Klasifikasi Cerita Bahasa Indonesia menggunakan Metode *Hybrid* PSO-KNN (*Modified Binary Particle Swarm Optimization* dengan *K-Nearest Neighbor*)

Anita Rahayu[1], Naim Rochmawati[2],

[1] Jurusan Teknik Informatika/Teknik Informatika, Universitas Negeri Surabaya
[2] Jurusan Teknik Informatika/Teknik Informatika, Universitas Negeri Surabaya
[1]rahayuanita45681@gmail.com
[2]naimrochmawati@unesa.ac.id

***Abstrak*— Penentuan kategori suatu cerita merupakan hal yang penting agar cerita yang dibaca sesuai dengan keinginan pembaca. Selama ini proses penentuan kategori suatu cerita masih dilakukan secara manual sehingga perlu adanya pengklasifikasian kategori cerita secara otomatis. Metode klasifikasi atau kategorisasi teks merupakan proses yang secara otomatis meletakkan dokumen teks ke dalam suatu kategori berdasarkan isi dari teks tersebut. Pada penelitian ini peneliti mengusulkan sebuah metode hybrid PSO-KNN yaitu penggabungan metode Modified Binary Particle Swarm Optimization dengan K-Nearest Neighbor. Metode PSO-KNN akan mengatasi permasalahan pengklasifikasian teks sekaligus mengatasi kelemahan KNN yang menggunakan seluruh fitur saat proses pembentukan model (learning). PSO-KNN akan mengurangi dimensi dari dokumen dengan memilih token-token sebagai fitur yang paling baik namun isi yang dikandung dokumen tetap terjaga karena fitur yang dipilih sangat merepresentasikan dokumen tersebut. Penerapan metode PSO-KNN berhasil mengkategorikan 5 kategori cerita Bahasa Indonesia sebanyak 150 data dengan tingkat akurasi sebesar 53% dan total fitur optimal sebanyak 88 fitur. Berdasarkan hasil penelitian yang dilakukan dapat disimpulkan bahwa metode PSO-KNN berhasil melakukan pengklasifikasi kategori cerita pendek serta mengurangi fitur saat proses pembentukan model dan meningkatkan nilai akurasi.**

Kata Kunci*— Klasifikasi teks, cerita pendek, *modified binary particle swarm optimization, k-nearest neighbor, loocv

I. PENDAHULUAN

Penentuan kategori suatu cerita merupakan hal yang penting agar cerita yang dibaca sesuai dengan keinginan pembaca [1]. Namun selama ini proses pengklasifikasi kategori cerita ini masih dilakukan secara manual yakni dengan membaca keseluruhan isi cerita kemudian menentukannya masuk dalam kategori apa. Dengan perkembangan teknologi saat ini yang semakin pesat menuntut adanya penyampaian informasi secara tepat dan cepat sehingga dibutuhkan teknik tertentu untuk mengolah dokumen teks tersebut sehingga dapat melakukan pengklasifikasian cerita secara otomatis Untuk mengatasi masalah ini bisa dilakukan dengan memanfaatkan salah satu cabang text mining yakni klasifikasi.

Klasifikasi teks atau kategorisasi teks merupakan proses yang secara otomatis meletakkan dokumen teks ke dalam suatu kategori berdasarkan isi dari teks tersebut. Salah satu metode klasifikasi teks yakni metode KNN. KNN ini sangat umum digunakan dalam pengkategorian teks karena algoritmanya yang mudah dan efisien untuk klasifikasi teks [2]. Namun metode KNN ini memiliki kelemahan yakni beratnya komputasi karena KNN ini menggunakan seluruh fitur yang diperlukan untuk perhitungan jarak.

Penelitian mengenai klasifikasi teks sebelumnya telah dilakukan oleh Oman Somantri dan Mohammad Khambali [3] dengan judul "*Feature Selection* Klasifikasi Kategori Cerita Pendek menggunakan Naïve Bayes dan Algoritme Genetika" dari penelitian ini GA terbukti dapat meningkatkan akurasi klasifikasi dengan *feature selection*. Hasil akurasi akhir yang dihasilkan yakni 84,29%. Penelitian lain mengenai klasifikasi teks juga dilakukan oleh Vishwanath Bijalwan, et.al [4] dengan judul "*Machine Learning Approach for Text and Document Mining*" yang menggunakan metode KNN untuk mengklasifikasikan artikel ke dalam lima kategori. Hasil penelitian ini menunjukkan bahwa metode KNN memiliki nilai akurasi tertinggi jika dibandingkan dengan metode naïves bayes dan term-graph. Selanjutnya penelitian mengenai seleksi fitur dilakukan oleh Sheba K.U [5] dengan judul "*A Modified Binary PSO based Feature Selection for Automatic Lesion Detection in Mammograms*" yang mana penelitian ini menggunakan PSO yang telah dimodifikasi yakni *Modified Binary Particle Swarm Optimization* (MBPSO) untuk mencari subset fitur yang optimal dan dievaluasi menggunakan KNN. Dimana metode PSO-KNN yang diusulkan berhasil mengurangi kompleksitas komputasional dan mampu menunjukkan efisiensi yang tinggi dbandingkan metode seleksi fitur lainnya dengan tingkat akurasi mencapai 97,2% dengan jumlah fitur optimal sebanyak 6 yang sebelumnya jumlah fitur asli sebanyak 117 dengan tingkat akurasi 87,3%.

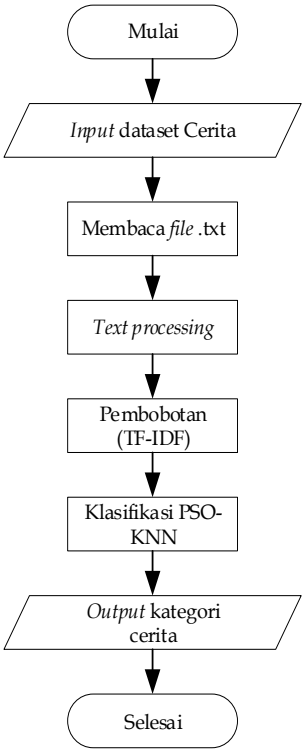
Dari beberapa penelitian yang telah dilakukan belum ada yang menggunakan metode gabungan KNN dengan MBPSO dalam pengklasifikasian teks, sehingga penelitian yang akan dilakukan akan menggunakan PSO-KNN (*Modified Binary Particle Swarm Optimization* dengan *K-Neareset Neighbor*) sebagai metode pengklasifikasian teks cerita sehingga akan menghasilkan sistem klasifikasi cerita Bahasa Indonesia dengan nilai akurasi yang lebih baik dan menghasilkan fitur yang paling optimal.

II. METODE PENELITIAN

Dalam penelitian ini terdapat beberapa tahapan penelitian yakni meliputi identifikasi, studi literatur, pengumpulan data, perancangan penelitian, implemetasi dan pengujian, analisis

hasil serta kesimpulan. Alur jalannya penelitian dpat dilihat pada Gbr. 1.

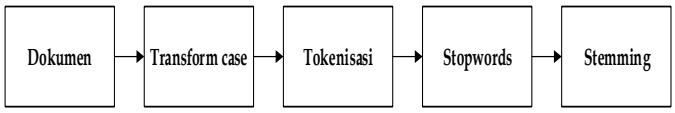
Tujuan dari penelitian ini yaitu untuk menerapkan salah satu metode dari data mining klasifikasi yaitu *K-Nearest Neighbor* yang digabung dengan metode *Modified Binary Particle Swarm Optimization* yang merupakan modifikasi dari PSO tradisional sebagai dasar untuk mengetahui kategori suatu cerita. Alur pengklasifikasian PSO-KNN dapat dilihat pada Gbr. 1.



Gbr. 1 Alur jalannya penelitian

A. *Text Processing*

Text processing merupakan suatu proses untuk menyiapkan data sehingga data akan menjadi data terstruktur dan dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut. Dalam proses text processing ada beberapa tahapan yakni sebagai berikut :



Gbr. 2 Alur proses *text processing*

1. *Transform case* atau c*ase folding* adalah mengubah semua kata yang ada pada dokumen menjadi huruf kecil.
2. Tokenisasi adalah proses pemenggalan kata dalam suatu dokumen.
3. *Stopwords removal* atau *filtering* adalah proses menghilangkan kata yang tidak penting.
4. *Stemming* adalah proses mengubah kata menjadi bentuk kata dasar. Dalam penelitian ini peneliti mengggunakan *Sastrawi Stemmer* yang merupakan *library stemmer* yang disediakan python yang berbasis algoritma Nazief dan Adriani yang ditingkatkan lagi menjadi *Modified ECS* (*Enhanced Confix Striping*).

B. *Pembobotan TF-IDF*

Pembobotan adalah proses pemberian bobot pada kata yang ada pada setiap dokumen berdasarkan jumlah kemunculan kata atau token dalam dokumen tersebut. Berikut adalah penjelasan proses pembobotan TF-IDF :

1. Hitung frekuensi tiap term dalam dokumen yang disebut tf dengan formula TF murni (raw TF) yakni nilai tf diberikan berdasarkan jumlah kemunculan suatu term dalam dokumen.
2. Hitung frekuensi dokumen yang mengandung sebuah term tr pada setiap term yang disebut df.
3. Hitung nilai idf untuk setiap term mengacu pada (1)
$$IDF_{(t)} = log\left(\frac{N}{nt}\right) \quad (1)$$
4. Hitung bobot TFIDF dengan mengalikan frekuensi term tf dengan nilai idf pada setiap dokumen mengacu pada (2)
$$TF - IDF = f_{(t,d)} * log\left(\frac{N}{nt}\right) \quad (2)$$

Maka dihasilkan bobot yang akan dijadikan vektor fitur tiap dokumen berupa matriks.

C. *PSO-KNN*

PSO-KNN merupakan gabungan antara metode *Modified Binary Particle Swarm Optimization* dengan *K-Nearest*, berikut adalah proses penentuan fitur menggunakan PSO-KNN :

1. Masukkan dataset pelatihan yang telah mengalami preprocessing dan pembobotan.
2. Inisialisasi learning rate c1=c2=1,49618 dan c3=0,5, Vmax=6 dan Vmin=-6, ωmax = 0,995 dan ωmin=0,5 (mengacu pada jurnal), jumlah fitur (D), total fitur optimal (d), total iterasi (T) dan jumlah partikel (n).
3. Bangkitkan n partikel awal secara random yang merupakan vektor biner dengan panjang vektor D lalu jumlah total biner bernilai 1 pada setiap vektor adalah d. Jadikan ini sebagai posisi awal n partikel (Xi) dimana Xij berupa 0 atau 1.
4. Inisialisasi kecepatan awal (Vi) n dengan 0.
5. Inisialisasi Pbest, Itbest, dan Gbest dengan 0.
6. Hitung nilai fitness tiap partikel dengan KNN-LOOCV.
7. Untuk setiap partikel pada iterasi selanjutnya hingga iterasi maksimum lakukan langkah berikut :
   a. Perbarui nilai ω dengan mengacu pada (3)
   $$\omega = \omega_{max} - \frac{(\omega_{max} - \omega_{min})}{iter\ max} \times iter \quad (3)$$
   b. Bangkitkan δ dengan nilai random antara 0-1
   c. Perbarui kecepatan partikel Vij mengacu pada (4)

$$V_{ij}^{t} = \omega V_{ij} + c_1 r_1 \left(P_{best} - X_{ij}\right) + c_2 r_2 \left(It_{best} - X_{ij}\right) + c_3 r_3 \left(G_{best} - X_{ij}\right) \quad (4)$$

- V_{ij}^{t} : *velocity* atau kecepatan saat ini
- V_{ij} : *velocity* sebelumnya
- X_{ij} : lokasi partikel saat ini
- t : nomor iterasi
- $r_{1,2,3}$: bilangan acak [0,1]
- $c_{1,2,3}$: *learning rate*

Cek apakah V_{ij} masih berada di batas Vmax dan Vmin (V_{max} = 6 dan V_{min} = -6) serta perbarui posisi partikel dengan (5)

$$S_{ij} = \frac{1}{1+e^{-v_{ij}}} \quad (5)$$

d. Hitung nilai *fitness* menggunakan posisi partikel yang telah diperbarui seperti langkah 6, selanjutnya adalah perbarui nilai P_{best} It_{best} dan G_{best} dengan kondisi seperti berikut :
   1) Jika((*fitness*(X_i)>P_{best}) ‖ (*fitness*(X_i)=P_{best}) && (|X|< |P_{best}|)) maka P_{best} = nilai *fitness* X_i sehingga (P_{best1}, P_{best2},⋯P_{bestD}) = (X_{i1}, X_{i2},⋯ X_{iD})
   2) Jika((P_{best}(t)>It_{best}) ‖ (P_{best}(t)=It_{best}) && (|P_{best}(t)| < |It_{best}|)) maka It_{best} = P_{best2} sehingga (It_{best1}, It_{best2},…It_{bestD}) = (X_{i1}, X_{i2},…X_{iD})
   3) Jika((P_{best}>G_{best}) ‖ (P_{best}=G_{best}) && (|P_{best}|< |G_{best}|)) maka G_{best} = P_{best} sehingga (G_{best1}, G_{best2},…G_{bestD}) = (X_{i1}, X_{i2},…X_{iD})

8. Ulangi hingga mencapai iterasi maksimum.

Hasil dari proses PSO-KNN ini adalah berupa fitur-fitur optimal, dimana fitur-fitur ini yang akan digunakan sebagai model untuk menentukan kategori suatu cerita. Alur proses PSO-KNN dapat dilihat pada Gbr. 3.

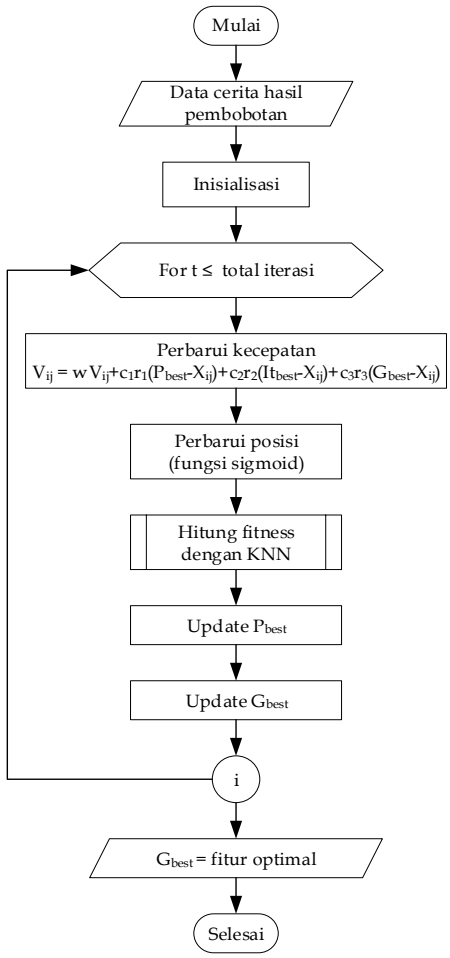
Proses pencarian nilai *fitness* adalah dengan menggunakan algoritma KNN yang divalidasi dengan LOOCV, berikut ini adalah proses pencarian nilai *fitness* :

1. Hitung jarak *Euclidean* antara objek yang akan diprediksi (n) dengan semua objek pelatihan (m-n) mengacu pada (6), namun hanya fitur yang bernilai biner 1 saja yang dihitung

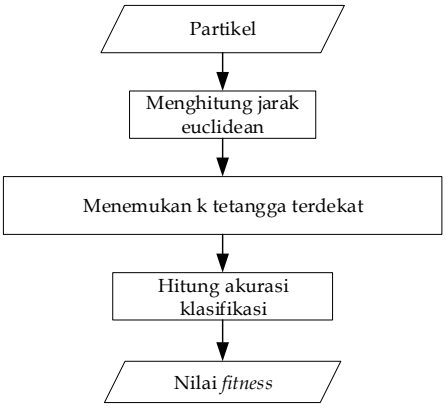
$$D(x_i x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_i) - a_r(x_j)\right)^2} \quad (6)$$

   - n : dimensi objek
   - a_r : data ke-*r*
   - $(x_i)(x_j)$: *record* ke-*i* , *record* ke-*j*
   - i, j : 1,2,3,…n
2. Temukan k tetangga terdekat dari objek yang mempunyai jarak terkecil . Kelas/label yang paling banyak diantara k tetangga akan menjadi label untuk objek yang diprediksi.
3. Hitung nilai fitness dengan menghitung akurasi pengklasifikasian dengan LOOCV yakni jumlah klasifikasi benar dibagi dengan jumlah sampel pelatihan. Hasil nilai fitness ini akan menjadi nilai P_{best}.

Alur proses pencarian nilai *fitness* dapat dilihat pada Gbr. 4



Gbr. 3 *Flowchart* PSO-KNN



Gbr. 4 *Flowchart* PSO-KNN

III. HASIL DAN PEMBAHASAN

A. *Dataset*

Data yang digunakan dalam penerapan sistemini adalah data cerita pendek yang didapatkan dari situs www.cerpenmu.com dengan jumlah data total sebanyak 750 data yang terbagi menjadi 5 kategori yakni fabel, horor, humor, romantis dan misteri. Deskripsi mengenai data yang digunakan dapat dilihat pada Tabel I.

TABEL I
DESKRIPSI DATA CERITA PENDEK

No	Kategori	*Training*	*Test*	Total
1	Horor	120	30	150
2	Misteri	120	30	150
3	Fabel	120	30	150
4	Humor	120	30	150
5	Romantis	120	30	150
Total Dokumen				750

B. *Penentuan Nilai k*

Penentuan nilai k dilakukan memakai dataset dengan 5 kategori (fabel, horor, humor, misteri dan romantis).

TABEL IIII
HASIL PENCARIAN NILAI K

Nilai *k*	Akurasi
k = 1	39,83%
k = 2	39,83%
k = 3	43,50%
k = 4	44,37%
***k* = 5**	**45,83%**
k = 6	45,67%
k = 7	45%
k = 8	45,67%
k = 9	45,67%
k = 10	44,83%
⋮	⋮
⋮	⋮
⋮	⋮
k = 199	22%
k = 200	22,17%

Dari hasil pengujian didapatkan nilai k terbaik adalah 5 dengan nilai akurasi sebesar 45,8%, sehingga nilai k inilah yang akan digunakan selama proses *learning* menggunakan PSO-KNN.

C. *Pengujian Model PSO-KNN*

Pengujian model ini dilakukan untuk menentukan model mana yang paling optimal untuk proses pengklasifikasi kategori teks cerita Bahasa Indonesia yang dihasilkan saat proses *learning* menggunakan PSO-KNN. tabel hasil pengujian dari 20 model PSO-KNN dengan menggunakan dataset 5 kategori cerita terdapat pada Tabel V.

Model yang paling optimal adalah model yang dihasilkan oleh 30 partikel pada iterasi ke-200 dengan total fitur optimal adalah 88 yang memiliki nilai akurasi tertinggi yakni 53%.

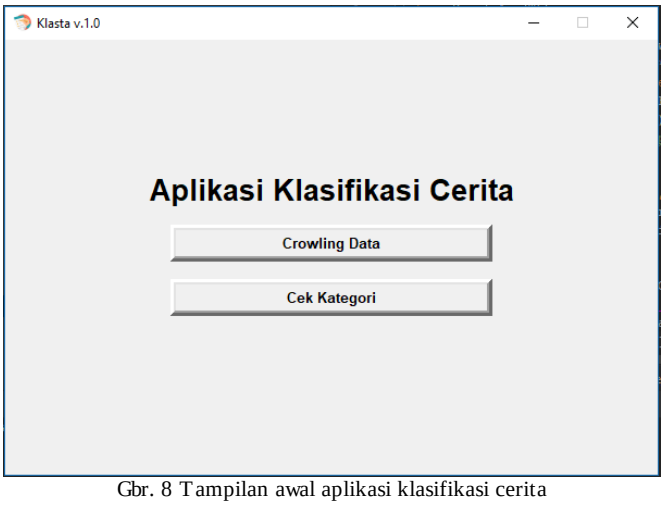
TABEL V
HASIL PENGUJIAN MODEL PSO-KNN 5 KATEGORI

No	**Partikel**	**Iterasi**	**Fitur Optimal**	**Akurasi *Testing***
1	10	50	239	43%
2		100	175	47%
3		150	124	51%
4		200	125	52%
5	20	50	110	38%
6		100	104	40%
7		150	102	41%
8		200	98	43%
9	**30**	50	94	45%
10		100	93	45%
11		150	91	45%
12		**200**	**88**	**53%**
13	40	50	53	48%
14		100	46	45%
15		150	40	47%
16		200	37	48%
17	50	50	107	48%
18		100	104	49%
19		150	101	49%
20		200	101	49%

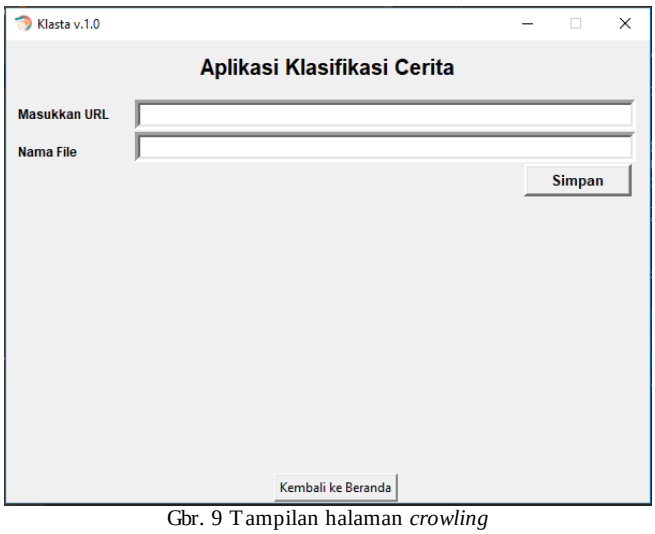
D. *Antarmuka Aplikasi*

Antarmuka yang dihasilkan berbasis desktop yang dalam penelitian ini peneliti sebut sebagai Klasta (Aplikasi Klasifikasi Cerita). Tampilan aplikasi Klasta ditunjukkan pada Gbr 8.

Pada halaman awal aplikasi ini pengguna memilih menu yakni "*crowling* data" atau "cek kategori". Jika pengguna ingin mengambil data cerita dari situs www.cerpenmu.com maka pengguna memilih menu "*crowling* data", jika pengguna ingin melakukan prediksi kategori cerita maka pengguna memilih menu "cek kategori". Tampilan halaman *crowling* data adalah sebagai berikut pada Gbr 9.



Gbr. 8 Tampilan awal aplikasi klasifikasi cerita

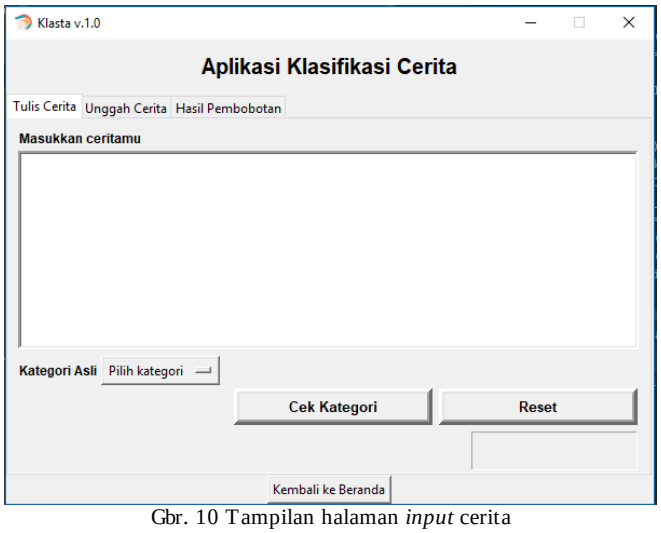


Gbr. 9 Tampilan halaman *crowling*

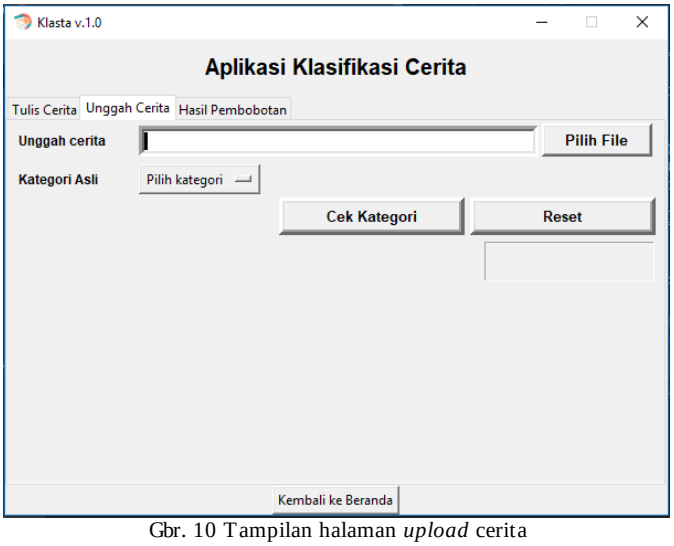
Pada halaman ini pengguna memasukkan url atau alamat situs yang akan diambil datanya, kemudian pengguna menekan tombol "simpan" untuk menyimpan data dalam format .txt dan secara otomatis pengguna akan diarahkan ke halaman cek kategori.

Halaman selanjutnya yakni halaman cek kategori, pada halaman ini terdapat dua cara untuk melakukan prediksi kategori cerita yakni upload file cerita dan *input* cerita. Tampilan halaman unggah file cerita dapat dilihat pada Gbr. 10.

Pada halaman ini pengguna memasukkan cerita yang akan diprediksi kategorinya pada kolom "masukkan ceritamu" . Untuk memulai proses prediksi pengguna menekan tombol cek kategori., maka hasil prediksi akan muncul.



Gbr. 10 Tampilan halaman *input* cerita



Gbr. 10 Tampilan halaman *upload* cerita

Pada halaman ini pengguna memasukkan *file* cerita dengan format .txt yang akan diprediksi kategorinya pada kolom "unggah cerita". Untuk memulai proses prediksi pengguna menekan tombol cek kategori., maka hasil prediksi akan muncul.

KESIMPULAN

Kesimpulan yang dapat diambil dari keseluruhan penelitian yang telah dilakukan adalah pengklasifikasian teks dengan menggunakan metode *Modified Binary Particle Swarm Optimization* dengan *K-Nearest Neighbor* (PSO-KNN) telah berhasil dibuat serta memiliki tampilan antarmuka berbasis desktop. Proses pembuatan sistem dimulai dari proses *text processing*, pembobotan TF-IDF, pemilihan fitur oleh MBPSO dan terakhir validasi fitur menggunakan KNN-LOOCV. Hasil uji coba metode *Modified Binary Particle Swarm Optimization* yang diterapkan pada metode KNN sebagai seleksi fitur telah terbukti dapat mengurangi fitur awal yang dihasilkan oleh model KNN. Hasil fitur yang dihasikan oleh model PSO-KNN

memiliki akurasi sebesar 53% dengan total fitur optimal hanya 88. Saran untuk penelitian selanjutnya adalah dengan melakukan percobaan menggunakan dataset yang lain serta mencoba algoritma pengukuran jarak selain *Euclidean distance* seperti *Manhattan distance* atau *Minkowski distance*.

UCAPAN TERIMA KASIH

Puji syukur penulis ucapkan kepada Allah SWT. yang telah memberikan ridho-Nya dan senantiasa memberikan kemudahan dan kelancaran dalam pengerjaan jurnal ini serta tak lupa penulis ucapkan terimakasih kepada semua pihak yang telah ikut berperan membantu proses pengerjaan jurnal ini sehingga jurnal ini dapat terselesaikan dengan baik.

REFERENSI

[1] Somantri, Oman. 2017. Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes (NB). Jurnal Telematika, Volume 12.

[2] Hua, K.-L. & al, e., 2015. Computer-aided Classification of Lung Modules on Computed Tomograhy Image fo Deep Learning Technique. Onco Tragets and Therapy, Volume 8.

[3] Somantri, Oman. & Khambali, Mohammad. 2017. Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritme Genetika. JNTETI, Volume 6, pp. 301-306.

[4] Bijalwan, Vishwanath. et al., 2014. Machine Learning Approach for Text and Document Mining. International Journal of Database Theory and Application, Volume 7, pp. 61-70.

[5] Sheba, Raj, G. & Ramachandran, 2018. A Modified Binary PSO based Feature Selection for Automatic Lesion Detection in Mammograms. International Journal of Computer Science & Information Technology (IJCSIT), Volume 10, pp. 39-55.