



Klasifikasi Genre *Line Webtoon* menggunakan *K-Nearest Neighbor* berbasis *Particle Swarm Optimization* berdasarkan Ringkasan Cerita

Line Webtoon Genre Classification using K-Nearest Neighbor with Particle Swarm Optimization based on Story Summary

Siti Nur Khaizah, Prodi Matematika FMIPA UNY
Bambang Sumarno Hadi Marwoto*, Prodi Matematika FMIPA UNY
*e-mail: bambang@uny.ac.id

Abstrak

Penelitian ini bertujuan untuk membangun model pengklasifikasian genre Webtoon berdasarkan ringkasan cerita. Dari model yang dihasilkan, akan dieksplorasi proses seleksi fitur pada perbaikan algoritma klasifikasi yang dilihat melalui nilai akurasi dan waktu komputasi. Algoritma klasifikasi yang digunakan adalah *K-Nearest Neighbors* (KNN) serta algoritma yang digunakan untuk seleksi fitur adalah *Particle Swarm Optimization* (PSO). Proses ini diawali dengan pemilihan data dan *preprocessing* yang terdiri dari *cleaning*, *filtering*, *lemmatization* dan *stemming*, serta *tokenizing*. Kemudian, diikuti proses *transformation* dan seleksi fitur dengan *package Pyswarm*. Fitur terpilih kemudian digunakan pada proses *data mining* yaitu klasifikasi. selanjutnya hasil dievaluasi dan diinterpretasi. Hasil penelitian menunjukkan bahwa algoritma PSO-KNN terbukti lebih baik daripada algoritma KNN dalam mengklasifikasi genre Webtoon. Diperoleh nilai akurasi dari klasifikasi dengan algoritma KNN yaitu sebesar 88% sedangkan dengan algoritma PSO-KNN diperoleh nilai akurasi sebesar 100%. Selain itu waktu komputasi yang digunakan oleh algoritma PSO-KNN dalam mengkasifikasi terbukti lebih singkat dari algoritma KNN.

Kata kunci: *text mining, k-nearest neighbors, particle swarm optimization, seleksi fitur.*

Abstract

This study aimed to develop a Webtoon classification model genre based on the story summary. From the model, it would be explored the selection feature on the algyorhythm improvement based on accuracion score and computation times. The classification algorithm used is K-Nearest Neighbors (KNN) and the algorithm used for feature selection is Particle Swarm Optimization (PSO). This process begins with data selection and preprocessing consisting of cleaning, filtering, lemmatization and stemming, and tokenizing. Then, followed by the transformation process and feature selection with the Pyswarm package. The selected features used in the data mining process, namely classification. Then the results are evaluated and interpreted. The results showed that the PSO-KNN algorithm proved to be better than the KNN algorithm in classifying the Webtoon genre. The accuracy value obtained from the classification with the KNN algorithm is 88% while the PSO-KNN algorithm obtained an accuracy value of 100%. In addition, the computational time used by the PSO-KNN algorithm in classifying is proven to be shorter than the KNN algorithm.

Keywords: *text mining, k-nearest neighbors, particle swarm optimization, feature selection.*

PENDAHULUAN

Perkembangan teknologi telah mengubah gaya hidup, aktivitas dan kebiasaan masyarakat, di mana sebagian besar masyarakat mulai terbiasa dengan kegiatan digital dalam kehidupan mereka (Lestari & Irwansyah, 2020). Salah satunya pada industri penerbitan buku yang berhasil memunculkan istilah baru yaitu komik digital atau *digital comic* (Jin, 2015). Salah satu *platform* yang menyediakan komik digital adalah LINE Webtoon atau disebut Webtoon yang dapat diakses secara gratis melalui *website* resmi Webtoon atau aplikasi LINE secara langsung.

Setiap karya sastra yang mengandung unsur cerita umumnya harus memiliki alur cerita dan jenis cerita/genre yang jelas. Genre menjadi salah satu unsur cerita yang sangat berpengaruh dalam mendeskripsikan bagaimana cerita yang akan dibawakan dalam suatu karya. Penelitian Rizky & Stellarosa pada tahun 2017 menunjukkan sebesar 96% responden setuju bahwa genre dalam film membantu penonton dalam memilih film yang akan ditonton serta sebesar 86% responden akan menonton film berdasarkan genre yang disukai (Rizky & Stellarosa, 2017).

Upaya untuk mendeskripsikan suatu genre secara tekstual bukan suatu hal yang mudah dilakukan karena diperlukan perbedaan yang jelas antara satu genre dengan genre yang lain, disamping itu sering terjadi genre yang tercampur atau tumpang tindih (Chandler, 2004). Pada dasarnya, genre dalam sastra mempunyai persyaratan tersendiri pada teknik kesusastraan yang dapat dilakukan dengan menganalisis unsur intrinsik, seperti tema, tokoh, metode penceritaan, dan lain-lain (Purwaningsih, 2017). Terlebih lagi, Webtoon memberikan kebebasan semua orang untuk berkarya dan mengunggah karyanya melalui Webtoon. Tentunya tidak semua orang mampu memahami bagaimana menentukan genre dengan tepat, sehingga berakibat pada pengelompokan genre Webtoon yang belum tentu tepat sesuai dengan ceritanya. Oleh karena itu, diharapkan penelitian ini mampu mempermudah orang awam dalam menentukan genre karya mereka melalui ringkasan cerita.

Penelitian tentang pengklasifikasian genre pernah dilakukan pada studi kasus film dengan menggunakan data teks sinopsis film. Penelitian tersebut dilakukan oleh Saputra, dkk pada tahun 2019 dengan menggunakan metode *Support Vector Machine* (SVM) dan diperoleh nilai akurasi sebesar 45% (Saputra et al., 2019). Masih seputar pengklasifikasian data teks, penelitian serupa juga pernah dilakukan dengan menggunakan algoritma Naive Bayes yang dikolaborasikan dengan *Particle Swarm Optimization* (PSO) untuk mengklasifikasi emosi pada lirik lagu dan diperoleh nilai akurasi sebesar 90%-96% (Ramadhan et al., 2021). Lebih lengkap, algoritma yang pernah dilakukan untuk klasifikasi data teks antara lain algoritma *Rocchio*, *bagging and boosting*, regresi logistik, *decision tree*, *random forest*, *Naïve Bayes Classifier* (NBC), *K-Nearest Neighbor* (K-NN), *Support Vector Machine* (SVM) (Kowsari et al., 2019).

Telah diakui bahwa KNN merupakan salah satu algoritma paling sederhana untuk menyelesaikan berbagai kasus klasifikasi termasuk dual/multi klasifikasi karena KNN menjadi salah satu algoritma nonparametrik yang tidak memerlukan asumsi atau model awal serta *output* yang dihasilkan pun cukup mudah untuk dipahami dan diartikan (Primartha, 2018). Perlu diingat bahwa semua algoritma dapat bekerja lebih optimal jika didampingi dengan kemampuan dalam memahami metode ekstraksi fitur dan cara mengevaluasinya dengan baik dan benar (Kowsari et al., 2019). Oleh sebab itu, banyak peneliti yang memodifikasi ataupun mengkolaborasikan dua algoritma secara langsung. Seperti pada penelitian yang dilakukan oleh Muslimah dan Wihandika, penelitian ini membuktikan bahwa algoritma *improved K-Nearest Neighbor* (KNN) mencapai nilai akurasi yang lebih baik dan lebih stabil dibandingkan dengan algoritma KNN saja (Muslimah & Wihandika, 2019).

KNN memiliki kekurangan pada pembelajaran berdasarkan jarak, di mana jenis jarak yang digunakan kerap tidak jelas ditambah perhitungan jarak dari tiap data uji cukup memakan banyak waktu komputasi (Primartha, 2018). Oleh sebab itu, upaya-upaya perbaikan masih perlu dilakukan untuk meningkatkan kinerja algoritma KNN. Seperti pada penelitian yang dilakukan

(Chantar & Corne, 2011), mereka menambah optimasi PSO untuk seleksi fitur dan telah terbukti mampu meningkatkan nilai akurasi dari algoritma klasifikasi SVM dan naïve bayes. Selain itu, PSO juga mudah untuk diimplementasikan karena PSO tidak menggunakan asumsi-asumsi atau model awal dan memiliki parameter yang relatif kecil (Ramadhan et al., 2021). Metode PSO untuk seleksi fitur pernah dikolaborasikan dengan teknik *bagging* untuk mendeteksi sel kanker serviks yang dilihat dari normal atau tidaknya sel pada uji *pap smear* dan diperoleh hasil akurasi terbaik sebesar 95.05% dengan menggunakan algoritma klasifikasi KNN sedangkan dengan Naïve Bayes hanya diperoleh nilai akurasi sebesar 64.24% (Arifin, 2017). Sehingga dapat diketahui bahwa algoritma PSO lebih efektif jika dikolaborasikan dengan KNN dari pada *Naïve Bayes*.

Dengan pertimbangan tersebut, maka pada penelitian ini akan digunakan algoritma KNN yang dikolaborasikan dengan *Particle Swarm Optimization* (PSO) untuk menyeleksi fitur sehingga diharapkan dapat diperoleh hasil akurasi yang lebih tinggi untuk mengklasifikasikan genre Webtoon berdasarkan ringkasan cerita serta mampu mempersingkat waktu komputasi karena tidak semua fitur digunakan dalam klasifikasi.

METODE

Penelitian klasifikasi genre *line Webtoon* menggunakan *k-nearest neighbor* berbasis *particle swarm optimization* berdasarkan ringkasan cerita merupakan jenis penelitian dengan pendekatan kuantitatif. Data yang digunakan pada penelitian ini adalah *dataset* Webtoon sebanyak 1309 serial berbahasa Inggris yang berasal dari Webtoon Originals dengan 10 atribut yaitu, *id, name, writer, likes, genre, rating, subscriber, summary, update, dan reading link*. Penelitian ini menggunakan analisis pengolahan data *text mining* sesuai dengan kerangka KDD yang dijelaskan oleh (Adiwijaya, 2006) dengan tambahan seleksi fitur menggunakan PSO yang menjadi fokus dari penelitian ini. Secara keseluruhan, penelitian ini terdiri dari enam tahapan, yaitu:

- (1) Pemilihan data (*data selection*),
- (2) Pra-pemrosesan data (*data preprocessing*),
- (3) Transformasi data (*data transformation*),
- (4) Seleksi fitur (*feature selection*),
- (5) Penambangan data (*data mining*), dan
- (6) Interpretasi dan evaluasi hasil (*interpretation and evaluation*).

Dari keenam proses tersebut, terdapat satu proses yang memiliki sub-proses yaitu proses *preprocessing*. *Preprocessing* diawali dengan proses pembersihan data (*cleaning data*), kemudian dilanjutkan dengan *case folding, filtering* atau penghapusan *stopword, stemming* dan *lemmatization*, serta *tokenizing*. Pada proses pemilihan data (*data selection*), Pra-pemrosesan data (*preprocessing*) kecuali *tokenizing*, akan dilakukan dengan bantuan R Studio. Kemudian dilanjutkan dengan *notebook jupyter* (Python) hingga proses interpretasi dan evaluasi (*interpretation and evaluation*).

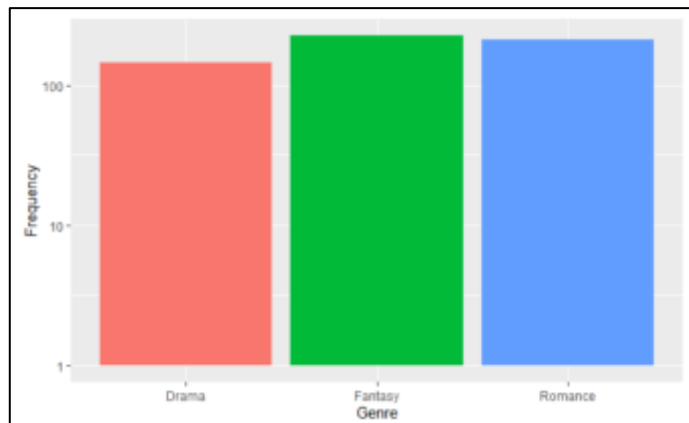
HASIL DAN PEMBAHASAN

Hasil

1. Pemilihan Data (*Data Selection*)

Sebelum dilakukan pemilihan data, terlebih dahulu dilakukan *import* data csv ke dalam R Studio. Kemudian, pilih data yang akan digunakan untuk penelitian ini yaitu data serial Webtoon dengan atribut *Summary* dan Genre saja. Sehingga atribut *id, Name, Writer, Likes, Rating, Subscribers, Update, dan Reading Link* dihapus pada proses ini. Selanjutnya, dari data tersebut akan dipilih 3 genre dengan jumlah terbanyak yaitu *fantasy* sebanyak 227 serial, *romance* sebanyak 211 serial, dan *drama* sebanyak 145 serial. Hasil dari proses seleksi ini adalah data teks dengan 2 atribut, yaitu 583 ringkasan

cerita (*Summary*) dan 3 label genre. Visualisasi data hasil seleksi ditampilkan pada Gambar 1.



Gambar 1. Visualisasi Genre Romance, Fantasy, dan Drama

2. Pra-pemrosesan Data (*Data Preprocessing*)

a. *Cleaning*

Data yang berhasil dikumpulkan kemudian dilakukan penghapusan data duplikat dengan *package* yang disediakan R Studio yaitu *package* 'Dplyr'. Proses ini mendeteksi sebanyak 73 data duplikat, sehingga tersisa sebanyak 510 data dari 583 data awal. Kemudian proses pembersihan dilanjut dengan mengembalikan singkatan, menghapus angka, tanda baca/*punctuation*, serta tanda dan karakter yang tidak terpakai. Perbedaan data sebelum dan setelah proses *cleaning* terlihat pada Gambar 2 dan Gambar 3.

Sebelum:

```
[10] Though she comes from a rich family, Gabin is a
bullied loner at school. That doesn't mean she puts up
with bullying from other students, though -- Gabin is
infamous for her fierce personality. worried about
Gabin's well-being, her family secretly offers Tae-ung
$500,000 to transci-fier schools and act as her
bodyguard. will Tae-ung be able to protect her from the
school bullies, all while keeping his identity hidden
from Gabin and the other students?
```

Gambar 2. Output sebelum proses *cleaning*

Sesudah:

```
[10] Though she comes from a rich family Gabin is a
bullied loner at school That does not mean she puts up
with bullying from other students though Gabin is
infamous for her fierce personality worried about Gabin
s well being her family secretly offers Tae ung to
transci fier schools and act as her bodyguard will Tae
ung be able to protect her from the school bullies all
while keeping his identity hidden from Gabin and the
other students
```

Gambar 3. Output sesudah proses *cleaning*

b. *Case Folding*

Pada proses ini, seluruh kata akan diubah ke bentuk huruf kecil karena di dalam R, huruf kapital dan huruf kecil memiliki nilai atau arti yang berbeda. Perbedaan sebelum dan setelah proses *case folding* terlihat pada Gambar 4 dan Gambar 5.

Sebelum:

[1] She is young single and about to achieve her dream of creating incredible videogames. But then life throws her a one two punch a popular streamer gives her first game a scathing review Even worse she finds out that same troublesome critic is now her new neighbor A funny sexy and all too real story about gaming memes and social anxiety Come for the plot stay for the doggo
[2] After binge watching beauty videos online a shy comic book fan masters the art of makeup and sees her social standing skyrocket as she becomes her school s prettiest pretty girl overnight But will her elite status be short lived How long can she keep her real self a secret And what about that cute boy who knows her secret
[3] After making a grisly discovery in the countryside a small town book editor s life gets entangled with a young Mafia lord and his intimidating bodyguard even as every step she takes draws her deeper into the dangerous underworld of the city
[4] She is a hopeless romantic who is turning s and is not super happy about it he is a reclusive billionaire who is hired her to be his assistant and he is not too happy about that either Together they rewrite the rules of friendship love work and the BEST way to clean someone s apartment
[5] A runaway werewolf Lumine meets a witch boy named Kody and is being hired to work as his bodyguard Lumine being determined to befriend Kody tries everything to get Kody to trust him while Kody tries to keep Lumine away from him Their life gets even more complicated after an unfortunate event that leaves them to try to survive on their own

Gambar 4. Output sebelum proses case folding

Setelah:

[1] she is young single and about to achieve her dream of creating incredible videogames but then life throws her a one two punch a popular streamer gives her first game a scathing review even worse she finds out that same troublesome critic is now her new neighbor a funny sexy and all too real story about gaming memes and social anxiety come for the plot stay for the doggo
[2] after binge watching beauty videos online a shy comic book fan masters the art of makeup and sees her social standing skyrocket as she becomes her school s prettiest pretty girl overnight but will her elite status be short lived how long can she keep her real self a secret and what about that cute boy who knows her secret
[3] after making a grisly discovery in the countryside a small town book editor s life gets entangled with a young mafia lord and his intimidating bodyguard even as every step she takes draws her deeper into the dangerous underworld of the city
[4] she is a hopeless romantic who is turning s and is not super happy about it he is a reclusive billionaire who is hired her to be his assistant and he is not too happy about that either together they rewrite the rules of friendship love work and the best way to clean someone s apartment
[5] a runaway werewolf lumine meets a witch boy named kody and is being hired to work as his bodyguard lumine being determined to befriend kody tries everything to get kody to trust him while kody tries to keep lumine away from him their life gets even more complicated after an unfortunate event that leaves them to try to survive on their own

Gambar 5. Output setelah proses case folding

c. *Filtering*

Proses ini akan menghapus semua *stopword* yaitu kata-kata yang tidak dapat dijadikan acuan klasifikasi atau dengan kata lain, kata-kata yang tidak mencirikan salah satu label genre. Perbedaan sebelum dan sesudah proses *filtering* dapat dilihat dengan menggunakan visualisasi *wordcloud* pada Gambar 6 dan Gambar 7.

Sebelum:



Gambar 6. Wordcloud sebelum filtering

Setelah:



Gambar 7. Wordcloud setelah filtering

d. *Stemming* dan *Lemmatization*

Proses *lemmatization* dilakukan terlebih dahulu dengan pertimbangan bahwa proses ini bekerja lebih baik dalam mengubah bentuk *irregular* bila dibandingkan dengan *stemming*, serta tidak merubah kata secara sembarangan yang berakibat merusak kata seperti pada *stemming*. Perbedaan sebelum dan setelah proses *stemming* dan *lemmatization* terlihat pada Gambar 8 dan Gambar 9.

Sebelum:

```
[1] young single achieve dream creating incredible videogames life throws one two punch popular streamer gives first game scathing review even worse finds troublesome critic now new neighbor funny sexy real story gaming memes social anxiety come plot stay doggo [2] binge watching beauty videos online shy comic book fan masters art makeup sees social standing skyrocket becomes school prettiest pretty girl overnight will elite status short lived long can keep real self secret cute boy knows secret [3] making grisly discovery countryside small town book editor life gets entangled young mafia lord intimidating bodyguard even every step takes draws deeper dangerous underworld city
```

Gambar 8. Output sebelum *stemming* dan *Lemmatization*

Setelah:

```
[1] young singl achiev dream creat incred videogam life throw one two punch popular streamer give first game scath review even bad find troublesom critic now new neighbor funni sexi real stori game meme social anxieti come plot stay doggo [2] bing watch beauti video onlin shi comic book fan master art makeup see social stand skyrocket becom school pretti pretti girl overnight will elit status short live long can keep real self secret cute boy know secret [3] make grisli discoveri countrysid small town book editor life get entangl young mafia lord intimid bodyguard even everi step take draw deep danger underworld citi
```

Gambar 9. Output setelah *stemming* dan *Lemmatization*

e. *Tokenizing*

Proses *Tokenizing* dilakukan dengan *notebook* jupyter. Sehingga, untuk melanjutkan proses dari R Studio ke *notebook* python, data hasil *stemming* disimpan terlebih dahulu ke komputer dengan format CSV. Kemudian data tersebut akan di *import* ke dalam *notebook* jupyter. Dengan menggunakan python, proses pemisahan kata atau *tokenizing* digabungkan secara langsung dengan proses pembobotan atau *transformation*. Dari proses *tokenizing*, diperoleh fitur sebanyak 3273 kata. Hasil ini merupakan list kata yang unik artinya meskipun terdapat kata yang sama dalam *summary* serial yang berbeda, kata tersebut tetap terhitung satu.

3. Transformasi Data/Pembobotan kata (*Data Transformation*)

Proses ini bertujuan untuk mengubah data teks menjadi data numerik dengan menghitung bobot tiap kata dalam suatu dokumen. Pada proses ini digunakan TF-IDF untuk merubah data *summary* menjadi numerik serta label *encoding* untuk merubah data genre menjadi numerik dengan 0 untuk genre *Romance*, 1 untuk genre *Fantasy*, dan 2 untuk genre *Drama*. Hasil dari proses *transformation* ditunjukkan pada Gambar 10..

	abandon	abbey	abbott	abili	abl	abroad	absolut	absorb	abus	academi	...	yuzuki	Zahn	zap	zecallon	zedekiah	zeha	zerlasha	zero	zeroni	zombi
D1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...
506	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
507	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
508	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
509	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
510	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

0 rows x 3273 columns

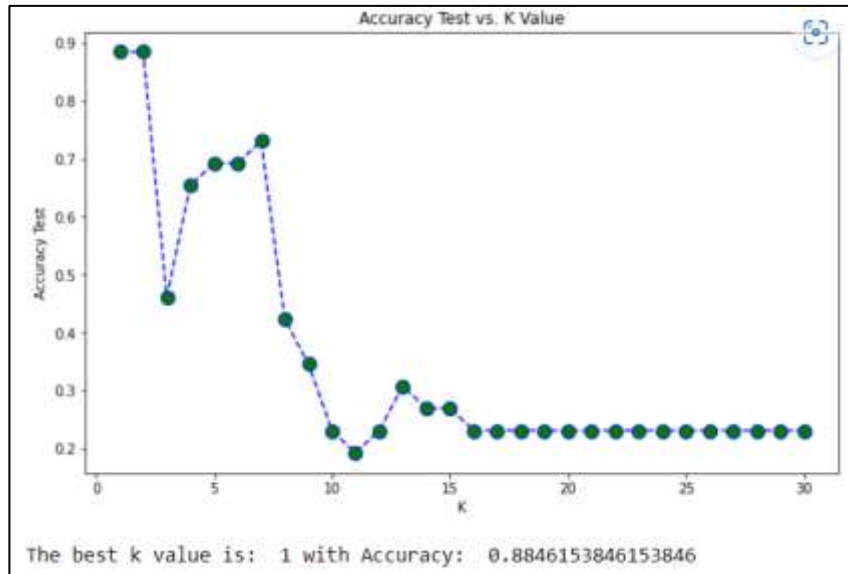
Gambar 10. Output hasil transformation

4. Penambangan Data (*Data Mining*)

Penambangan data dimaksudkan pada algoritma yang akan digunakan dalam klasifikasi genre Webtoon. Dilakukan 2 skema penambangan data, yaitu klasifikasi genre Webtoon dengan algoritma KNN dan klasifikasi genre Webtoon dengan algoritma KNN yang dilengkapi dengan seleksi fitur PSO. Sebelum kedua skema ini diproses, hasil pembobotan data *summary* dan hasil *encoder* label genre akan dibagi terlebih dahulu dengan rasio 95%:5% dari keseluruhan dataset 510 serial. Sehingga diperoleh data latih (*data train*) sebanyak 484 data dan data uji (*data test*) sebanyak 26 data. Kemudian, kedua skema tadi akan diproses sebagai berikut.

a. Skema 1 : Klasifikasi dengan algoritma KNN

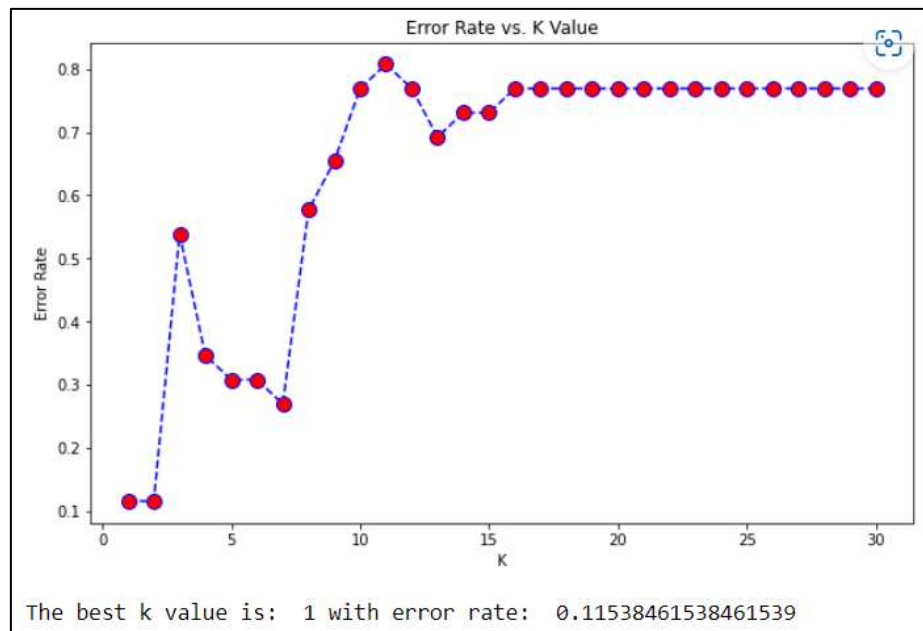
Dalam skema 1, KNN langsung diterapkan pada data train atau lebih tepatnya X_{train} yang merupakan hasil pembobotan TF-IDF dengan nilai k pada interval 1 sampai 31. Kemudian dihitung nilai akurasi yang dihasilkan dari membandingkan hasil prediksi data tes dengan nilai label yang sesungguhnya. Hasil ini akan memperlihatkan nilai k yang memberikan *error rate* terkecil serta akurasi terbaik yang kemudian nilai k tersebut akan digunakan untuk mengklasifikasi data latih dan uji. Gambar 11 menunjukkan plot nilai k dan akurasi pada data uji untuk skema 1.



Gambar 11. Plot Nilai K dan Akurasi Data Uji dengan Skema 1

Plot ini memperlihatkan bahwa nilai akurasi terbaik tercapai pada nilai $k = 1$ yaitu sebesar 88%.

Gambar 12 menunjukkan plot nilai k dan *error rate* pada data uji untuk skema 1. Plot ini memperlihatkan nilai *Error rate* terendah tercapai pada nilai $k = 1$ yaitu sebesar 12%.



Gambar 12. Plot Nilai K Dan Error Rate Data Uji dengan Skema 1

Gambar 13 merupakan hasil prediksi data latih pada skema 1. Hasil tersebut memperlihatkan bahwa data latih dikelompokkan ke dalam label dengan menggunakan skema 1 selama 0.281 detik.


```

RESULT[DATA TRAIN]: [1 2 0 2 2 1 2 1 2 1 2 0 2 1 2 2 2 2 2 0 0 1 1 1 1
0 0 0 1 0 1 2 2 1 2 2 0
1 2 2 0 1 2 0 0 2 0 0 0 1 2 0 2 0 2 1 2 1 1 1 1 0 2 2 2 0 2 1 2 1 1 1
1 1
1 1 1 1 2 1 1 2 2 2 0 2 0 0 1 2 2 2 0 0 0 0 1 2 0 0 1 0 2 2 1 0 1 0 2
2 1
1 1 1 2 1 0 2 0 1 2 2 2 2 0 1 1 2 2 1 1 2 0 0 2 0 2 2 2 2 0 0 1 1 1 0
1 2
2 2 1 1 0 1 0 2 1 2 1 0 1 1 2 1 0 1 0 2 1 0 1 1 2 1 0 1 0 0 1 1 2 1 2
2 2
1 1 1 1 2 2 1 2 0 1 1 2 2 2 1 2 1 2 2 2 1 1 2 2 0 2 0 2 0 0 2 0 2 2 2
1 1
1 1 2 2 1 2 0 1 2 1 1 2 1 1 2 0 0 0 1 0 1 2 2 1 1 0 2 1 2 2 0 0 1 2 1
2 2
2 2 2 1 1 1 0 1 2 2 1 0 2 1 1 1 1 0 1 2 2 1 0 2 2 1 2 2 0 2 1 2 1 1
1 2
1 0 1 2 2 1 0 0 0 1 0 2 1 2 1 2 2 0 1 1 0 0 2 0 0 2 1 1 1 1 2 0 1 1 0
2 1
0 2 0 2 1 2 0 1 1 1 2 0 0 2 1 0 1 1 0 1 2 1 2 2 1 2 1 2 1 2 2 0 1 0 0
1 2
1 0 1 2 1 2 1 1 2 2 2 1 2 2 0 1 0 2 1 1 1 1 1 1 2 2 1 2 1 0 0 1 2 2 0
0 1
2 1 1 0 1 0 0 0 0 1 0 1 2 2 2 2 1 0 0 2 1 0 1 1 2 2 2 1 0 0 0 1 1 2 1
1 1
1 1 1 1 2 0 2 2 2 2 1 1 1 2 0 2 0 1 1 1 1 1 0 0 2 0 1 1 2 1 1 1 2 1 0
0 1
1 2 2]
CPU times: total: 281 ms
Wall time: 226 ms
    
```

Gambar 13. Output Hasil Prediksi Klasifikasi Data Latih dengan Skema 1

Hasil prediksi data uji pada skema 1 ditunjukkan pada Gambar 14. Hasil tersebut memperlihatkan bahwa data uji dikelompokkan ke dalam label dengan menggunakan skema 1 selama 0.0469 detik.

```

RESULT[DATA TEST]: [2 2 2 1 2 0 2 0 1 1 2 2 0 2 1 2 1 2 2 0 0 0 2 2 2
1]
CPU times: total: 46.9 ms
Wall time: 81 ms
    
```

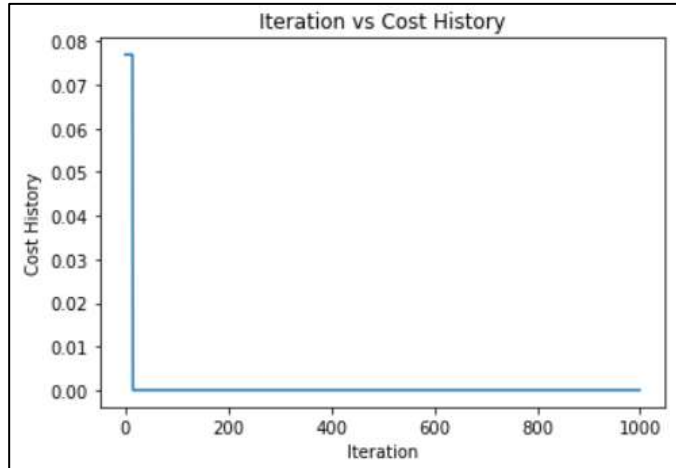
Gambar 14. Output Hasil Prediksi Klasifikasi Data Uji dengan Skema 1

- b. Skema 2 : Klasifikasi dengan algoritma KNN dan PSO
 Skema ini tidak jauh berbeda dengan skema yang pertama, hanya saja terdapat tambahan PSO setelah proses pembobotan dan sebelum proses klasifikasi KNN. Penambahan PSO diawali dengan *install package* ‘pyswarm’ yang merupakan paket algoritma PSO untuk seleksi fitur. Kemudian dilakukan beberapa persiapan seperti mengimpor *library* PSO, mendefinisikan fungsi-fungsi dasar PSO, inialisasi nilai C1, C2, w, k, dan nilai p. Fungsi yang dibentuk langsung diterapkan pada data latih sebanyak 1000 iterasi hingga diperoleh hasil *best cost* dan *best position*/fitur-fitur kata yang terpilih seperti pada Gambar 15 dan Gambar 16.

```

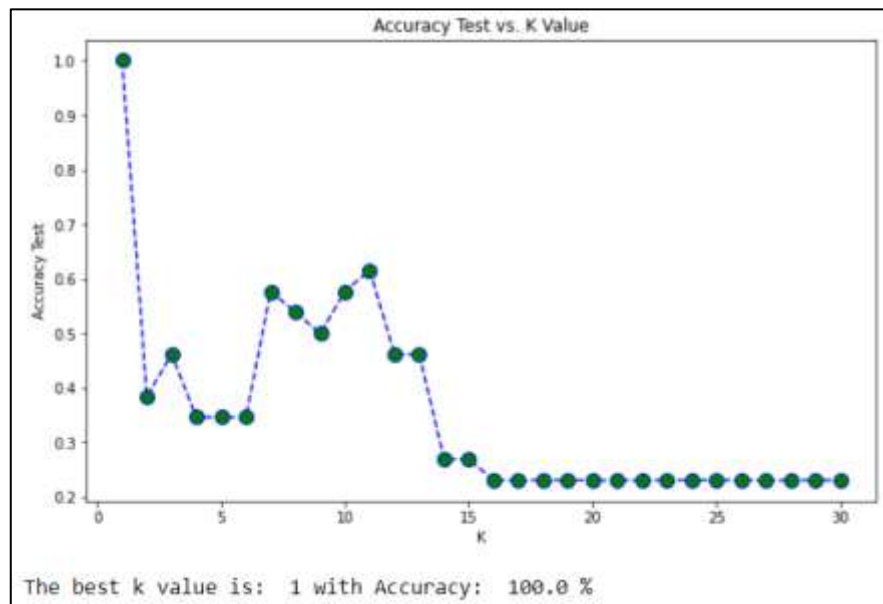
2022-10-06 11:32:57,571 - pyswarms.discrete.binary - INFO - Optimize fo
r 1000 iters with {'c1': 1.774, 'c2': 1.902, 'w': 0.9, 'k': 7, 'p': 2}
pyswarms.discrete.binary: 100%|██████████████████████████████████████
██████████████████████████████████████|1000/1000, best_cost=0
2022-10-06 11:43:58,328 - pyswarms.discrete.binary - INFO - Optimizatio
n finished | best cost: 0.0, best pos: [1 1 0 ... 0 1 0]
    
```

Gambar 15. Output Proses PSO



Gambar 16. Plot *Cost History* dan Iterasi

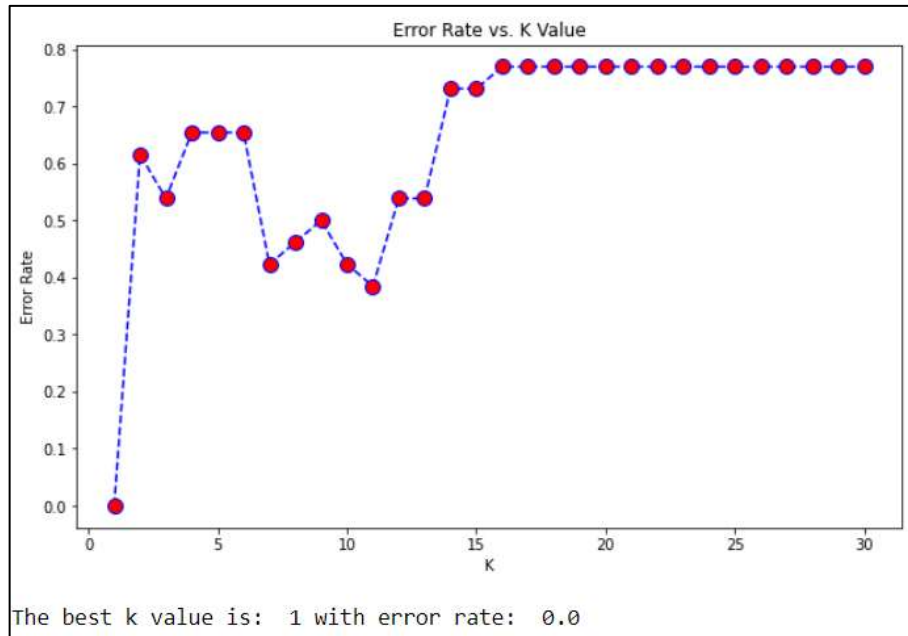
Selanjutnya, diperoleh fitur kata sejumlah 1653 kata dari 3273 total kata. Fitur-fitur tersebut akan langsung digunakan untuk membuat subset fitur pada data uji atau lebih tepatnya *X_{test}*. Kemudian, data latih dan data uji yang telah diseleksi akan diproses dengan algoritma KNN seperti pada skema 1.



Gambar 17. Plot Nilai K dan Akurasi Data uji dengan Skema 2

Gambar 17 merupakan plot nilai *k* dan akurasi pada data uji untuk skema 2. Plot ini memperlihatkan bahwa nilai akurasi tertinggi klasifikasi data uji dengan skema 2 tercapai pada nilai *k* = 1 yaitu sebesar 100%.

Gambar 18 merupakan plot nilai *k* dan *error rate* pada data uji untuk skema 2. Plot ini memperlihatkan nilai *Error rate* terendah tercapai pada nilai *k* = 1 yaitu sebesar 0%.



Gambar 18. Plot Nilai K dan *Error Rate* pada Data uji dengan Skema 2

Gambar 19 merupakan hasil prediksi dari pengklasifikasian data latih dengan algoritma KNN dan PSO.

```

RESULT TRAIN: [1 2 0 2 2 1 2 1 2 1 2 0 2 1 2 2 2 2 2 0 0 1 1 1 1 0 0 0
1 0 1 2 2 1 2 2 0
1 2 2 0 1 2 0 0 2 0 0 0 1 2 0 2 0 2 1 2 1 1 1 1 0 2 2 2 0 2 1 2 1 1 1
1 1
1 1 1 1 2 1 1 2 2 2 0 2 0 0 1 2 2 2 0 0 0 0 1 2 0 0 1 0 2 2 1 0 1 0 2
2 1
1 1 1 2 1 0 2 0 1 2 2 2 2 0 1 1 2 2 1 1 2 0 0 2 0 2 2 2 2 0 0 1 1 1 0
1 2
2 2 1 1 0 1 0 2 1 2 1 0 1 1 2 1 0 1 0 2 1 0 1 1 2 1 0 1 0 0 1 1 2 1 2
2 2
1 1 1 1 2 2 1 2 0 1 1 2 2 2 1 2 1 2 2 2 1 1 2 2 0 2 0 2 0 0 2 0 2 2 2
1 1
1 1 2 2 1 2 0 1 2 1 1 2 1 1 2 0 0 0 1 0 1 2 2 1 1 0 2 1 2 2 0 0 1 2 1
2 2
2 2 2 1 1 1 0 1 2 2 1 0 2 1 1 1 1 1 0 1 2 2 1 0 2 2 1 2 2 0 2 1 2 1 1
1 2
1 0 1 2 2 1 0 0 0 1 0 2 1 2 1 2 2 0 1 1 0 0 2 0 0 2 1 1 1 1 2 0 1 1 0
2 1
0 2 0 2 1 2 0 1 1 1 2 0 0 2 1 0 1 1 0 1 2 1 2 2 1 2 1 2 1 2 2 0 1 0 0
1 2
1 0 1 2 1 2 1 1 2 2 2 1 2 2 0 1 0 2 1 1 1 1 1 1 2 2 1 2 1 0 0 1 2 2 0
0 1
2 1 1 0 1 0 0 0 0 1 0 1 2 2 2 2 1 0 0 2 1 0 1 1 2 2 2 1 0 0 0 1 1 2 1
1 1
1 1 1 1 2 0 2 2 2 2 1 1 1 2 0 2 0 1 1 1 1 1 0 0 2 0 1 1 2 1 1 1 2 1 0
0 1
1 2 2]
CPU times: total: 266 ms
wall time: 136 ms
    
```

Gambar 19. Output Hasil Prediksi Klasifikasi Data Latih dengan Skema 2

Hasil di atas memperlihatkan bahwa data latih dikelompokkan ke dalam label dengan menggunakan skema 1 selama 0.266 detik.

Gambar 20 merupakan hasil prediksi data uji pada skema 2. Hasil tersebut memperlihatkan bahwa data uji dikelompokkan ke dalam label dengan menggunakan skema 2 selama 0.0469 detik.

```
RESULT TEST: [2 2 2 1 2 0 0 0 1 1 2 2 0 0 1 2 1 0 2 0 0 0 2 2 2 1]
CPU times: total: 46.9 ms
Wall time: 42 ms
```

Gambar 20. Output Hasil Prediksi Klasifikasi Data Uji dengan Skema 2

5. Evaluasi dan Interpretasi (Evaluation and Interpretation)

a. Evaluasi dan Interpretasi klasifikasi genre Webtoon dengan algoritma KNN (Skema 1).

1) Data latih

a) *Confusion Matrix*

[[116	0	0]
[0	194	0]
[1	0	173]]

Gambar 21. Output *Confusion Matrix* Klasifikasi Data Latih dengan Skema 1

Sesuai dengan cara membaca tabel *confusion matrix* dimana baris sebagai label aktual dan kolom sebagai label prediksi, maka diperoleh hasil sebagai berikut.

- sebanyak 116 data bergenre 0 atau *Romance* diklasifikasikan dengan **benar** sebanyak 116 data sebagai genre *Romance*
- sebanyak 194 data bergenre 1 atau *Fantasy* diklasifikasikan dengan **benar** sebanyak 194 data sebagai genre *Fantasy*
- sebanyak 1 data bergenre 2 atau *Drama* diklasifikasikan dengan **salah** sebanyak 1 data sebagai genre *Romance*.
- sebanyak 173 data bergenre 2 atau *Drama* diklasifikasikan dengan **benar** sebanyak 173 data sebagai genre *Drama*.

b) *Performance*

	precision	recall	f1-score	support
0	0.99	1.00	1.00	116
1	1.00	1.00	1.00	194
2	1.00	0.99	1.00	174
accuracy			1.00	484
macro avg	1.00	1.00	1.00	484
weighted avg	1.00	1.00	1.00	484

Gambar 22. Output *Performance* Klasifikasi Data Latih dengan Skema 1

Hasil ini memperlihatkan nilai akurasi, nilai presisi, dan f1-score yang diperoleh dari klasifikasi genre Webtoon dengan algoritma KNN pada data latih adalah 100%.

2) Data uji

a) *Confusion Matrix*

[[6	0	3]
[0	6	0]
[0	0	11]]

Gambar 23. Output *Confusion Matrix* Klasifikasi Data Uji dengan Skema 1

Sesuai dengan cara membaca tabel *confusion matrix* dimana baris sebagai label aktual dan kolom sebagai label prediksi, maka diperoleh hasil sebagai berikut.

- terdapat 6 data bergenre 0 atau *Romance*, diklasifikasi dengan **benar** sebanyak 6 pada genre *Romance*.
- terdapat 3 data bergenre 0 atau *Romance*, diklasifikasi **salah** sebanyak 3 pada genre *Drama*.
- terdapat 6 data bergenre 1 atau *Fantasy*, diklasifikasi dengan **benar** sebanyak 6 pada genre *Fantasy*.
- terdapat 11 data bergenre genre 2 atau *Drama*, diklasifikasi dengan **benar** sebanyak 11 pada genre *Drama*.

b) *Performance*

	precision	recall	f1-score	support
0	1.00	0.67	0.80	9
1	1.00	1.00	1.00	6
2	0.79	1.00	0.88	11
accuracy			0.88	26
macro avg	0.93	0.89	0.89	26
weighted avg	0.91	0.88	0.88	26

Gambar 24. *Output Performance* Klasifikasi Data Uji dengan Skema 1

Hasil ini memperlihatkan nilai akurasi, nilai presisi, dan f1-score yang diperoleh dari klasifikasi genre Webtoon dengan algoritma KNN pada data latih sebagai berikut.

- Sebesar 100% data yang benar bergenre 0 atau *Romance* dari seluruh data yang diprediksi bergenre 0 atau *Romance*.
- Sebesar 100% data yang benar bergenre 1 atau *Fantasy* dari seluruh data yang diprediksi bergenre 1 atau *Fantasy*.
- Sebesar 79% data yang benar bergenre 2 atau *Drama* dari seluruh data yang diprediksi bergenre 2 atau *Drama*.
- Sebesar 67 % data yang diprediksi bergenre 0 atau *Romance* dari seluruh data yang sebenarnya bergenre 0 atau *Romance*.
- Sebesar 100 % data yang diprediksi bergenre 1 atau *Fantasy* dari seluruh data yang sebenarnya bergenre 1 atau *Fantasy*.
- Sebesar 100 % data yang diprediksi bergenre 2 atau *Drama* dari seluruh data yang sebenarnya bergenre 2 atau *Drama*.
- Nilai F1-score pada semua genre menunjukkan hasil yang cukup baik yaitu 0.80, 1.00, dan 0.88. hal ini mengindikasikan bahwa model klasifikasi KNN memiliki nilai presisi dan recall yang cukup baik pula.
- Nilai akurasi dari klasifikasi data uji dengan skema 1 adalah sebesar 88%.

b. Evaluasi dan Interpretasi klasifikasi genre Webtoon dengan algoritma KNN dan PSO (Skema 2).

1) Data latih

a) *Confusion Matrix*

[[116	0	0]
[0	194	0]
[1	0	173]]

Gambar 25. *Output Confusion Matrix* Klasifikasi Data Latih dengan Skema 2
 Sesuai dengan cara membaca tabel *confusion matrix* dimana baris sebagai label aktual dan kolom sebagai label prediksi, maka diperoleh hasil sebagai berikut.

- sebanyak 116 data bergenre 0 atau *Romance* diklasifikasikan dengan **benar** sebanyak 116 data sebagai genre *Romance*
- sebanyak 194 data bergenre 1 atau *Fantasy* diklasifikasikan dengan **benar** sebanyak 194 data sebagai genre *Fantasy*
- sebanyak 1 data bergenre 2 atau *Drama* diklasifikasikan dengan **salah** sebanyak 1 data sebagai genre *Romance*
- sebanyak 173 data bergenre 2 atau *Drama* diklasifikasikan dengan **benar** sebanyak 173 data sebagai genre *Drama*.

b) *Performance*

	precision	recall	f1-score	support
0	0.99	1.00	1.00	116
1	1.00	1.00	1.00	194
2	1.00	0.99	1.00	174
accuracy			1.00	484
macro avg	1.00	1.00	1.00	484
weighted avg	1.00	1.00	1.00	484

Gambar 26. *Output Performance* Klasifikasi Data Latih dengan Skema 2

Hasil ini memperlihatkan nilai akurasi, nilai presisi, dan f1-score yang diperoleh dari klasifikasi genre Webtoon dengan algoritma KNN dan PSO pada data latih adalah 100%.

2) Data uji

a) *Confusion Matrix*

[[9	0	0]
[0	6	0]
[0	0	11]]

Gambar 27. *Output Confusion Matrix* Klasifikasi Data Uji dengan Skema 2

Sesuai dengan cara membaca tabel *confusion matrix* dimana baris sebagai label aktual dan kolom sebagai label prediksi, maka diperoleh hasil sebagai berikut.

- terdapat 9 data bergenre 0 atau *Romance*, diklasifikasi dengan **benar** pada genre 0 atau *Romance*.

- terdapat 6 data bergenre 1 atau *Fantasy*, diklasifikasi dengan **benar** pada 1 atau *Fantasy*.
- terdapat 11 data bergenre genre 2 atau *Drama*, diklasifikasi dengan **benar** pada genre 2 atau *Drama*.

b) *Performance*

	precision	recall	f1-score	support
0	1.00	1.00	1.00	9
1	1.00	1.00	1.00	6
2	1.00	1.00	1.00	11
accuracy			1.00	26
macro avg	1.00	1.00	1.00	26
weighted avg	1.00	1.00	1.00	26

Gambar 28. Output *Performance* Klasifikasi Data Uji dengan Skema 2

Hasil ini memperlihatkan nilai akurasi, nilai presisi, dan f1-score yang diperoleh dari klasifikasi genre Webtoon dengan algoritma KNN dan PSO pada data uji adalah 100%.

Pembahasan

1. Pembuatan model klasifikasi genre Webtoon dengan algoritma PSO KNN dimulai dari proses Pemilihan data (*data selection*), Pra-pemrosesan data (*data preprocessing*) yang terdiri dari *cleaning, case folding, filtering, lemmatization dan stemming, serta tokenizing*, kemudian dilanjutkan proses transformasi data (*data transformation*) dengan TF-IDF, seleksi fitur (*feature selection*) dengan algoritma PSO, penambahan data (*data mining*) yaitu klasifikasi dengan algoritma KNN serta interpretasi dan evaluasi hasil (*interpretation and evaluation*) dengan *confusion matrix* dan *performance*. Sebelum proses seleksi fitur, data hasil pembobotan sebelumnya telah dibagi menjadi data latih dan data uji dengan rasio 95:5. Selain itu, banyaknya iterasi yang digunakan saat proses seleksi fitur ditetapkan sebanyak 1000 kali. Kedua ketentuan ini terbukti memberikan hasil *best cost* yang cukup konsisten yaitu sebesar 0,00.
2. Nilai akurasi yang dihasilkan dari penggunaan model PSO KNN untuk klasifikasi genre Webtoon dengan berdasarkan ringkasan cerita menunjukkan nilai akurasi sebesar 100% baik pada data latih maupun data uji. Hal ini berarti bahwa model yang dibangun telah berhasil mengklasifikasi genre dengan baik dan tepat. Perbandingan nilai akurasi data latih dan data uji klasifikasi dengan algoritma KNN dapat dilihat pada gambar di bawah.

Accuracy of train data: 0.998
Accuracy of test data: 0.885

Gambar 29. Perbandingan nilai akurasi data latih dan data uji pada Skema 1

Gambar di atas menunjukkan nilai akurasi data latih dengan algoritma KNN adalah 99,8% atau 100%. Sedangkan nilai akurasi untuk data uji sebesar 88,5% atau 86%. Kemudian, perbandingan nilai akurasi data latih dan data uji klasifikasi dengan model PSO KNN dapat dilihat pada gambar di bawah.

Accuracy of train data: 0.998 Accuracy of test data: 1.000

Gambar 30. Perbandingan nilai akurasi data latih dan data uji pada Skema 2

Gambar di atas menunjukkan nilai akurasi data latih dengan algoritma PSO KNN adalah 99,8% atau 100%. Sedangkan nilai akurasi untuk data uji sebesar 100%.

Dari kedua gambar di atas, dapat dikatakan bahwa nilai akurasi yang diperoleh dengan algoritma PSO KNN lebih tinggi dari algoritma KNN saja. Hal ini membuktikan bahwa algoritma PSO mampu memperbaiki algoritma klasifikasi KNN.

Di samping itu, waktu komputasi yang digunakan dengan algoritma PSO KNN terbukti lebih singkat daripada algoritma KNN saja. Model PSO KNN membutuhkan waktu 0,266 detik untuk data latih sedangkan model KNN membutuhkan waktu 0,281 detik untuk data latih. Sedangkan pada data uji membutuhkan waktu komputasi yang sama baik dengan algoritma KNN maupun PSO KNN yaitu selama 0,0469 detik. Meskipun hasil ini selalu berubah setiap dilakukan *running*, rata-rata hasil menunjukkan bahwa waktu komputasi model PSO KNN lebih singkat dari model KNN.

SIMPULAN

Dari hasil penelitian tersebut, penulis memperoleh kesimpulan bahwa:

1. Klasifikasi genre Webtoon dengan menggunakan *Particle Swarm Optimization* (PSO) sebagai algoritma seleksi fitur dan *K-Nearest Neighbors* (KNN) sebagai algoritma klasifikasi dimulai dengan melakukan pemilihan atribut hingga pembagian dataset sebesar 95% untuk data latih (*data train*) dan sebesar 5% untuk data uji (*data test*). Semakin banyak data latih yang digunakan, maka akan semakin lengkap kamus fitur/kata yang terbentuk sehingga model yang terbentuk dapat mengklasifikasikan data uji lebih baik. Seleksi fitur dengan iterasi sebanyak 1.000 kali. Banyaknya iterasi cukup memengaruhi hasil *best cost* yang berdampak pada fitur-fitur yang akan terpilih. Dengan jumlah iterasi ini, model cukup konsisten memunculkan hasil *best cost* sebesar 0.00 dan diperoleh 1.653 kata terpilih dari 3.273 total kata. Setelah itu, dilanjutkan dengan membentuk subset baru dari fitur-fitur yang terdapat dalam data latih dan data uji berdasarkan fitur terpilih. Hasil subset ini kemudian digunakan untuk proses klasifikasi.
2. Algoritma KNN menghasilkan nilai akurasi tertinggi untuk data uji sebesar 88% dengan nilai akurasi data latihnya sebesar 99,8% atau 100%. Sedangkan Algoritma KNN setelah dilakukan seleksi fitur menggunakan PSO menghasilkan nilai akurasi sebesar 100% dengan nilai akurasi data latihnya 99,8% atau 100%. Sehingga dapat disimpulkan bahwa PSO mampu meningkatkan nilai akurasi pada algoritma klasifikasi KNN. Selain itu, Algoritma PSO juga mampu mempercepat proses klasifikasi pada data train sebesar 0,015 detik.

UCAPAN TERIMA KASIH

Terima kasih kepada koordinator Prodi Matematika dan semua Dosen Prodi Matematika yang telah memberikan ilmu dan bimbingannya sehingga artikel ini dapat diselesaikan.

DAFTAR PUSTAKA

- Adiwijaya, I. (2006). Text Mining dan Knowledge Discovery. *Kolokium Bersama Komunitas Datamining Indonesia & Soft-Computing Indonesia*, 1–9. [http://web.ipb.ac.id/~ir-lab/pdf/tm \(text summarization\).pdf](http://web.ipb.ac.id/~ir-lab/pdf/tm(text%20summarization).pdf)

- Arifin, T. (2017). *Implementasi Algoritma PSO Dan Teknik Bagging Untuk Klasifikasi Sel Pap Smear*. 4(2), 155–162.
- Chandler, D. (2004). An Introduction to Genre Theory. [Http://Www.Aber.Ac.Uk/Media/Documents/Intgenre/Chandler_genre_theory.Pdf](http://www.aber.ac.uk/media/Documents/intgenre/Chandler_genre_theory.pdf), 1–15. papers2://publication/uuid/84DD8640-73B4-4264-A37E-52AE23DED1D8
- Chantar, H. K., & Corne, D. W. (2011). Feature subset selection for Arabic document categorization using BPSO-KNN. *Proceedings of the 2011 3rd World Congress on Nature and Biologically Inspired Computing, NaBIC 2011*, 546–551. <https://doi.org/10.1109/NaBIC.2011.6089647>
- Jin, D. Y. (2015). Digital convergence of Korea's webtoons: transmedia storytelling. *Communication Research and Practice*, 1(3), 193–209. <https://doi.org/10.1080/22041451.2015.1079150>
- Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information (Switzerland)*, 10(4), 1–68. <https://doi.org/10.3390/info10040150>
- Lestari, A. F., & Irwansyah. (2020). Line Webtoon Sebagai Industri Komik Digital. *Jurnal Ilmu Komunikasi*, 6(2), 134–148. <http://180.250.41.45/jsourc/article/view/1609/1726>
- Muslimah, N., & Wihandika, R. C. (2019). *Klasifikasi Film Berdasarkan Sinopsis dengan Menggunakan Improved K-Nearest Neighbor (K-NN)*. 3(1), 196–204.
- Primartha, R. (2018). *Belajar Machine Learning Teori dan Praktik*. Informatika Bandung.
- Purwaningsih, N. S. (2017). *Analisis genre dalam karya Pramoedya Ananta Toer periode 1950-an*. 5, 1–17.
- Ramadhan, G. S., Irawan, B., Setianingsih, C., & Telkom, U. (2021). *Klasifikasi emosi pada lirik lagu menggunakan algoritma naïve bayes dan particle swarm optimization classification of emotions on song lyrics using naïve bayes algorithm and particle swarm optimization*. 8(5), 6293–6306.
- Rizky, M. Y., & Stellarosa, Y. (2017). *PREFERENSI PENONTON TERHADAP FILM INDONESIA*. 15–34.
- Saputra, A. C., Sitepu, A. B., Stanley, P, Y. S. P. W., Tetuko, P. G. S., & Nugroho, G. C. (2019). The Classification of the Movie Genre based on Synopsis of the Indonesian Film. *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIIT)*, 201–204.