Identifikasi Teks Gereflekter pada Buku Anak dengan Algoritma *k-Nearest Neighbor*

I Kadek Ananda Prana Widya[1], Adiwijaya[2], Widi Astuti[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]ikadekanandapw@students.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id,
[3]widiwdu@telkomuniversity.ac.id

Abstrak

Buku anak merupakan salah satu sumber pengetahuan bagi pembaca, khususnya anak. Ketika buku itu dibaca, anak akan berusaha memaknai setiap kata dan kalimat di dalamnya. Terdapat permasalahan ketika ditemukan kesalahan konten pada buku tersebut. Konten yang dimaksud yaitu kata maupun kalimat yang memiliki makna kurang sopan, seksual, serta kata kasar. Bagi anak-anak di tingkat sekolah dasar konten tersebut menjadi hal yang bermakna gereflekter (tabu). Berdasarkan permasalah tersebut, maka dilakukan penelitian tugas akhir terhadap cerita anak yang diambil dari buku fiksi dan buku pelajaran. Penelitian ini dilakukan dengan membangun sistem untuk mendeteksi konten gereflekter pada teks cerita yang dijadikan sebagai *dataset*. Penelitian dilakukan dengan membangun model menggunakan algoritma klasifikasi teks *k-Nearest Neighbor* dengan pendekatan *distance measure*. *Distance measure* yang digunakan adalah *Euclidean Distance* dan *Manhattan Distance*. Sistem dievaluasi dengan menggunakan *precision*, *recall*, dan F1 *score*. Berdasarkan hasil evaluasi, skenario pengujian menggunakan *Euclidean distance* dan *Manhattan distance* mendapatkan nilai *precision* 0.915, *recall* 0.845, dan F1 *score* 0.895.

Kata kunci : buku anak, *distance measure*, gereflekter, *k-Nearest Neighbor*

Abstract

Children's books are one source of knowledge for readers, especially children. When the book is read, the child will try to make sense of every word and sentence in it. There was a problem when a content error was found in the book. The content in question is words and sentences that have meanings that are not polite, sexual, and rude words. For children at the elementary school level, the content becomes meaningful reflectivity (taboo). Based on these problems, a final assignment research was carried out on children's stories taken from fiction books and textbooks. This research was conducted by building a system for detecting reflector content on story text that is used as a dataset. The study was conducted by building a model using the k-Nearest Neighbor text classification algorithm with a distance measure approach. Distance measure used is Euclidean Distance and Manhattan Distance. The system is evaluated using precision, recall, and F1 score. Based on the evaluation results, testing scenarios using Euclidean distance and Manhattan distance get a precision value of 0.915, recall 0.845, and F1 score 0.895.

Keywords: children's book, distance measure, gereflekter, k-Nearest Neighbor

1. Pendahuluan

Latar Belakang

Dunia pendidikan didukung dengan fasilitas yang menunjang pendidikan tersebut. Adapun fasilitas yang digunakan adalah buku pelajaran, salah satunya buku anak. Buku anak merupakan buku yang digunakan dalam mendukung kegiatan belajar yang berisi uraian mengenai materi tertentu yang disusun secara sistematis dengan tujuan tertentu [1].

Anak memiliki tingkat keingintahuan yang tinggi, sehingga akan menyerap informasi yang didapatkan. Informasi diperoleh dari bahan bacaan, seperti buku anak. Anak akan berusaha memahami serta memaknai setiap kata dan kalimat yang terdapat dalam buku anak tersebut. Apabila buku anak memiliki konten yang tidak sesuai dan vulgar, akan berpengaruh kepada pembaca, khususnya anak. Pada forum diskusi *online*, ditemukan bahwa terdapat buku anak pada tahun 2013 dan 2014 memuat konten yang tidak sesuai untuk anak. Sebagai contoh, pada Juli 2013, ditemukan cerita yang memuat cerita tak senonoh dalam buku pelajaran Bahasa Indonesia untuk SD dan MI Kelas VI di halaman 57-60 [2].

Berdasarkan permasalahan tersebut, penelitian tugas akhir ini merancang sistem pengklasifikasi teks pada buku anak yang diambil dari beberapa buku cerita dan buku pelajaran tingkat sekolah dasar (SD) tahun 2011 hingga 2019 (khususnya buku pelajaran tematik Kurikulum 2013 revisi tahun 2017). Teks pada buku anak dipecah menjadi potongan paragraf yang digunakan sebagai *dataset* dan dimasukkan kedalam kelas kata yang mengandung makna gereflekter maupun non-gereflekter. Makna gereflekter merupakan makna bersifat tabu yang muncul akibat reaksi seseorang terhadap makna yang lain, dan berhubungan dengan seksual, kepercayaan,

serta kebiasaan [3]. Dalam penelitian ini, bagi anak SD adalah konten dewasa berupa kata-kata vulgar dan atau cerita tak senonoh.

Algoritma klasifikasi yang digunakan adalah algoritma *k-Nearest Neighbor*. Algoritma *k-Nearest Neighbor* (k-NN) merupakan pendekatan untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut [4]. Algoritma ini menemukan sejumlah *k* objek yang paling dekat dengan titik pada data [5]. Untuk menentukan jarak pada tiap objek data, akan menggunakan *distance measure* yang menentukan akurasi jarak antar data terbaik. Dalam tugas akhir ini menggunakan algoritma k-NN dengan penentuan *distance measure* terdiri dari *euclidean distance* dan *manhattan distance* yang ditujukan untuk mengklasifikasikan atau mengolah kalimat sehingga akan diketahui kalimat apakah mengandung makna gereflekter atau non-gereflekter pada buku anak. Dengan menggunakan metode ini, diharapkan mendapatkan tingkat akurasi yang cukup tinggi dalam mengklasifikasikan kalimat tersebut.

Topik dan Batasannya

Rumusan masalah pada penelitian tugas akhir ini adalah teks cerita anak SD sejak tahun 2011 ditemukan bermakna gereflekter. Makna gereflekter merupakan makna yang berhubungan dengan kata atau ungkapan yang tabu berupa kata-kata vulgar dan atau cerita tak senonoh. Dari permasalahan tersebut, pertanyaan penelitian pada penelitian ini adalah: bagaimana hasil pengujian sistem yang dibangun dengan menggunakan *Euclidean distance* dan *Manhattan distance* dalam melakukan klasifikasi teks.

Berdasarkan latar belakang di atas, pada tugas akhir ini, telah dilakukan studi dan implementasi sebuah metode untuk mendeteksi adanya konten gereflekter pada teks cerita anak-anak. Metode klasifikasi yang digunakan adalah *k-Nearest Neighbor*.

Batasan masalah dalam penelitian tugas akhir ini adalah teks cerita yang dijadikan *dataset* adalah teks berbahasa Indonesia untuk anak SD (berumur 6 – 12 tahun). Pada penelitian ini, teks cerita yang mengandung makna gereflekter relatif sedikit sehingga dilakukan pemecahan teks cerita menjadi potongan paragraf dan juga dianggap bahwa beberapa paragraf pada teks cerita tersebut sudah dapat merefleksikan adanya gereflekter. Maka dari itu dataset adalah potongan paragraf dari teks cerita yang diambil dari buku cerita dan buku pelajaran yang paling banyak memuat teks cerita wacana yaitu buku pelajaran tematik Kurikulum 2013 revisi tahun 2017 kelas IV tema 8 dengan judul Daerah Tempat Tinggalku.

Tujuan

Tujuan dari penelitian tugas akhir ini yaitu mendapatkan hasil pengujian sistem yang dibangun dengan menggunakan *Euclidean distance* dan *Manhattan distance*.

2. Studi Terkait

Makna Gereflekter

Makna gereflekter (Belanda: *gereflecteerde betekenis*) adalah makna yang muncul dalam hal makna konseptual yang jamak akibat reaksi kita terhadap makna lain (Leech, I, 1974:33-35) [3]. Tidak saja muncul karena sugesti emosional, tetapi makna gereflekter juga berhubungan dengan kata atau ungkapan yang tidak boleh diucapkan atau tabu seperti kata-kata bersetubuh, ereksi, ejakulasi [3]. Beberapa contoh kata sebelumnya merupakan hal yang tabu dalam masyarakat Indonesia untuk digunakan dalam percakapan sehari-hari, terlebih lagi pada pada anak yang sedang duduk di bangku sekolah dasar (umur 6-12 tahun). Dengan demikian, dalam penelitian tugas akhir ini, konten berupa kata kata vulgar dan kalimat tak senonoh dianggap sebagai makna gereflekter bagi anak-anak.

Klasifikasi Teks

Klasifikasi teks adalah sebuah pekerjaan untuk menentukan apakah sebuah dokumen adalah milik dari sebuah kategori yang telah ditentukan sebelumnya [6]. Pada klasifikasi teks, data yang berupa teks akan dikonversikan menjadi bobot yang siap untuk digunakan. Tahapan dalam klasifikasi teks antara lain sebagai berikut [7]:

a. *Preprocessing.*
   Merupakan tahapan untuk merpresentasikan dokumen dalam bentuk fitur vektor, yang berarti harus memisahkan teks menjadi kata terpisah. Dalam tahap *preprocessing*, dilakukan penghapusan *stopwords* pada dokumen, dengan tujuan untuk menghapus kata-kata umum dan tak bermakna yang disesuaikan dengan kosakata bahasa yang digunakan. Setelah *stopwords* dihapus, dilakukan tahapan *stemming* yang digunakan untuk mencari kata dasar dari kata yang telah diekstraksi dari dokumen.
b. Rekayasa Fitur
   Pada tahap ini merupakan tahapan latih yang terdiri dari tahapan seleksi fitur, *dictionary construction*, dan *feature weighting*. Tujuan dari rekayasa fitur adalah untuk menghapus semua fitur yang tidak relevan dan selalu muncul pada semua dokumen.

c. Generasi Model Klasifikasi
Tahap ini merupakan tahap untuk membangun algoritma klasifikasi, dalam penelitian ini menggunakan metode *k-Nearest Neighbor* (k-NN) atau juga melakukan generasi model classifier berdasarkan hasil pelatihan oleh dokumen sebelumnya yang akan digunakan untuk mengklasifikasikan dokumen yang tidak diketahui kategorinya.

d. Pengkategorian Dokumen
Merupakan tahapan untuk melakukan klasifikasi dari dokumen baru yang tidak diketahui asal kategori dari dokumen tersebut, dengan catatan bahwa dokumen baru tersebut telah melewati tahap *preprocessing* dan *feature weighting*.

K-Nearest Neighbor

Metode *k-Nearest Neighbor* (k-NN) merupakan algoritma yang mengkategorikan objek berdasarkan ruang fitur terdekat pada himpunan latih. Himpunan latih dipetakan dalam ruang fitur yang multidimensi [8]. Ruang fitur terbagi dengan basis area sesuai dengan kategori himpunan latih. Titik dari ruang fitur ditetapkan sebagai bagian dari kategori apabila titik tersebut sering dekat dengan kategori tertentu pada data latih dengan *k* terdekat [9].

Adapun langkah yang diterapkan oleh *k-Nearest Neighbour* adalah sebagai berikut [10]:

1. Muat data
2. Inisialisasi K untuk jumlah tetangga pilihan Anda
3. Untuk setiap contoh dalam data
   a. Hitung jarak antara contoh kueri dan contoh saat ini dari data.
   b. Tambahkan jarak dan indeks contoh ke koleksi yang dipesan
4. Urutkan koleksi jarak dan indeks yang terurut dari terkecil ke terbesar (dalam urutan menaik) berdasarkan jarak
5. Pilih entri K pertama dari koleksi yang diurutkan
6. Dapatkan label entri K yang dipilih
7. Jika regresi, kembalikan rata-rata label K
8. Jika klasifikasi, kembalikan mode label K

Distance Measure

Pengelompokkan data diukur dengan menentukan dua objek mirip atau tidak mirip. Untuk menentukan kemiripan tersebut dapat digunakan pengukuran yang disebut dengan *distance measure*. Berikut merupakan jenis-jenis pada *distance measure* [11].

a. *Euclidean distance*

$$D_{L2}(X_2, X_1) = \|X_2 - X_1\|_2 = \sqrt{\sum_{j=1}^{p} (X_{2j} - X_{1j})^2} \quad (1)$$

Keterangan: 0
p = Dimensi Data
$X1$ = Posisi titik 1
$X2$ = Posisi titik 2

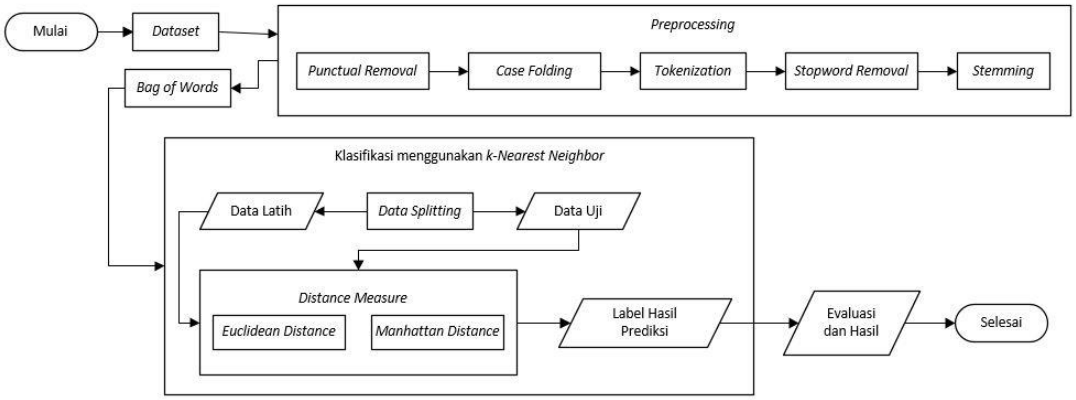
b. *Manhattan distance*

$$D_{L1}(X_2, X_1) = \|X_2 - X_1\|_{11} = \sum_{j=1}^{p} |x_{2j} - x_{1j}| \quad (2)$$

Keterangan:
p = Dimensi Data
$X1$ = Posisi titik 1
$X2$ = Posisi titik 2
|| = nilai absolut

3. Sistem yang Dibangun

Sistem yang dibangun berfungsi untuk mengklasifikasikan makna gereflekter pada teks cerita yang dimuat di dalam buku cerita anak dan buku pelajaran SD dari berbagai sumber sejak tahun 2011 hingga buku pelajaran tematik SD Kurikulum 2013 revisi tahun 2017. Sistem ini mengelompokkan potongan paragraf (yang dijadikan sebagai dataset) ke dalam dua kelas, yaitu kelas positif (gereflekter) dan kelas negatif (non-gereflekter) dengan menggunakan algoritma *k-Nearest Neighbor* sebagai teknik klasifikasinya. Alur kerja sistem yang telah dibangun digambarkan pada Gambar 1.



Gambar 1 Sistem yang Dibangun

3.1 *Dataset*

Pada penelitian tugas akhir ini, *dataset* adalah potongan paragraf yang diambil dari teks cerita wacana berbahasa Indonesia, seperti cerita pendek dan dongeng. Teks cerita tersebut diambil dalam beberapa buku pelajaran SD dan buku cerita anak dari beberapa sumber. Hal pertama yang dilakukan adalah mengumpulkan berita dari beberapa situs resmi berita daring yang menerbitkan berita tentang beredarnya buku pelajaran maupun buku bacaan anak berisi teks cerita yang memuat konten gereflekter sehingga meresahkan para orang tua. Konten gereflekter pada penelitian ini adalah kata-kata vulgar dan atau cerita tak senonoh.

Terkumpul sebanyak 12 cerita yang memuat konten gereflekter dari berbagai macam sumber, namun hanya enam cerita yang memenuhi batasan masalah yang digunakan dalam tugas akhir ini. Teks cerita yang mengandung makna gereflekter relatif sedikit sehingga mengakibatkan terjadinya *imbalance data*. Rincian buku yang digunakan adalah sebagai berikut, dua buah buku pelajaran tahun 2011, satu buah buku pelajaran tahun 2013, satu buah buku pelajaran tahun 2015, dua buah buku cerita anak tahun 2017, dan buku pelajaran tematik Kurikulum 2013 revisi tahun 2017 kelas IV tema 8 dengan judul Daerah Tempat Tinggalku.

Setelah dipecah menjadi potongan paragraf, *dataset* untuk kelas gereflekter sebanyak 14 data. Keempat belas data tersebut dianggap sudah dapat merefleksikan adanya kelas gereflekter. Sedangkan untuk kelas non-gereflekter, sebanyak 107 buah data potongan paragraf telah diambil dari teks cerita wacana pada buku pelajaran tematik Kurikulum 2013 revisi tahun 2017 kelas IV tema 8 dengan judul Daerah Tempat Tinggalku. *Dataset* yang digunakan dapat diakses oleh publik dan tersedia *online*[1].

3.2 *Preprocessing*

Preprocessing merupakan tahapan yang umum dalam melakukan klasifikasi terhadap teks [12]. Tahap ini bertujuan untuk mengolah data yang tadinya hanya berupa teks menjadi data yang siap untuk diklasifikasikan. *Preprocessing* yang digunakan pada penelitian ini adalah *punctual removal*, *case folding*, *tokenization*, *stopword removal* dan *stemming*.

Punctual removal adalah proses untuk menghilangkan tanda baca. Pada proses ini setiap karakter selain angka dan huruf akan dihapus. Misalnya pada kalimat “Kasuari adalah hewan yang diminati setiap orang.” menjadi “Kasuari adalah hewan yang diminati setiap orang”.

Case folding adalah proses untuk mengubah setiap huruf besar menjadi huruf kecil. Misalnya pada kalimat “Kasuari adalah hewan yang diminati setiap orang” menjadi “kasuari adalah hewan yang diminati setiap orang”.

Tokenization adalah proses untuk mengubah kalimat menjadi token atau kata. Misalnya pada kalimat “kasuari adalah hewan yang diminati setiap orang” menjadi [‘kasuari’, ’adalah’, ‘hewan’, ’yang’, ’diminati’, ’setiap’, ‘orang’].

[1] https://drive.google.com/drive/folders/1MFFT52043TjZFeB9QmbMdxgt193iuzt8?usp=sharing

Stopword removal adalah proses untuk menghapus kata yang dianggap tidak memiliki pengaruh pada suatu kalimat. Misalnya pada [‘kasuari’, ’adalah’, ‘hewan’, ’yang’, ’diminati’, ’setiap’, ‘orang’] menjadi [‘kasuari’, ‘hewan’,’diminati’, ’setiap’, ‘orang’].

Stemming adalah proses untuk mengubah kata dengan imbuhan menjadi kata dasar. Misalnya pada [‘kasuari’, ‘hewan’,’diminati’, ’setiap’, ‘orang’] menjadi [‘kasuari’, ‘hewan’,’minat’, ’setiap’, ‘orang’].

3.3 *Bag of Words*

Bag of words (BoW) merupakan sebuah cara dalam *Natural Language Processing* (NLP) yang digunakan untuk mendapatkan ekstrasi fitur dari data teks dan dapat digunakan dalam proses pembelajaran mesin [13]. BoW diterapkan agar data teks memiliki nilai berupa integer dan dapat diterapkan pada proses algoritma setelahnya .

Contoh dari BoW adalah sebagai berikut. Diasumsikan terdapat tiga dokumen (D1, D2, dan D3) dengan data teks sebagai berikut.

D1: “Aku senang sekali ayah dapat mengunjungiku.”
D2: “Hewan itu besar sekali!”
D3: “Aku sangat pandai.”

Berdasarkan dokumen tersebut, daftar kosa kata dapat dikonstruksikan dengan kata-kata yang berbeda sebagai berikut.

{
“Aku”: 1
“senang”: 2
“sekali”: 3
“ayah”: 4
“dapat”: 5
“mengunjungiku”: 6
“hewan”: 7
“itu”: 8
“besar”: 9
“sangat”: 10
“pandai”: 11
}

Dokumen akan direpresentasikan kedalam 11 vektor.
D1 = [1 1 1 1 1 1 0 0 0 0 0]
D2 = [0 0 1 0 0 0 1 1 1 0 0]
D3 = [1 0 0 0 0 0 0 0 0 1 1]

Vektor yang didapatkan, akan direpresentasikan kedalam gambar dibawah ini.

	0	1	2	3	4	5	6	7	8	9	10
0	3	2	1	3	2	2	2	1	2	2	1
1	0	0	0	0	0	0	0	1	0	0	0
2	0	3	0	7	3	2	4	2	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0
4	2	3	1	4	1	1	1	0	0	0	0
5	0	7	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	0
7	2	0	0	0	0	0	0	3	0	1	2
8	0	0	0	0	0	0	0	1	0	0	0
9	0	0	0	0	0	0	0	2	0	0	0
10	0	0	0	0	0	0	0	0	0	1	1

Gambar 2 *Bag of Words*

Pada gambar diatas, kolom menandakan dokumen yang digunakan, dan baris menandakan jumlah kata yang terdapat pada masing-masing dokumen. Seperti pada dokumen indeks ke-0 (dokumen pertama), memuat kata pertama sebanyak 3 kata, kata kedua sebanyak 2 kata, dan seterusnya.

3.4 Klasifikasi menggunakan *k-Nearest Neighbor*

Sebelum melakukan proses klasifikasi, data terlebih dahulu dibagi menjadi data latih dan data uji. Pembagian data dilakukan dengan perbandingan 80% pada data latih dan 20% pada data uji. Setelah melakukan pembangian data, dilanjutkan dengan penerapan algoritma *k-Nearest Neighbor*.

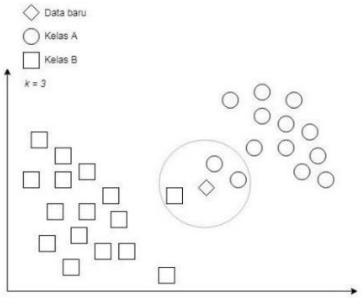
Penerapan algoritma *k-Nearest Neighbor* menggunakan *distance measure* seperti *Euclidean distance* pada persamaan (1) dan *Manhattan distance* pada persamaan (2).

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} \quad (1)$$

$$d(a,b) = |a_n - b_n| \quad (2)$$

Pada persamaan (1) dan (2), a dan b adalah data teks yang digunakan. Pada persamaan tersebut a dan b adalah data uji dan data latih. Sedangkan $a1$ sampai an merupakan fitur dari data uji begitu juga dengan $b1$ sampai bn yang juga merupakan fitur dari data latih.

Setelah mendapatkan jarak kedekatan antar data, maka dilakukan k terdekat dengan data uji yang sedang diklasifikasi. Setelah data sebanyak nilai k terdekat didapatkan, dilakukan penghitungan terhadap kelas pada sebanyak *k* data yang didapatkan dan kelas terbanyak menjadi hasil klasifikasi dari data tersebut. Ilustrasi klasifikasi menggunakan k-NN dengan nilai *k* = 3 ditunjukan pada gambar 3 berikut.



Gambar 3 Klasifikasi *k-Nearest Neighbor*

Penelitian ini menggunakan nilai *k* dimulai dari 3 sampai mendapatkan nilai *k* yang memiliki nilai konstant. Hasil dari percobaan ini, nilai *k* = 24 memiliki nilai yang konstan. Maka penelitian menggunakan nilai *k* dari 3 – 24. Adapun contoh penerapan *k-Nearest Neighbor* terdapat pada tabel 2.

Tabel 1 Contoh Penerapan *k-Nearest Neighbor*

Dokumen	Label
D1	0
D2	0
D3	1
D4	0
D5	?

Berdasarkan dokumen diatas, terdapat dokumen D1 – D4 yang telah memiliki label dan D5 yang belum memiliki label. Seperti yang telah disebutkan sebelumnya, pada proses klasifikasi akan dihitung *distance measure* dari D5 keseluruh data.

Setelah *distance measure* dihitung, maka diambil sebanyak *k* data terdekat. Pada ilustrasi diatas nilai *k* yang digunakan adalah 3. Berdasarkan ilustrasi penghitungan diatas, tiga data terdekat dengan D5 adalah D2, D3 dan D4. Selanjutnya akan dihitung label terbanyak, berdasarkan contoh diatas terdapat dua label 0 dan satu label 1 sehingga D5 akan diklasifikasi sebagai label 0.

3.5 Evaluasi dan Hasil

Setelah mendapatkan label klasifikasi baru, maka dilakukan proses prediksi label serta akurasi pada sistem yang dibangun. Untuk mengevaluasi akurasi yang dihasilkan, menggunakan penerapan *precision*, *recall*, dan *F1 score*. Evaluasi tersebut akan mendapatkan kesimpulan serta saran pada penelitian tugas akhir ini.

4. Evaluasi

4.1 Hasil Pengujian

4.1.1 Skenario Pertama menggunakan *Euclidean distance*

Pada skenario pertama menggunakan *Euclidean distance* sebagai *distance measure*. Digunakan nilai *k* dari 3 hingga 24 agar mendapatkan akurasi dari sistem. Pada skenario ini juga menerapkan *precision*, *recall*, dan *F1 Score* untuk mengevaluasi sistem yang dibangun. Tabel 3 akan menjelaskan hasil skenario pertama, dengan penulisan tebal merupakan nilai terbaik.

Tabel 2 Skenario Pertama menggunakan *Euclidean distance*

Nilai *k*	*Euclidean distance*			
	Akurasi	***Precision***	***Recall***	***F1 Score***
3	0.66	0.87	0.66	0.74
4	0.66	0.87	0.66	0.74
5	0.54	0.92	0.54	0.63
6	0.54	0.92	0.54	0.63
7	0.41	0.92	0.41	0.50
8	0.50	0.92	0.50	0.59
9	0.41	0.92	0.41	0.50
10	0.41	0.92	0.41	0.50
11	0.33	0.92	0.33	0.40
12	0.37	0.92	0.46	0.37
13	0.33	0.92	0.33	0.40
14	0.33	0.92	0.33	0.40
15	0.25	0.92	0.25	0.29
16	0.29	0.92	0.29	0.35
17	0.25	0.92	0.25	0.29
18	0.41	0.92	0.41	0.50
19	0.33	**0.93**	0.33	0.40
20	0.62	**0.94**	0.62	0.70
21	0.62	0.93	0.62	0.70
22	0.79	0.93	0.79	0.83
23	**0.70**	0.93	**0.70**	**0.77**
24	**0.92**	0.84	**0.92**	**0.88**
Rata-rata	0.48	0.91	0.48	0.55

4.1.2 Skenario Kedua menggunakan *Manhattan distance*

Pada skenario kedua menggunakan *Manhattan distance* sebagai *distance measure*. Digunakan nilai *k* dari 3 hingga 24 agar mendapatkan akurasi dari sistem. Pada skenario ini juga menerapkan *precision*, *recall*, dan *F1 score* untuk mengevaluasi sistem yang dibangun. Tabel 4 akan menjelaskan hasil skenario pertama, dengan penulisan tebal merupakan nilai terbaik.

Tabel 3 Skenario Kedua menggunakan *Manhattan distance*

Nilai *k*	*Manhattan distance*			
	Akurasi	***Precision***	***Recall***	***F1 Score***
3	0.37	0.92	0.37	0.46
4	0.41	0.92	0.41	0.50
5	0.29	0.92	0.29	0.35
6	0.29	0.92	0.29	0.35
7	0.25	0.92	0.25	0.29
8	0.29	0.92	0.29	0.35
9	0.21	0.92	0.21	0.23
10	0.25	0.92	0.29	0.25
11	0.16	0.92	0.16	0.16
12	0.25	0.92	0.25	0.29
13	0.16	0.92	0.16	0.16
14	0.29	0.92	0.29	0.35
15	0.20	0.92	0.20	0.23

Nilai *k*	*Manhattan distance*			
	Akurasi	***Precision***	***Recall***	***F1 Score***
16	0.29	0.92	0.29	0.35
17	0.16	0.92	0.16	0.16
18	0.66	0.93	0.66	0.74
19	0.54	0.92	0.54	0.63
20	0.83	**0.94**	**0.83**	0.86
21	**0.83**	**0.94**	0.83	**0.86**
22	0.83	0.89	0.83	0.85
23	0.83	0.94	0.83	0.86
24	**0.91**	0.84	**0.91**	**0.87**
Rata-rata	0.42	0.91	0.42	0.46

Masing-masing hasil terbaik pada skenario pengujian yang telah disebutkan diatas, akan dihitung rata-rata nilai pada masing-masing *distance measure*. Hasil rata-rata dapat dilihat pada tabel 5 berikut, nilai dengan penulisan tebal merupakan nilai terbaik.

Tabel 4 Hasil Akhir pada Pengujian Sistem

Skenario ke-	Nilai *k*	*Distance measure*	Akurasi	*Precision*	*Recall*	*F1 Score*	Rata-rata
1	19	*Euclidean distance*	0.33	**0.93**	0.33	0.40	0.49
1	20	*Euclidean distance*	0.62	**0.94**	0.62	0.70	0.72
2	20	*Manhattan distance*	0.83	**0.94**	**0.83**	0.86	0.86
2	21	*Manhattan distance*	**0.83**	**0.94**	0.83	**0.86**	0.86
1	23	*Euclidean distance*	**0.70**	0.93	**0.70**	**0.77**	0.75
1	24	*Euclidean distance*	**0.92**	0.84	**0.92**	**0.88**	**0.89**
2	24	*Manhattan distance*	**0.91**	0.84	**0.91**	**0.87**	**0.88**

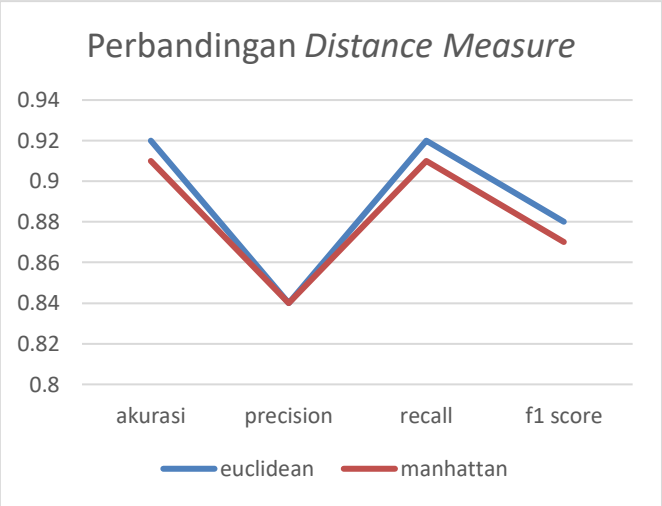
4.2 Analisis Hasil Pengujian

Pada skenario pertama dengan menggunakan *Euclidean distance*¸ nilai *k* mengalami peningkatan nilai pada *k* ke-19 hingga ke-24, baik dari nilai akurasi, *precision*, *recall*, maupun F1 *Score*. Pada skenario kedua dengan menggunakan *Manhattan distance*¸ nilai *k* mengalami peningkatan nilai pada *k* ke-20 hingga ke-24, baik dari nilai akurasi, *precision*, *recall*, maupun F1 *Score*.

Skenario pertama memiliki kemiripan dalam nilai akurasi, *precision*, *recall*, serta *F1 score*. Dalam melakukan perubahan pada nilai *k* mulai dari 19, nilai evaluasi mengalami peningkatan. Hal tersebut terjadi karena semakin besar nilai *k* yang dimasukkan, maka rentang jarak antar data mengalami perlebaran. Sehingga jangkauan data yang belum memiliki label menjadi lebih luas. Karena luasnya jangkauan data tersebut, maka menyebabkan nilai evaluasi meningkat.

Nilai *k* yang digunakan antara nilai *k* ke-3 hingga ke-24. Apabila menggunakan nilai *k* ganjil, maka sistem akan mendeteksi kelas yang memiliki jumlah paling tinggi. Seperti nilai *k* = 3, maka sistem mendeteksi kelas negatif = 2 serta kelas positif = 1. Maka pada data baru akan dikelompokkan menjadi kelas negatif. Apabila menggunakan nilai *k* = 4, maka sistem dapat mendeteksi kelas negatif = 2 serta kelas positif = 2. Hal ini menyebabkan data baru tidak diklasifikasikan secara langsung. Bahasa pemrograman *Python* mengambil data pada indeks paling atas. Sehingga, sistem mengikuti kelas teratas.

Adapun perbandingan nilai evaluasi antara *Euclidean distance* dan *Manhattan distance* dapat dilihat pada gambar 4.



Gambar 4 Perbandingan *Distance Measure*

Euclidean dan *Manhattan distance* dapat digunakan mencari kedekatan antar data. Tingkat dimensi pada data mempengaruhi jarak yang dihasilkan. Pada gambar diatas, terlihat perbandingan antara *Euclidean* dan *Manhattan* memiliki perbedaan nilai yang sedikit. Namun, *Euclidean* memiliki nilai yang lebih tinggi daripada *Manhattan*. Hal ini dikarenakan *Euclidean* melakukan perhitungan dengan dimensi data yang lebih banyak.

5. Kesimpulan

Distance measure seperti *Euclidean distance* dan *Manhattan distance* dapat digunakan untuk mengukur tingkat kedekatan antar data. Pada skenario pertama dengan menggunakan *Euclidean distance*¸ nilai *k* optimal pada nilai ke-19 hingga ke-24. Pada saat nilai *k* > 24, nilai akurasi menjadi tetap. Nilai akurasi yang didapat yaitu 0.92, nilai *precision* 0.84, nilai *recall* 0.92, dan nilai F1 *score* 0.88.

Pada skenario kedua dengan menggunakan *Manhattan distance*, nilai *k* optimal pada nilai ke-20 hingga ke-24. Pada saat nilai *k* > *24*, nilai akurasi menjadi tetap. Nilai akurasi yang didapat yaitu 0.91, nilai *precision* 0.84, nilai *recall* 0.91, dan nilai F1 *score* 0.87.

Klasifikasi gereflekter pada buku anak akan memberikan hasil yang lebih baik apabila penelitian yang akan dilakukan selanjutnya dapat menambahkan *dataset* yang mengandung makna gereflekter sehingga rasio kelas positif dan kelas negatif menjadi lebih seimbang.

Daftar Pustaka

[1] G. Rahmawati, "Buku Teks Pelajaran Sebagai Sumber Belajar Siswa Di Perpustakaan Sekolah Di Sman 3 Bandung," *Edulib*, vol. 5, no. 1, pp. 102–113, 2016.

[2] "6 Buku Pelajaran yang Pernah Bikin Geger Dunia Pendidikan Indonesia | KASKUS." [Online]. Available: https://www.kaskus.co.id/thread/5441ba4adc06bd784d8b457a/6-buku-pelajaran-yang-pernah-bikin-geger-dunia-pendidikan-pict/. [Accessed: 01-Oct-2019].

[3] M. Pateda, *Semantik Leksikal*, 2nd ed. Jakarta: PT Rineka Cipta, 2010.

[4] A. Nugraha, pratama dwi,, Said al faraby, "Klasifikasi Dokumen Menggunakan Metode Knn Dengan Information Gain," *eProceedings Eng.*, vol. 5, no. 1, pp. 1541–1550, 2018.

[5] M. Ramya and J. A. Pinakas, "Different Type of Feature Selection for Text Classification," *Int. J. Comput. Trends Technol.*, vol. 10, no. 2, pp. 102–107, 2014.

[6] S. Rahamat Basha, "Impact of feature selection techniques in Text Classification: An Experimental study," *J. Mech. Contin. Math. Sci.*, vol. 1, no. 3, pp. 39–51, 2019.

[7] A. K. Nikhath, K. Subrahmanyam, and R. Vasavi, "Building a K-Nearest Neighbor Classifier for Text Categorization," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 1, pp. 254–256, 2016.

[8] E. H. S. Han, G. Karypis, and V. Kumar, "Text categorization using weight adjusted k-nearest neighbor classification," in *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 2001, vol. 2035, pp. 53–65.

[9] M. Azam, T. Ahmed, F. Sabah, and M. I. Hussain, "Feature Extraction based Text Classification using K-Nearest Neighbor Algorithm," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 18, no. 12, pp. 95–101, 2018.

[10] T. Jo, "Improving K Nearest Neighbor into String Vector Version for Text Categorization," *Int. Conf. Adv. Commun. Technol. ICACT*, vol. 2019-February, no. 1, pp. 1091–1097, 2019.

[11] L. Greche, M. Jazouli, N. Es-Sbai, A. Majda, and A. Zarghili, "Comparison between Euclidean and Manhattan distance measure for facial expressions classification," *2017 Int. Conf. Wirel. Technol. Embed. Intell. Syst. WITS 2017*, pp. 2–5, 2017.

[12] J. Kaur and J. Saini, "A Study of Text Classification Natural Language Processing Algorithms for Indian Languages," *VNSGU J. Sci. Technol.*, vol. 4, no. 1, pp. 162–167, 2015.

[13] B. A. Eclarin, A. C. Fajardo, and R. P. Medina, "A novel feature hashing with efficient collision resolution for bag-of-words representation of text data," *ACM Int. Conf. Proceeding Ser.*, pp. 12–16, 2018.

[14] Daeli, N.O.F, Adiwijaya. Sentiment analysis on movie reviews using Information gain and K-nearest neighbor. Journal of Data Science and Its Applications, 3(1), 2020.

[15] Asriyanti Indah Pratiwi, Adiwijaya. 2018. On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis. Applied Computational Intelligence and Soft Computing, 2018.

[16] Arifin, A.H.R.Z., Mubarok, M.S. and Adiwijaya, A., Learning struktur bayesian networks menggunakan novel modified binary differential evolution pada klasifikasi data. In Indonesia Symposium on Computing (IndoSC) 2016.

[17] Naf'an, M. Z., Bimantara, A. A., Larasati, A., Risondang, E. M., & Nugraha, N. A. S. Sentiment Analysis of Cyberbullying on Instagram User Comments. Journal of Data Science and Its Applications, 2(1), 88-98, 2019

[18] Sari, P. K., & Purwadinata, A. Analysis Characteristics of Car Sales In E-Commerce Data Using Clustering Model. Journal of Data Science and Its Applications, 2(1), 68-77., 2019

Lampiran 1: Contoh *dataset* yang digunakan

Potongan Cerita	Label
Suatu siang, Devina mengajak Ira, Ovi, Tata, dan murid baru di kelasnya, Jenny, untuk mengadakan pesta di rumah pohonnya. Tentu saja mereka menerima ajakan itu dengan senang hati. Kebetulan, besok sekolah libur karena para guru akan rapat. Sore hari, teman-teman Devina datang. Mereka pun mengadakan pesta jambu sampai malam berganti pagi. Teman-teman Devina segera pulang ke rumahnya masing-masing. Sebelum pulang, mereka berbincang-bincang dulu. ”Rasanya, aku ingin memiliki rumah pohon yang banyak,” ucap Jenny. “Ya, aku juga. Bagaimana kalau kita membuat taman yang berisi banyak pohon.	0
Suatu sore Si Pitung melihat kelakuan anak buah Babah Liem yang sewenang-wenang. Babah Liem adalah tuan tanah di daerah tempat tinggal Si Pitung. Dia dan anak buahnya sering merampas harta rakyat dan menarik pajak tinggi. Sebagian hasil rampasan itu diberikan kepada pemerintah Belanda.	0
Sungguh sangat disayangkan, mereka binasa dalam keganasan banjir bandang itu. Ki Kerti Pejok tak tahu bahwa selama ini Sultan Agung memang melarang para abdinya memandikan gajah di hilir sungai. Karena ia tahu bahaya bisa datang sewaktu-waktu di sana. Ki Sapa Wira berduka. Ia sangat sedih karena kehilangan adik ipar dan gajah kesayangannya. Untuk mengenang kejadian itu, Sultan Agung menamakan sungai itu Kali Gajah Wong. Kali berarti sungai, gajah wong berarti gajah dan orang. Kali Gajah Wong ini terletak di sebelah timur Kota Yogyakarta.	0
Tanggung jawab terhadap keluarga. Contoh: seorang ibu hidup dengan tiga anak. karena suaminya meninggal dia harus bekerja untuk memenuhi kebutuhan hidup anak-anaknya. walaupun harus menjadi pelacur sekalipun. karena demi memberikan kehidupan dan bertanggung jawab atas ketiga anaknya.	1
Tidak lama berselang, terdengar pekikan tanda permusuhan. Ternyata benar, pekikan itu berasal dari suku Kuala yang mengajak berperang. Caadara memerintahkan teman-temannya pergi ke bukit yang tinggi dan membentuk benteng pertahanan. Tetapi peperangan tidak terelakkan lagi. Caadara dan teman-temannya berperang dengan suku Kuala. Pekikan mengerikan di sela suara senjata-senjata yang beradu tidak henti-hentinya terdengar. Namun, Caadara tidak gentar. Dia berhasil mengalahkan pasukan suku Kuala. Berkat petunjuk Caadara, teman-temannya pun berhasil mengalahkan musuh.	0
Beberapa saat kemudian, Ayah dan Dita sampai di ladang. Ternyata, Ayah sudah ditunggu beberapa orang yang akan membantu beliau. “Dita, itu beberapa orang yang akan membantu Ayah. Ada yang bertugas memanen sayuran, ada yang bertugas memanggul hasil panen ke aliran sungai untuk dicuci, dan ada yang membantu memindahkan sayuran ke atas mobil pengangkut. Mereka semua orang-orang yang sudah terlatih. Mereka memiliki otot kuat untuk melakukan pekerjaanpekerjaan tersebut,” terang Ayah kepada Dita. Dita mendengarkan penjelasan ayahnya. Dita mendengarkan perkataan Ayah sambil memperhatikan orang-orang yang bekerja.	1
Bergairahlah lelakiku. Aku ingin sekali menyempurnakan keinginanmu.	1
Bertahun silam, seorang mandor penebangan kayu melihatnya sedang mandi di sebuah telaga. Akhirnya terjadilah peristiwa yang merenggut kegadisannya sekaligus menimbulkan tumbuhnya janin di perutnya. Dia tadinya tidak bisa terima. Begitu lahir, bayi itu ditinggalkannya dengan kedua orangtuanya sementara dia lari ke kota. Kini dia sadar harus berbuat sesuatu untuk menghidupi anak yang pernah dikandungnya. Walau bagaimanapun dia adalah darah dagingnya. Dia ibu dari anak itu. Dari tempat paling hina di dunia ini, warung remang-remang tempat dia menjajakan badan, dia selalu diingatkan pada hal itu. Apapun. Apapun harus ia lakukan demi kehidupannya dan anak itu.	0

Dataset yang digunakan diambil dari penelitian sebelumnya oleh Mayya Tannia Wewengkan. *Dataset* dapat diakses pada *link* berikut:
https://drive.google.com/drive/folders/1MFFT52043TjZFeB9QmbMdxgt193iuzt8?usp=sharing