasifikasi, algoritma *K-Nearest Neighbor* melihat jarak terdekat dari objek dengan masing-masing kelompok [20]. Tahapan dalam algoritma *K-Nearest Neighbor* adalah sebagai berikut:

1. Menentukan jumlah pada tetangga *k*.
2. Menghitung jarak objek dengan menggunakan *euclidean distance*.

$$d\,(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad [1]$$

di mana *d* adalah jarak, *x* adalah data latih dan *y* adalah data uji.

3. Urutkan jarak tersebut dan tentukan tetangga mana yang terdekat berdasarkan jarak minimum ke-*k*.
4. Gunakan kategori mayoritas sebagai nilai prediksi dari data yang baru

E. Particle Swarm Optimization

Particle Swarm Optimization merupakan metode pencarian populasi yang terinspirasi dari pergerakan burung dan ikan dalam mencari makan. *Particle Swarm Optimization* banyak digunakan untuk memecahkan masalah optimasi dan masalah seleksi fitur [21]. Berikut adalah langkah-langkah dalam *Particle Swarm Optimization*:

1. *Inisialiasi*: Inisialisasi kecepatan awal pada iterasi ke-0, dapat dipastikan bahwa nilai kecepatan awal semua partikel adalah 0. Inisialisasi posisi awal partikel pada iterasi ke-0, posisi awal partikel dibangkitkan dengan persamaan

$$x = x_{min} + rand[0,1] \times (x_{max} - x_{min}) \qquad [2]$$

Keterangan:
x = posisi partikel
x_{min} = posisi partikel minimal
x_{min} = posisi partikel maximal
$rand[0,1]$= nilai *random* antara 0 dan 1 berdistribusi uniform dalam interval 0 dan 1
Inisialisasi *pBest* dan *gBest* pada iterasi ke-0, *pBest* akan disamakan dengan nilai posisi awal partikel. Sedangkan *gBest* dipilih dari satu *pBest* dengan *fitness* tertinggi.

2. *Perbaharui Kecepatan:* Perbaharui kecepatan dengan rumus

$$v_{i,j}^{t+1} = w.v_{i,j}^{t} + c_1.r_1\left(pBest_{i,j}^{t} - x_{i,j}^{t}\right) + c_2.r_2\left(gBest_{g,j}^{t} - x_{i,j}^{t}\right) \qquad [3]$$

Keterangan:
$v_{i,j}^{t+1}$ = kecepatan partikel *i* dimensi *j* pada iterasi *t*
w = bobot inersia
$v_{i,j}^{t}$ = kecepatan partikel *i* dimensi *j* pada iterasi *t*
w = bobot inersia
c_1 = konstanta kecepatan 1
c_1 = konstanta kecepatan 2
r_1, r_2= nilai acak antara 0 dan 1
$pBest_{i,j}^{t}$ = posisi terbaik dari partikel *i* dimensi *j* pada iterasi *t*
$gBest_{g,j}^{t}$ = global optimum dari partikel *g* dimensi *j* pada iterasi *t*
$x_{i,j}^{t}$= posisi partikel *i* dimensi *j*

3. *Perbaharui Posisi dan Hitung Fitness:* Perbaharui posisi dengan rumus

$$x_{i,j}^{t+1} = x_{i,i}^{t} + v_{i,j}^{t+1} \qquad [4]$$

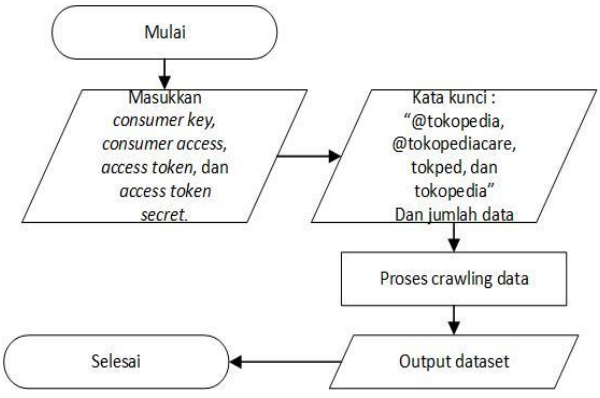
Keterangan:
$x_{i,j}^{t+1}$= posisi partikel *i* dimensi *j* pada iterasi *t*
$x_{i,j}^{t}$= posisi partikel *i* dimensi *j*
$v_{i,j}^{t+1}$ = kecepatan partikel *i* dimensi *j* pada iterasi *t*
dan hitung *fitness* dengan nilai akurasi dari model terbaik.

4. *Perbaharui pBest dan gBest:* Dilakukan perbandingan antara *pBest* pada iterasi sebelumnya dengan hasil dari *update* posisi. *Fitness* yang lebih tinggi akan menjadi *pBest* yang baru. *pBest* terbaru yang memiliki nilai *fitness* tertinggi akan menjadi *gBest* yang baru.

IV. METODOLOGI PENELITIAN

A. Teknik Pengumpulan Data

Pengumpulan data pada penelitian ini dilakukan dengan proses *crawling* dari Twitter menggunakan *Twitter API key* dengan bantuan *tools* RStudio. Berikut adalah alur dari tahap pengumpulan data.

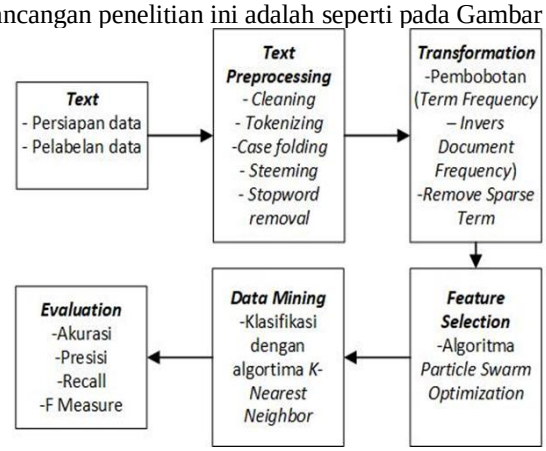


Gambar 4.Teknik Pengumpulan Data

Gambar 4 merupakan alur dari teknik pengumpulan data. Dimulai dari memasukkan *Twitter API Key*, kemudian memasukkan kata kunci dan jumlah data, lalu proses *crawling* data, dan *output* berupa *dataset*.

B. Rancangan Penelitian

Rancangan penelitian ini adalah seperti pada Gambar 5.



Gambar 5. Alur metode penelitian

Berikut adalah penjelasan dari Gambar 5.

1. *Text:* Data yang telah diambil akan dilakukan penyeleksian. Data yang telah diambil kemudian dilakukan pelabelan data secara manual oleh Ahli Bahasa Indonesia yaitu Bapak Priyanto, S.Pd., M.Pd. Dosen Program Studi Pendidikan Bahasa Indonesia Universitas Negeri Jambi.

2. *Text Preprocessing:* Tahap *text preprocesssing* merupakan tahap yang penting untuk dilakukan, bertujuan untuk mengurangi atribut yang kurang berpengaruh terhadap proses *data mining*.

3. *Transformation:* Setelah melakukan tahap *text prepocessing*, tahap selanjutnya yaitu *dataset* ditransformasikan menjadi matriks yang berisi bobot kata. Bobot kata kemudian dihitung menggunakan *Term Frequency–Invers Document Frequency*. Lalu dilakukan tahap penghapusan kata yang jarang muncul dengan *Remove Sparse Term*.

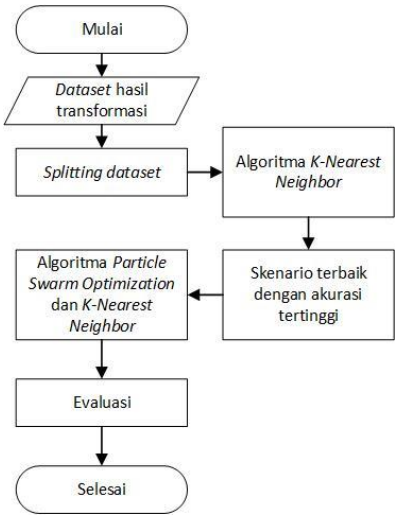
4. *Feature Selection*: Setelah dilakukan tahap transformasi, tahap selanjutnya yaitu seleksi fitur dengan menggunakan algoritma *Particle Swarm Optimization*. Tujuannya adalah untuk meningkatkan hasil evaluasi. Pengujian dilakukan dengan menggunakan parameter 10, 20, 30, 40, 50 iterasi dengan masing-masing yaitu 10 dan 20 partikel. Proses pengerjaan *feature selection* akan dikerjakan pada tahap *data mining* yaitu setelah menemukan skenario dengan nilai akurasi tertinggi. Nilai ini akan digunakan sebagai *fitness function* pada algoritma *Particle Swarm Optimization*.

5. *Data mining:* Pada tahap ini yaitu dilakukan proses *data mining* menggunakan Algoritma *K-Nearest Neighbor* dan memvisualisasikan ke dalam *word cloud*. *Dataset* dibagi menjadi dua, yaitu data training dan data testing. Data training digunakan untuk membangun model sedangkan data testing digunakan untuk mengetahui keakuratan dari model yang dibangun, umumnya digunakan untuk memprediksi kelas atau label. Pembagian *dataset* diterapkan teknik *percentage split*. Terdapat 4 skenario yang akan digunakan seperti pada TABEL I yaitu:

TABEL I
SKENARIO PEMBAGIAN *DATASET*

Skenario	***Data Training***	***Data Testing***
Skenario 1	60%	40%
Skenario 2	70%	30%
Skenario 3	80%	20%
Skenario 4	90%	10%

Selain itu, nilai k yang akan digunakan dalam proses *data mining* menggunakan algoritma *K-Nearest Neighbor* yaitu 1, 2, 3, 5, 6, 7, 8, 9, 10, dan 15.



Gambar 6. Alur *data mining*

Pada Gambar 6, *dataset* yang telah dilakukan *text preprocessing* selanjutnya akan dibagi menjadi data training dan data testing. Setelah itu dilakukan dua pemodelan, pertama dengan algoritma *K-Nearest Neighbor* tanpa seleksi fitur dan yang kedua dilakukan seleksi fitur dengan algoritma *Particle Swarm Optimization*. Setelah itu dilakukan tahap evaluasi.

6. *Evaluation:* Tahap ini adalah tahap evaluasi untuk mengukur hasil *data mining*. Evaluasi dapat dilakukan dengan perhitungan dari hasil *confusion matrix*. Nilai yang dapat dihitung dari *confusion matrix* adalah nilai *accuracy*, *precission*, *recall*, dan *f-measure*. Evaluasi juga dilakukan untuk membandingkan hasil *data mining* dengan *feature*

selection algoritma *Particle Swarm Optimization* dan tanpa seleksi fitur.

V. HASIL DAN PEMBAHASAN

A. Text

Pada data yang telah dikumpulkan kemudian dilakukan tahap penyeleksian. Dari 90 variabel, hanya variabel *text* yang digunakan dalam analisis sentimen. Variabel ini berisi *tweet* dari pengguna Twitter mengenai Tokopedia. Data *text* tersebut kemudian dilakukan penyeleksian lagi karena ada data yang tidak terbaca oleh RStudio, *tweet* berbahasa asing, serta data yang hanya berisi *mention*, *link*, dan *hashtag*. Hal ini dilakukan untuk memudahkan dalam pemberian *class.*

Setelah dilakukan penyeleksian, data yang semula berjumlah 17.298 menjadi 8.878 data. Tahap selanjutnya yaitu pemberian label atau *class*. *Class* data dibagi menjadi 3 yaitu positif, netral, dan negatif. Pemberian *class* dilakukan manual oleh peneliti kemudian divalidasi oleh ahli Bahasa Indonesia. Berikut ini perbandingan jumlah *class* antara peneliti dan ahli Bahasa Indonesia dapat dilihat pada TABEL II.

TABEL II
PERBANDINGAN JUMLAH *CLASS DATASET*

	Positif	Netral	Negatif	Jumlah
Peneliti	2.067	6.004	807	8.878
Ahli Bahasa Indonesia	1.452	6.716	710	8.878

Berikut adalah contoh *dataset* yang telah divalidasi oleh ahli Bahasa Indonesia dapat dilihat pada TABEL III.

TABEL III
CONTOH *DATASET* YANG TELAH DIVALIDASI

No	Teks	*Class*
1	makasih tokopedia karena diskon makeup cuma jadi 49,5 kalo 50k udah w ambil karena gratis ongkir wkwk	positif
2	karena iklan tokped gw jadi terngiang lagunya BTS, thanks tokped <U+0001F642><U+0001F64F><U+0001F3FB>	positif
3	@CiaClarissa2 <U+261D><U+0001F3FB> saya setiap liat iklan tokped kalo lagi nonton bioskop	netral
4	@TokopediaCare Kak mohon segera balas ya. Aku sudah mengajukan pengaduan juga di https://t.co/2kgsO625tO	netral
5	@TokopediaCare pesanan sudah 5 hari statusnya diproses terus tapi belum dikirim, apakah bisa dicancel?	negatif
6	@TokopediaCare Menyerah saya min, cmn buang2 waktu aja mau masukin barang ke etalase dari pagi ga bisa, udah instal ulang applikasi tetep ga bisa. https://t.co/oEhMjM1rzu	negatif

B. Text Preprocessing

Data yang telah melalui proses validasi oleh Ahli Bahasa Indonesia selanjutnya melalui tahap *text* preprocessing. Tahap ini bertujuan untuk menghilangkan atribut yang kurang berpengaruh terhadap hasil klasifikasi. Di bawah ini adalah tahap-tahap dari *text preprocessing*. Berikut ini merupakan hasil dari tahap *text preprocessing* dapat dilihat pada TABEL IV.

TABEL IV
CONTOH *TEXT PREPROCESSING*

	Sebelum	Sesudah
Cleaning URL	Di tokped banyak mb https://t.co/9CKQ4PnjRr	Di tokped banyak mb
Cleaning Username	Min @TokopediaCare moh on dibalas DM urgent	Min mohon dibalas DM urgent
Cleaning hashtag	Mantul min #TokopediaxBTS bismillah	Mantul min bismillah
Cleaning Number	min kode verifikasi udah 2 jam ga d kriim2 ga bs login nih	min kode verifikasi udah jam ga d kriim ga bs login ini
Cleaning Emoticon	ke bandung dong miin <U+FD><U+FD>	ke bandung dong miin
Cleaning Punctuation	Gak ada DM masuk :)	Gak ada DM masuk
Cleaning New Line	Tokopedia Youtube Channel *Update* \n\nTokopediaxBTS Belanja Kebutuhan Kecantikan Tokopedia Saja Bebas Ongkir \n\n	Tokopedia Youtube Channel *Update* TokopediaxBTS Belanja Kebutuhan Kecantikan Tokopedia Saja Bebas Ongkir
Cleaning Non ASCII	Promo Valentine LUMIX Periode Februari For order and detail Kramat Gantung WA Sumber Bahagia juga tersedia di berbagai MarketPlace Tokopedia â€¦	Promo Valentine LUMIX Periode Februari For order and detail Kramat Gantung WA Sumber Bahagia juga tersedia di berbagai MarketPlace Tokopedia a
Case Folding	GILASEHHHHH TOKO PEDIA KERENNNN NJIRRRR	gilasehhhhh toko pedia kerennnn njirrrr
Tokenizing	halo saya kok gabisa cek saldo toko saya ya	“halo” “saya” “kok” “gabis”a “cek” “saldo” “toko” “saya” “ya”
Steeming	min tolong cek dm ada pertanyaan lagi dari saya thanks	min tolong cek dm ada tanya lagi dari saya thanks
Hapus Duplikasi Huruf	aku mauuuu dong mimin	aku mau dong mimin
Stopword Removal	kartu kredit kredit saya kena hack sebesar juta pembelian dari tokopedia mau lapor biar cancel gimana cara nya ya tokped tidak ada customer service yg bs di telp	kartu kredit kredit kena hack juta beli tokopedia lapor biar cancel gimana nya ya tokped customer service yg bs telp
Remove Whitespace	halo kali cek data transaksi salah ulang barusan *update* aplikasi tetep	halo kali cek data transaksi salah ulang barusan update aplikasi tetep

C. Transformation

Data yang telah melalui tahap *text preprocessing* selanjutnya dilakukan transformasi data menjadi *Document Term Matrix* (DTM). DTM merupakan sebuah matriks berukuran n × m dengan dokumen sebagai baris dan kata (*term*) sebagai kolom. Pembentukan DTM dilakukan sekaligus dengan pembobotan kata menggunakan *Term-Frequency-Invers Document Frequency* (TF-IDF). Berikut ini merupakan hasil dari *Term-Frequency-Invers Document Frequency* (TF-IDF). Berikut adalah hasil dari TF-IDF.

```
<<DocumentTermMatrix (documents: 8587, terms: 11302)>>
Non-/sparse entries: 65749/96984525
Sparsity           : 100%
Maximal term length: 43
Weighting          : term frequency - inverse document frequency (tf-idf)
Sample             :
     Terms
Docs   ambil     beli cashback juta klik kupon link min tokopedia  tokped
  1669     0 0.000000        0    0    0     0    0   0  0.000000 2.67777
  3097     0 0.000000        0    0    0     0    0   0  0.000000 0.00000
  3516     0 0.000000        0    0    0     0    0   0  1.718104 0.00000
  3631     0 0.000000        0    0    0     0    0   0  0.000000 2.67777
  3712     0 0.000000        0    0    0     0    0   0  0.000000 2.67777
  4177     0 0.000000        0    0    0     0    0   0  3.436209 0.00000
  4513     0 0.000000        0    0    0     0    0   0  1.718104 0.00000
  5219     0 3.576085        0    0    0     0    0   0  0.000000 2.67777
  6445     0 0.000000        0    0    0     0    0   0  1.718104 0.00000
  7906     0 0.000000        0    0    0     0    0   0  0.000000 0.00000
```

Gambar 7. Hasil Transformasi dengan TF-IDF

Gambar 7 merupakan hasil dari transformasi data yang menghasilkan matriks berukuran 8.587 × 11.302 atau 8.587 dokumen dan 11.302 *term*. Dalam matriks tersebut terdapat 96.984.525 sel berisi nol dan 65.749 sel berisi bukan nol, serta 100% mayoritas isi sel-sel di dalam matriks adalah nol. Nilai nol tersebut merepresentasikan *term* yang jarang muncul pada *Document Term Matrix*, untuk mengurangi hal tersebut dilakukan teknik komputasi bernama *RemoveSparseTerm*. Nilai kejarangan diatur sebesar 99%, artinya 99% *term* yang berisi nol di dalam matriks akan dihapus.

Hasil dari *RemoveSparseTerm* menjadi 748.098 sel berisi nol dan 24.732 sel berisi tidak nol, serta 90 *term* pada dokumen yang sering muncul. Berikut adalah 90 *term* pada Document *Term* Matrix yang ditunjukkan pada TABEL V.

TABEL V
DOCUMENT TERM MATRIX

Docs	***Term***						
	army	**min**	**...**	**harga**	**tokopedia**	**...**	**via**
1	0	0	...	0	0	...	0
2	6.608507	0	...	0	0	...	0
6	0	0	...	4.604414	1.718104	...	0
...	...	...	...	...	...	...	0
4441	6.608507	0	...	4.604414	0	...	0
8587	0	0	...	0	0	...	0

D. Feature Selection

Feature Selection berfungsi untuk meningkatkan hasil evaluasi pada tahap *data mining*. *Feature selection* yang digunakan pada penelitian ini yaitu dengan menggunakan algoritma *Particle Swarm Optimization*. Paramater yang digunakan pada algoritma *Particle Swarm Optimization* yaitu 10, 20, 30, 40, dan 50 iterasi dengan masing-masing partikel yaitu 10 dan 20. Proses pengerjaan *feature selection* akan dikerjakan pada tahap *data mining* yaitu setelah menemukan skenario dengan nilai akurasi tertinggi. Nilai ini akan digunakan sebagai *fitness function* pada algoritma *Particle Swarm Optimization*.

E. Data Mining

Tahap *data mining* terdiri dari dua pemodelan, yaitu dengan algoritma *K-Nearest Neighbor* tanpa seleksi fitur dan algoritma *K-Nearest Neighbor* dengan seleksi fitur algoritma *Particle Swarm Optimization*. Untuk pemodelan pertama yaitu dengan algoritma *K-Nearest Neighbor* tanpa seleksi fitur akan dilakukan empat skenario dengan perbandingan data latih dan data uji yaitu 60:40, 70:30, 80:20, dan 90:10 dengan teknik *random percentage split*, serta dengan nilai k yang digunakan yaitu 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, dan 15. Berikut adalah jumlah perbandingan antara data latih dan data uji dapat dilihat pada TABEL VI.

TABEL VI
JUMLAH PERBANDINGAN *DATASET*

Persentase	**Data Latih**	**Data Uji**
60% dan 40%	5.153	3.434
70% dan 30%	6.011	2.576
80% dan 20%	6.870	1.717
90% dan 10%	7.729	858

Setelah dilakukan pengujian dengan skenario yang telah ditentukan, didapatkan skenario dengan akurasi tertinggi yaitu skenario 4 dengan pembagian *dataset* 90:10 dengan teknik *random percentage split* dengan nilai *k* = 1 sebesar 88.11%. TABEL VII menunjukkan nilai akurasi dari masing-masing skenario.

TABEL VII
SKENARIO DENGAN AKURASI TERBAIK

Skenario	***k* terbaik**	**Akurasi**
Skenario 1	1	82.62%
Skenario 2	1	85.29%
Skenario 3	1	85.85%
Skenario 4	1	88.11%

Berikut adalah hasil dari *confusion matrix* untuk skenario 4 dengan nilai k = 1 seperti pada Gambar 8.

```
                Reference
Prediction negatif netral positif
   negatif       52      6       1
   netral        15    616      45
   positif        1     34      88

Overall Statistics

               Accuracy : 0.8811
```

Gambar 8. *Confusion Matrix* skenario 4 dengan nilai k = 1

Berdasarkan Gambar 8 dapat dilihat bahwa 52 kelas negatif benar diklasifikasikan sebagai kelas negatif, sementara itu 15 kelas negatif diklasifikasikan sebagai kelas

netral dan 1 sebagai kelas positif. Sedangkan 616 kelas netral benar diklasifikasikan sebagai kelas netral, sementara itu 6 kelas netral diklasifikasikan sebagai kelas negatif dan 34 sebagai kelas positif. Sedangkan untuk kelas positif terdapat 88 kelas positif benar diklasifikasikan sebagai positif, 1 kelas positif diklasifikasikan sebagai kelas negatif dan 45 kelas positif diklasifikasikan sebagai kelas netral. Akurasi yang dihasilkan sebesar 0.8811 atau 88.11%.

Selanjutnya nilai akurasi ini akan dijadikan nilai *fitness* pada *Particle Swarm Optimization*. Setelah dilakukan percobaan, didapatkan bahwa dengan 20 iterasi dan 10 partikel menghasilkan akurasi tertinggi. Gambar 9 merupakan hasil *confusion matrix* setelah diterapkan algoritma *Particle Swarm Optimization*.

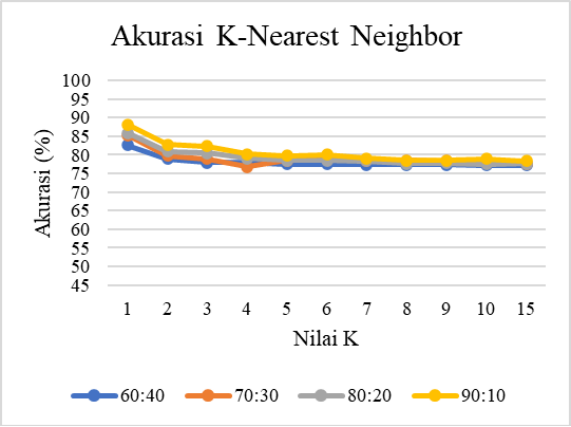
```
                Reference
Prediction negatif netral positif
   negatif       65      3       1
   netral         3    646       7
   positif        0      7     126

Overall Statistics

               Accuracy : 0.9755
```

Gambar 9. *Confusion Matrix* setelah fitur seleksi

F. Evaluation

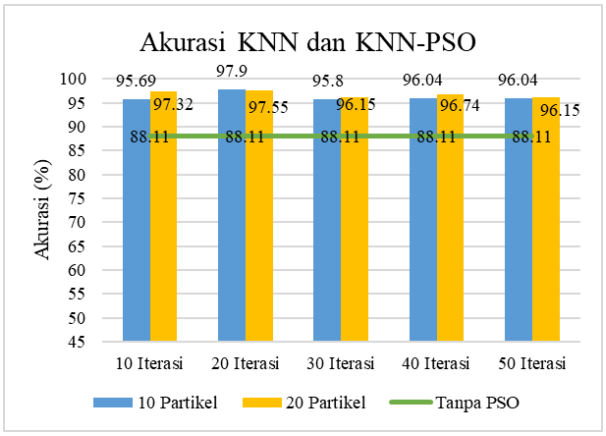
Setelah dilakukan tahap *data mining*, tahap selanjutnya yaitu evaluation untuk mengukur kinerja dari algoritma *K-Nearest Neighbor* dan *K-Nearest Neighbor* dengan seleksi fitur *Particle Swarm Optimization*. Skala pada grafik yaitu antara 45%-100%, hal ini bertujuan agar grafik terlihat lebih jelas.

1. *Akurasi:* Berikut merupakan nilai akurasi dari hasil *data mining* dengan menggunakan algoritma *K-Nearest Neighbor*.



Gambar 10. Perbandingan akurasi *K-Nearest Neighbor*

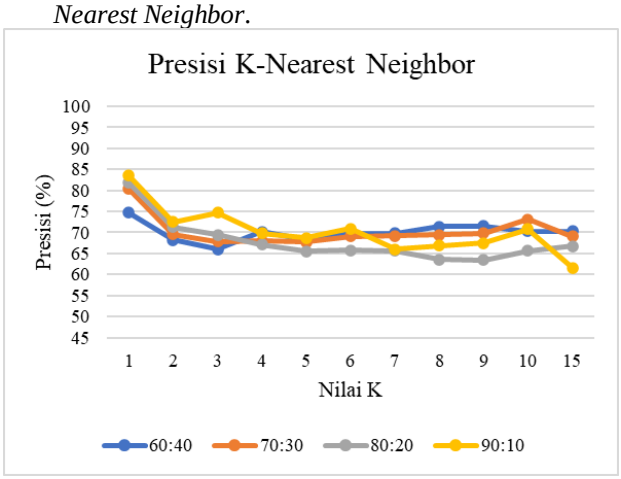
Gambar 10 menunjukkan bahwa semakin besar nilai k akurasinya cenderung menurun. Selain itu, dari nilai *k* = 7 sampai *k* = 15 tidak terdapat perubahan akurasi secara signifikan.



Gambar 11. Akurasi KNN dan KNN dengan PSO

Berdasarkan Gambar 11, akurasi tertinggi Algoritma *K-Nearest Neighbor* yaitu 88.11%. Setelah diterapkan seleksi fitur dengan algoritma *Particle Swarm Optimization*, nilai akurasi meningkat dengan tajam. Akurasi tertinggi yaitu dengan parameter 20 iterasi dan 10 partikel sebesar 97.9%.

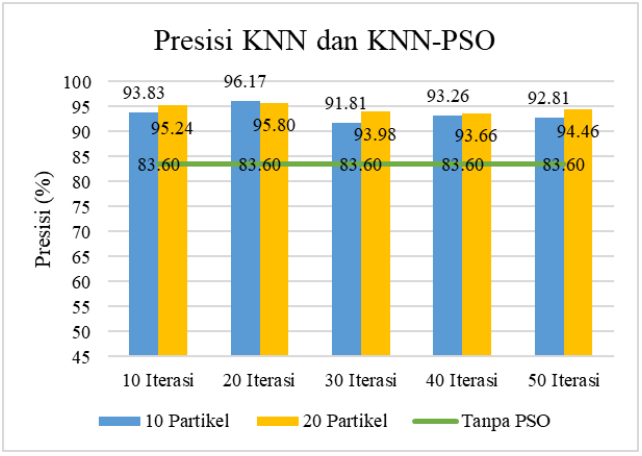
2. *Presisi:* Berikut merupakan nilai presisi dari hasil *data mining* dengan menggunakan algoritma *K-Nearest Neighbor*.



Gambar 12. Presisi *K-Nearest Neighbor*

Gambar 12 menunjukkan bahwa dengan k = 1 nilai presisinya lebih tinggi dibandingkan dengan k lainnya. Berikut adalah grafik perbandingan presisi *K-Nearest Neighbor* dan *K-Nearest Neighbor* dengan *Particle Swarm Optimization*.

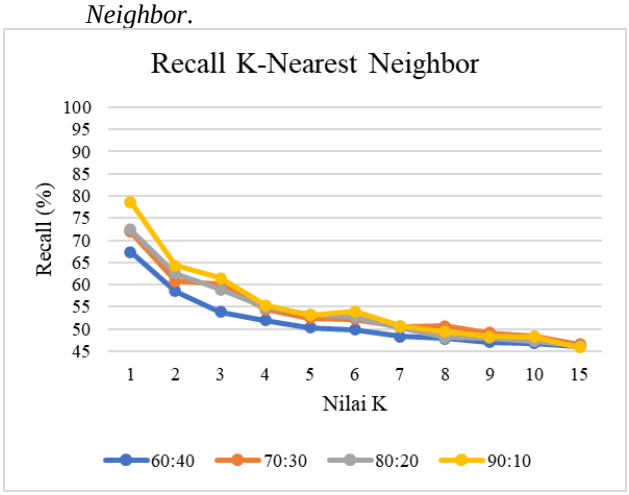
p-ISSN : 2443-2210
e-ISSN : 2443-2229



Gambar 13. Presisi KNN dan KNN dengan PSO

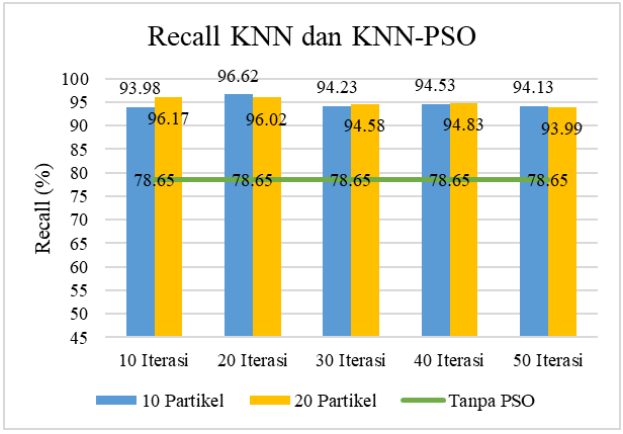
Berdasarkan Gambar 13, akurasi tertinggi Algoritma *K-Nearest Neighbor* yaitu 88.11%. Setelah diterapkan seleksi fitur dengan algoritma *Particle Swarm Optimization*, nilai akurasi meningkat dengan tajam. Akurasi tertinggi yaitu dengan parameter 20 iterasi dan 10 partikel sebesar 97.9%.

3. *Recall:* Berikut merupakan nilai *recall* dari hasil *data mining* dengan menggunakan algoritma *K-Nearest Neighbor*.



Gambar 14. *Recall K-Nearest Neighbor*

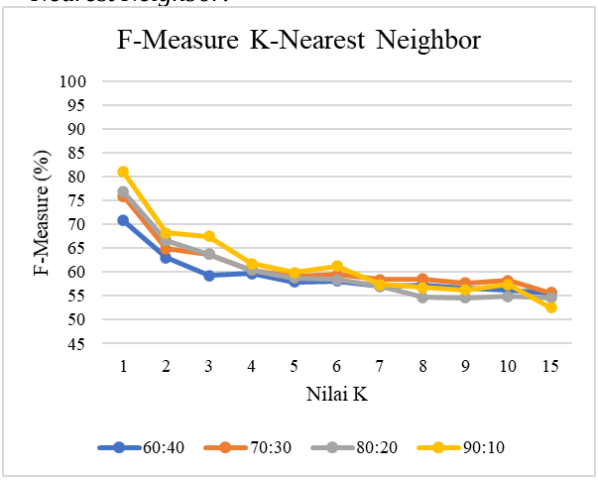
Gambar 14 menunjukkan bahwa semakin besar nilai k nilai *recall* cenderung menurun. Berikut adalah grafik perbandingan recall *K-Nearest Neighbor* dan *K-Nearest Neighbor* dengan *Particle Swarm Optimization.*



Gambar 15. *Recall* KNN dan KNN dengan PSO

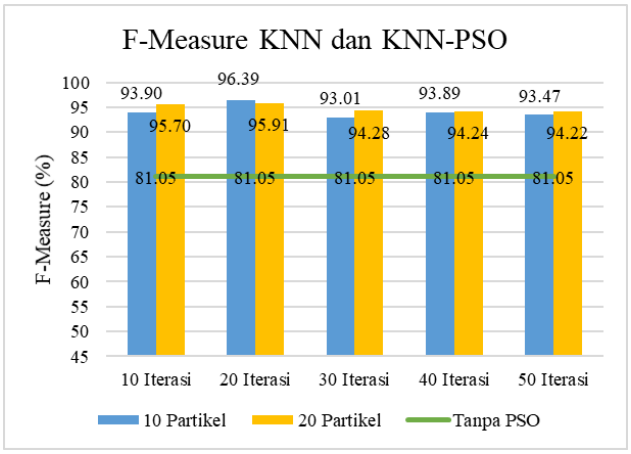
Berdasarkan Gambar 15, *recall* tertinggi Algoritma *K-Nearest Neighbor* yaitu 78.65%. Setelah diterapkan seleksi fitur dengan Algoritma *Particle Swarm Optimization, recall* meningkat dengan tajam. *Recall* tertinggi yaitu dengan parameter 20 iterasi dan 10 partikel sebesar 96.62%.

4. *F-measure:* Berikut merupakan nilai *f-measure* dari hasil *data mining* dengan menggunakan algoritma *K-Nearest Neighbor*.



Gambar 16. *F-measure K-Nearest Neighbor*

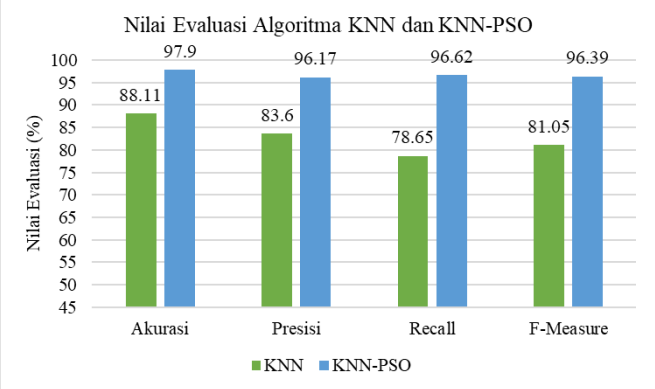
Gambar 16 menunjukkan bahwa semakin besar nilai k nilai *f-measure* cenderung menurun. Berikut adalah grafik perbandingan *f-measure K-Nearest Neighbor* dan *K-Nearest Neighbor* dengan *Particle Swarm Optimization.*



Gambar 17. *F-measure* KNN dan KNN dengan PSO

Berdasarkan Gambar 17, *f-measure* algoritma *K-Nearest Neighbor* yaitu 81.05%. Setelah diterapkan seleksi fitur dengan algoritma *Particle Swarm Optimization*, *f-measure* meningkat dengan tajam. *F-measure* tertinggi yaitu dengan parameter 20 iterasi dan 10 partikel sebesar 96.39%.

Gambar 18 adalah grafik perbandingan nilai evaluasi pada skenario 90:10 dengan nilai k = 1 antara algoritma *K-Nearest Neighbor* dan *K-Nearest Neighbor* dengan algoritma *Particle Swarm Optimization* dengan 20 iterasi dan 10 partikel.



Gambar 18.Perbandingan nilai evaluasi algoritma KNN dan KNN-PSO

Gambar 18 menunjukkan bahwa dengan menggunakan algoritma *Particle Swarm Optimization* dapat meningkatkan nilai evaluasi dari algoritma *K-Nearest Neighbor*.

Penelitian *K-Nearest Neighbor* Berbasis *Particle Swarm Optimization* untuk Analisis Sentimen Terhadap Tokopedia. Metodologi yang digunakan *adalah Knowledge Discovey in Database* (KDD) yang terdiri dari *text, text preprocessing, transformation, feature selection, data mining,* dan *evaluation.* Teknik pengumpulan data yang digunakan yaitu *crawling data* dengan *tool* Rstudio. Data yang berhasil dikumpulkan sebanyak 17.298 data. *Dataset* kemudian dilakukan tahap penyeleksian sehingga menjadi 8.588 data lalu dilakukan pemberian *class*. Selanjutnya dilakukan tahap *text preprocessing* untuk membersihkan *dataset* dari atribut yang mengganggu. Setelah itu dilakukan transformasi data dengan pembobotan TF-IDF dan *Remove Sparse Term*. Algoritma yang digunakan pada tahap *data mining* yaitu *K-Nearest Neighbor*. Pada tahap *data mining* dilakukan empat skenario dengan nilai *k* yang berbeda-beda, kemudian diambil skenario dengan nilai akurasi tertinggi untuk dilakukan seleksi fitur dengan algoritma *Particle Swarm Optimzation*. Tahap terakhir yaitu evaluasi dengan *confusion matrix* untuk melihat performa dari algoritma yang digunakan.

Berdasarkan empat skenario yang dilakukan, skenario dengan pembagian *dataset* 90% sebagai data latih dan 10% sebagai data uji dengan teknik *random percentage split* dan nilai k yang digunakan yaitu 1 menghasilkan akurasi tertinggi yaitu 88.11%. Skenario ini kemudian digunakan untuk dilakukan tahap seleksi fitur dengan algoritma *Particle Swarm Optimization*. Parameter yang digunakan yaitu 20 iterasi dan 10 partikel menghasilkan akurasi terbaik sebesar 97.9% dibandingkan parameter lain. Selain nilai akurasi, nilai evaluasi lainnya pun mengalami kenaikan yaitu presisi dari 83.60% menjadi 96.17%, *recall* dari 78.65% menjadi 96.62% dan *f-measure* dari 81.05% menjadi 96.39%.

Gambar 19. Visualisasi *Word Cloud*

Gambar 19 menunjukkan *word cloud* yaitu visualiasi dari kata-kata yang sering muncul pada setiap kelas sentimen. Kata "kupon", "*cashback*", dan "bagi-bagi" menjadi kata bersifat positif yang sering digunakan oleh pengguna Twitter untuk menunjukkan bahwa Tokopedia sedang membagikan kupon cashback bagi para pengguna. Selain itu juga ada kata "bebas", "ongkir", "bts", dan "iklan" yang mana hal ini berkaitan dengan iklan Tokopedia yang dibintangi oleh *boygroup* asal Korea Selatan yaitu BTS mendapat respon yang positif dikalangan pengguna Twitter. Pada sentimen netral, kata "min" yang merupakan singkatan dari kata "admin" mendominasi *word cloud*. Pada sentimen negatif, kata "barang" mendominasi dibandingkan kata-kata lainnya. Kata "komplain" dan "gagal" menunjukkan sentimen negatif dari pengguna Twitter terhadap Tokopedia.

VI. KESIMPULAN

Pada penelitian ini digunakan algoritma *K-Nearest Neighbor* untuk menganalisis sentimen terhadap Tokopedia dengan menggunakan metodologi *Knowledge Discovery in Database* (KDD) dengan tahapan *text, text preprocessing, transformation, feature selection, data mining,* dan

evaluation. Algoritma *Particle Swarm Optimization* digunakan untuk tahapan *feature selection* dengan melihat skenario terbaik dari tahap *data mining.*

Kinerja algoritma *K-Nearest Neighbor* dalam menganalisis sentimen terhadap Tokopedia dievaluasi dengan menggunakan *confusion matrix.* Akurasi terbaik yaitu pada skenario 90:10 dengan nilai *k* = 1 yaitu sebesar 88.11%. Hal ini disebabkan karena dengan menggunakan *k* = 1 perbedaan jarak antara data latih dan data uji tidak terlalu jauh jika dibandingkan dengan nilai *k* lainnya. Selanjutnya dilakukan *feature Selection* dengan menggunakan algoritma *Particle Swarm Optimization* dengan parameter 20 iterasi dan 10 partikel menghasilkan nilai evaluasi terbaik yaitu akurasi sebesar 97.9%, presisi sebesar 96.17%, *recall* sebesar 96.62% dan *f-measure* sebesar 96.39%.

DAFTAR PUSTAKA

[1] Y. Pratomo. (2019) APJII: Jumlah pengguna internet di indonesia tembus 171 juta jiwa. [Online]. Tersedia: https://tekno.kompas.com/read/2019/05/16/03260037/apjii-jumlah-pengguna-internet-di-indonesia-tembus-171-juta-jiwa

[2] H. Widowati. (2019) Indonesia jadi negara dengan pertumbuhan e-commerce tercepat di dunia. [Online]. Tersedia: https://databoks.katadata.co.id/datapublish/2019/04/25/indonesia-jadi-negara-dengan-pertumbuhan-e-commerce-tercepat-di-dunia.

[3] W. Pusparisa. (2019) 96% pengguna internet di indonesia pernah menggunakan e-commerce. [Online]. Tersedia: https://databoks.katadata.co.id/datapublish/2019/12/03/96-pengguna-internet-di-indonesia-pernah-gunakan-e-commerce

[4] D. H. Jayani. (2019) Tren pengguna e-commerce terus tumbuh. [Online]. Tersedia: https://databoks.katadata.co.id/datapublish/2019/10/10/tren-pengguna-e-commerce-2017-2023

[5] D. H. Jayani. (2019). 10 e-commerce dengan pengunjung terbesar kuartal iii-2019. [Online]. Tersedia: https://databoks.katadata.co.id/datapublish/2019/10/2/inilah-10-e-commerce-dengan-pengunjung-terbesar

[6] D. Adrianyah. (2019) 10 top brand di twitter di indonesia 2019. [Online]. Tersedia: https://blog.twitter.com/in_id/topics/insights/2019/10-top-brand-di-Twitter-di-Indonesia.html

[7] K. Gilang. (2019) Memahami cara kerja tim customer experiance tokopedia. [Online]. Tersedia: https://id.techinasia.com/customer-experience-tokopedia

[8] M. Allahyari *et al.*, "A brief survey of text mining: classification, clustering and extraction techniques," *Proceedings of KDD Bigdas*, 2017, arXiv preprint:1707.02919, pp. 1-13.

[9] U. Rofiqoh, R. S. Perdana, & M. A. Fauzi, "Analisis sentimen tingkat kepuasan pengguna penyedia layanan telekomunikasi seluler indonesia pada twitter dengan metode support vector machine dan lexion based feature," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Kompututer*, vol. 1, no. 12, pp. 1725–1732, Desember 2017.

[10] A. M. Zuhdi, E. Utami, & S. Raharjo, "Analisis sentiment twitter terhadap capres indonesia 2019 dengan metode k-nn," *Jurnal Informa Politek Indonusa Surakarta*, vol. 5, no. 2, pp. 1–7, 2019.

[11] L. D. Utami, "Komparasi algoritma klasifikasi pada analisis review hotel," *Jurnal Pilar Nusa Mandiri*, vol. 14, no. 2, pp. 261–266, 2018.

[12] M. W. Pertiwi, "Analisis sentimen opini publik mengenai sarana dan transportasi mudik tahun 2019 pada twitter menggunakan algoritma naïve bayes, neural network, knn dan svm," *Inti Nusa Mandiri*, vol. 14, no. 1, pp. 27–32, 2019.

[13] D. A. Kristiyanti, "Analisis sentimen review produk kosmetik melalui komparasi feature selection," *Konferensi Nasional Ilmu Pengetahuan dan Teknologi (KNIT) 2015*, 2015, pp. 69-76.

[14] Riswanti, I. Budiman, & A. R. Arrahimi, "Sentiment analysis svm dan svm-pso pada kolom komentar evaluasi dosen," *Seminar Nasional Ilmu Komputer (SOLITER)*, 2019, pp. 110-119.

[15] M. A. Maulana, A. Setyanto, & M. P. Kurniawan, "Analisis sentimen media sosial universitas amikom," *Seminar Nasional Teknologi Informasi dan Multimedia 2018*, 2018, pp. 55-59.

[16] A. Zanuardi & H. Suprayitno, "Analisa karakteristik kecelakaan lalu lintas di jalan ahmad yani surabaya melalui pendekatan knowledge discovery in database," *Jurnal Manajemen Aset Infrastruktur & Fasilitas*, vol. 2, no. 1, pp. 45–55, Maret 2018.

[17] S. Widaningsih & A. Suheri, "Klasifikasi jurnal ilmu komputer berdasarkan pembagian web of service dengan menggunakan text mining," *Seminar Nasional Teknologi Informasi dan Komunikasi 2018 (SENTIKA 2018)*, 2018, pp. 320-328.

[18] G. Ngurah, M. Nata, & P. P. Yudiastra, "Knowledge discovery pada email box sebagai penunjang email marketing knowledge discovery in the email box for support email marketing," *Jurnal Sistem dan Informatika*, pp. 26–37, 2017.

[19] W. Gata & Purnomo, "Akurasi text mining menggunakan algoritma k-nearest neighbour pada data content berita sms," *Jurnal Format*, vol. 6, no. 1, pp. 1–13, 2017.

[20] Y. I. Claudy, R. S. Perdana, & M. A. Fauzi, "Klasifikasi dokumen twitter untuk mengetahui karakter calon karyawan menggunakan algoritme k-nearest neighbor (KNN)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 8, pp. 2761–2765, Agustus 2018.

[21] F. Pramono, D. Rosiyadi, & W. Gata, "Integrasi n-gram, information gain, particle swarm optimation di naïve bayes untuk optimasi sentimen google classroom," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 3, pp. 383–388, 2019.