Syntax Literate: Jurnal Ilmiah Indonesia p–ISSN: 2541-0849 e-ISSN: 2548-1398
Vol. 7, No. 12, Desember 2022

DETEKSI PLAGIARISME BERBASIS PARAFRASE PADA TEKS BAHASA INDONESIA

Fauziyah Amini, Cahyo Crysdyan
Magister Informatika UIN Maulana Malik Ibrahim Malang, Indonesia
Email: 200605210002@student.uin-malang.ac.id, cahyo@ti.uin-malang.ac.id

Abstrak

Sistem pendeteksi plagiarisme konvensional perlu modifikasi agar mendapatkan hasil yang maksimal. Penelitian ini mengeksplorasi berbagai algoritma machine learning untuk memodelkan teks yang di parafrase dengan teks yang lain. Tujuan penelitian ini adalah untuk menganalisa kompleksitas masalah plagiarisme berdasarkan parafrase pada teks bahasa Indonesia dan membandingkan keunggulan metode-metode klasifikasi dalam mendeteksi plagiarisme berdasarkan parafrase pada teks bahasa Indonesia. Pada penelitian ini ada beberapa tahapan, yaitu : pengumpulan data, kemudian desain sistem, pada tahap desain sistem ini meliputi data extraction, text pre processing, dan term weighting. Berdasarkan hasil pengujian yang telah diuraikan maka dapat disimpulkan bahwa algoritma KNN dan SVM cukup optimal untuk mengklasifikasi dataset dalam penelitian ini karena menhasilkan akurasi yang memuaskan pada penelitian ini.

Kata Kunci: Deteksi Plagiarisme, Parafrase, Bahasa Indonesia.

Abstract

Conventional plagiarism detection systems need modifications in order to get maximum results. This research explores various machine learning algorithms to model paraphrased text with other texts. The purpose of this study is to analyze the complexity of plagiarism problems based on paraphrasing on Indonesian texts and compare the advantages of classification methods in detecting plagiarism based on paraphrasing on Indonesian texts. In this study there are several stages, namely: data collection, then system design, at the design stage this system includes data extraction, text pre-processing, and term weighting. Based on the test results that have been described, it can be concluded that the KNN and SVM algorithms are quite optimal for classifying datasets in this study because they produce satisfactory accuracy in this study.

Keywords*: Plagiarism Detection, Paraphrasing, Indonesian.*

Pendahuluan

Ketersediaan informasi digital yang semakin luas dan semakin mudah memberikan dampak peningkatan terhadap plagiarism (Iswara, 2020). Berdasarkan hasil dari beberapa

How to cite: Fauziyah Amini, Cahyo Crysdyan (2022) Deteksi Plagiarisme Berbasis Parafrase Pada Teks Bahasa Indonesia, *Syntax Literate : Jurnal Ilmiah Indonesia (7)12,* http://dx.doi.org/10.36418/syntax-literate.v7i12.10904
E-ISSN: 2548-1398
Published by: Ridwan Institute

survei penelitian menunjukkan peningkatan kasus plagiarisme baik dalam karya akademis maupun literatur ilmiah (Aziz, 2015). Plagiarisme dipandang sebagai pelanggaran ilmiah yang serius, pencurian terhadap ide-ide intelektual (Yudhana et al., 2017). Salah satu bentuk plagiarisme yang sering ditemukan saat ini adalah menjiplak karya orang lain dan menuliskannya dengan susunan kata yang berbeda atau yang biasa kita sebut dengan parafrase kalimat (Amini, 2022).

Penelitian yang membahas tentang plagiarisme sudah dilakukan sejak tahun 1990-an. Namun deteksi plagiarisme hanya didasarkan pada kata-kata yang sama kemudian dihitung presentase kemiripan antar dokumen (Yudiantoko, 2016). Metode seperti ini masih belum maksimal untuk menangani plagiarisme yang semakin marak saat ini (Damanik et al., 2021). Karena saat ini bentuk tindakan plagiarisme tidak hanya menjiplak karya orang lain, namun pelaku plagiarisme saat ini mengambil karya ilmiah atau ide orang lain kemudian dituliskan Kembali dengan makna yang sama tetapi dengan susunan kata yang berbeda (Isnaini, 2019). Tindakan plagiarisme seperti ini biasa kita sebut dengan parafrase kalimat.

Identifikasi Parafrase atau Natural Language Sentence Machine (NLSM) adalah salah satu hal yang menantang dalam pemrosesan teks. Dimana peneliti harus mengidentifikasi apakah sebuah kalimat adalah parafrase dari kalimat lain di pasangan kalimat yang diberikan (Julianto et al., 2017). Parafrase kalimat menyampaikan arti yang sama tetapi struktur dan urutan kata-katanya bervariasi. Ini adalah suatu hal menantang karena sulit untuk menyimpulkan konteks yang tepat dalam sebuah kalimat. Parafrase terjadi ketika teks dimodifikasi secara leksikal atau sintaksis (Clough & Stevenson, 2011) agar terlihat berbeda dari sumbernya, tetapi tetap memiliki makna yang sama. Parafrase itu sendiri legal bila dilakukan dengan benar seperti dalam penggunaan kembali teks Jurnalistik (Haryanto et al., 2020), tetapi ketika teks dimodifikasi dan digunakan tanpa menyebutkan sumbernya dengan benar, itu adalah termasuk tindakan plagiarisme.

Penelitian ini menghasilkan sistem pendukung keputusan yang membantu manajer dalam mempromosikan pegawai secara objektif. Aplikasi ini menghasilkan system pendukung keputusan yang menyajikan hasil nilai pegawai berupa grafik dengan menggunakan metode Multifactor Evalution Process (MFEP) (Handhika & Hendrawan, 2021).

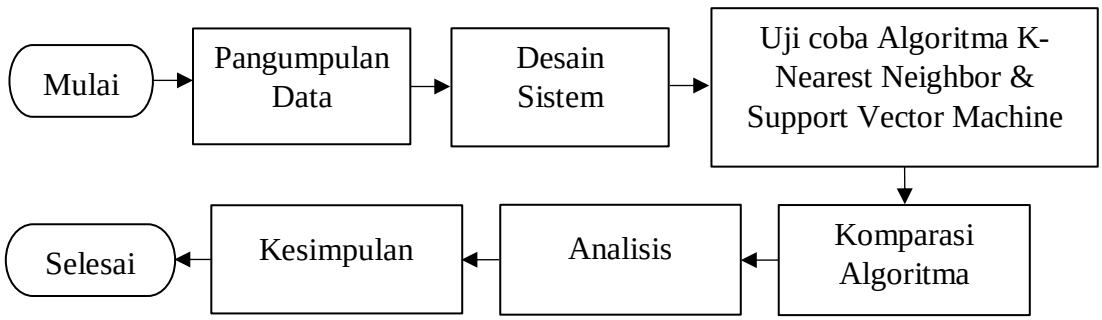
Berdasarkan permasalahan yang telah diuraikan sebelumnya, sistem pendeteksi plagiarisme konvensional perlu modifikasi agar mendapatkan hasil yang maksimal. Penelitian ini mengeksplorasi berbagai algoritma machine learning untuk memodelkan teks yang di parafrase dengan teks yang lain. Dengan adanya penelitian ini diharapkan menjadi solusi untuk mencegah tindakan plagiarisme yang semakin marak terjadi pada saat ini. Masalah dalam penelitian ini adalah plagiarisme berdasarkan parafrase pada teks Bahasa Indonesia sulit untuk diidentifikasi dan metode apa yang dapat secara optimal mendeteksi plagiarisme berdasarkan parafrase pada teks Bahasa Indonesia.

Tujuan penelitian ini adalah untuk menganalisa kompleksitas masalah plagiarisme berdasarkan parafrase pada teks bahasa Indonesia dan membandingkan keunggulan metode-metode klasifikasi dalam mendeteksi plagiarisme berdasarkan parafrase pada teks bahasa Indonesia.

Ada pun manfaat penelitian ini adalah dapat membantu instansi pendidikan dalam mencegah praktik plagiarisme dalam lingkungan instansi pendidikan, deteksi plagiarisme memberikan dukungan kepada pelaku akademik untuk membangun nilai-nilai kejujuran, keadilan, dan kepercayaan dalam mengejar integritas akademik. dapat membantu instansi pendidikan dalam mengembangkan konsep integritas pendidikan yang salah satunya mencakup deteksi plagiarisme.

Metode Penelitian

Pada penelitian ini ada beberapa tahapan, yaitu : pengumpulan data, kemudian desain sistem, pada tahap desain sistem ini meliputi data extraction, text pre processing, dan term weighting. Kemudian dilanjutkan dengan uji coba terhadap data menggunakan algoritma K-Nearest Neighbor & Support Vector Machine. Selanjutnya, kedua algoritma dikomparasi untuk melakukan analisis dan terakhir diambil kesimpulan dari penelitian ini. Alur proses desain penelitian dapat dilihat pada Gambar 1.



Gambar 1. Flowchart Prosedur Penelitian

Tahap pengumpulan data ini dilakukan untuk memperoleh informasi yang dibutuhkan dalam rangka mencapai tujuan pada penelitian ini. Dalam penelitian ini data berbentuk teks, yang didapatkan dari abstrak tugas akhir mahasiswa. Data primer ini didapat melalui *e-theses* uin malang yang diakses melalui (http://etheses.uin-malang.ac.id). Sample dataset yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. ***Sample Dataset***

Id	qid1	qid2	Dok1	Dok2	Parafrase
0	1	2	Metode fuzzy Sugeno untuk membantu Arduino Uno mengolah data menjadi suatu acuan suhu dan kelembaban yang ideal untuk perkembangan optimal jamur tiram. NodeMCU untuk membantu Arduino Uno mengirimkan data hasil pengolahan di website monitoring.	Tata cara fuzzy Sugeno buat menolong Arduino Uno mencerna informasi jadi sesuatu acuan temperatur serta kelembaban yang sempurna buat pertumbuhan maksimal jamur tiram. NodeMCU buat menolong Arduino Uno mengirimkan informasi hasil pengolahan di web monitoring.	1
1	3	4	Berkumpulnya orang di suatu tempat sehingga membentuk sebuah kerumunan merupakan hal yang lumrah saat ini. Memperkirakan jumlah orang dalam kerumunan merupakan masalah penting untuk berbagai tujuan mulai dari keselamatan umum hingga strategi industri.	Berkumpulnya orang di sesuatu tempat sehingga membentuk suatu kerumunan ialah perihal yang lumrah dikala ini. Memperkirakan jumlah orang dalam kerumunan ialah permasalahan berarti buat bermacam tujuan mulai dari keselamatan universal sampai strategi industri.	1
2	5	6	Sebagai sarana ujian saringan masuk pada perguruan tinggi islam, Pelaksanaan SSE-UMPTKIN tentu perlu mempersiapkan infrastruktur yang berhubungan dengan proses pelaksanannya. Sehingga faktor yang mempengaruhi terjadinya kendala pada pelaksanaan SSE-UMPTKIN dapat dicegah dan diatasi.	Selaku fasilitas tes saringan masuk pada akademi besar islam, Penerapan SSE-UMPTKIN pasti butuh mempersiapkan infrastruktur yang berhubungan dengan proses pelaksanannya. Sehingga aspek yang pengaruhi terbentuknya hambatan pada penerapan SSE- UMPTKIN bisa dicegah serta diatasi.	1

3	7	8	Proses penyusunan aksi rekontruksi rehabilitasi pasca terjadinya bencana merupakan hal penting, karena kegiatan ini dilakukan untuk mengetahui tingkat kerusakan dan tindakan yang perlu dilakukan setelah terjadinya bencana alam sesuai dengan data dilapangan langsung, maka perlunya dilakukan penelitian dengan metode Fuzzy-VIšekriterijumsko KOmpromisno Rangiranje (Fuzzy-VIKOR).	Proses penataan aksi rekontruksi rehabilitasi pasca terbentuknya musibah ialah perihal berarti, sebab aktivitas ini dicoba buat mengenali tingkatan kehancuran serta aksi yang butuh dicoba sehabis terbentuknya musibah alam cocok dengan informasi dilapangan langsung, hingga perlunya dicoba riset dengan tata cara Fuzzy-VIšekriterijumsko KOmpromisno Rangiranje(Fuzzy- VIKOR).	1
…	…	…	…	…	…
30	61	62	Di dalam kecerdasan buatan, agen cerdas (AI) adalah sebuah entitas otonom yang mengamati dan bertindak atas suatu lingkungan dan mengarahkan aktivitasnya tersebut untuk mencapai tujuan.	Chatbot merupakan sebuah program komputer yang dibangun untuk menampilkan percakapan atau komunikasi interaktif dengan pengguna (manusia) melalui teks, ucapan, dan atau Gambar.	0
31	63	64	Proses pemilihan aplikasi Point of Sale harus didasarkan pada kemampuan dan kebutuhan pembeli.	Ketika pembeli dihadapkan pada banyak pilihan merk POS dan berbagai spesifikasinya kebanyakan pembeli jadi kebingungan memilih aplikasi yang sesuai untuk usahanya.	0

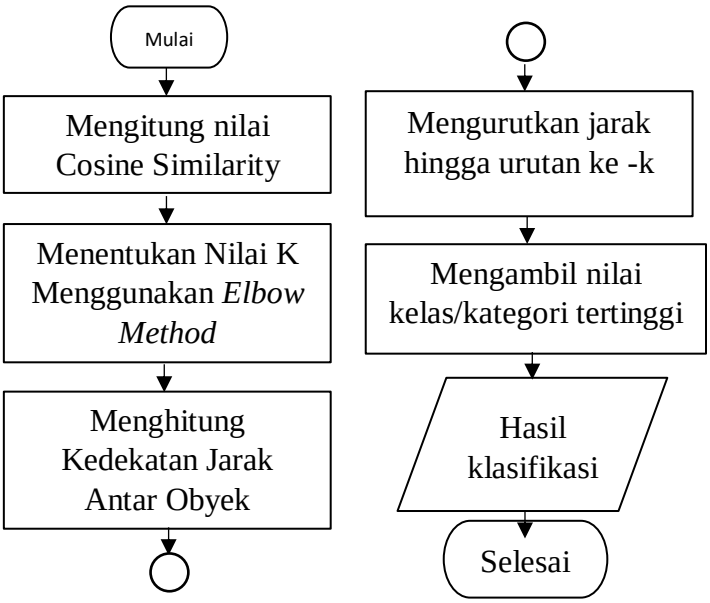
32	65	66	Toko Orisha Beauty merupakan toko yang menjual produk kecantikan yang berlokasi di Batu, Malang.	Toko ini khusus menjual produk bermerek Nu-Skin selain toko offline juga online, sehingga membutuhkan sebuah sistem yang mampu memudahkan dalam merekapitulasi pembelian dari konsumen.	0
33	67	68	Indonesia merupakan negara tropis dan rentan dengan bencana alam. Bencana alam di Indonesia sangat sering terjadi seperti gempa bumi, banjir, kebakaran, Angin Topan dan lain-lain.	Bencana alam menimbulkan dampak bagi warga terdampak, infrastruktur ataupun sektor yang terdapat di Indonesia. Pemerintah pusat, daerah, maupun kota berupaya untuk memberikan rehabilitasi pasca terjadinya bencana alam.	0

Dalam melakukan klasifikasi teks, penelitian ini menggunakan Algoritma K-Nearest Neighbor dan Support Vector Machine.

Hasil dan Pembahasan

A. Desain & Implementasi

K-Nearest Neighbor merupakan salah satu metode klasifikasi digunakan dalam *data mining* dan *machine learning* (Baharuddin et al., 2019). Kinerja klasifikasi dari KNN sangat bergantung pada metrik yang digunakan untuk menghitung jarak berpasangan antara titik data (Muflikhah & Mahmudy, 2021). Untuk menghitung k titik data tetangga terdekat yang diinginkan, pada penelitian ini menggunakan *cosine similarity* sebagai metrik kesamaan (Surenggana et al., 2022). Aturan klasifikasi KNN dibuat oleh sampel pelatihan saja, tanpa data tambahan lainnya. Klasifikasi k-NN, menemukan sekelompok k objek dalam set pelatihan yang paling dekat dengan objek uji, dan mendasarkan penetapan label pada dominasi kelas tertentu di lingkungan ini. Algoritma *K-Nearest Neighbor* (k-NN) adalah metode untuk mengklasifikasikan objek berdasarkan contoh pelatihan terdekat dalam ruang fitur (Zuhdi et al., 2019). Seperti yang telah disebutkan sebelumnya, pada penelitian ini menggunakan algoritma dalam melakukan klasifikasi teks. Berikut ini adalah *Flowchart* proses klasifikasi KNN ditunjukan pada Gambar 2.



Gambar 2. *Flowchart* Algoritma KNN

B. Uji Coba

Pada penelitian ini untuk melakukan uji coba algoritma K-Nearest Neighbor digunakan skenario eksperimen sebagai berikut :

1. Menentukan Nilai K yang paling optimal menggunakan *Elbow Method*

   Pada bagian sebelumnya telah diimplementasikan penentuan nilai K menggunakan Elbow Method. Hasil dari metode elbow method meunjukan bahwa nilai K yang paling optimal adalah K=4. Maka selanjutnya akan digunakan nilai K=4.
2. Menggunakan 4 metode pengukuran jarak, yaitu:
   a) Klasifikasi teks berdasarkan pengukuran *Manhattan Distance*
   b) Klasifikasi teks berdasarkan pengukuran *Euclidean Distance*
   c) Klasifikasi teks berdasarkan pengukuran *Minkowsky Distance*
   d) Klasifikasi teks berdasarkan pengukuran *Chebyshev Distance*

   Adapun hasil pengukuran jarak pada setiap metode pada pengujian penelitian ini di visualisasikan pada Tabel 4.12.

Tabel 2. Pengukuran *Euclidean, Manhattan, Minkowsky, Cebyshev Distance*

K	Euclidean	Manhattan	Minkowsky	Chebyshev
1	0.3400778	0.3400778	0.3400778	0.3400778
2	0.	0.	0.	0.
3	0.00461328	0.00461328	0.00461328	0.00461328
4	0.07810293	0.07810293	0.07810293	0.07810293

Berdasarkan Tabel 2 dapat dilihat bahwa pengukuran kedekatan jarak antar obyek dengan 4 metode menghasilkan jarak yang sama dalam setiap perhitungan jarak setiap pasangan kalimat. Maka dapat disimpulkan bahwa pada penelitian ini pemilihan

pengukuran jarak tidak memberikan pengaruh yang signifikan terhadap performa algoritma knn dan menghasilkan performa yang sama pada setiap metode pengukuran jarak.

3. Evaluasi performa dari masing- masing *algoritma K-Nearest Neighbor* akan dihitung Menggunakan *Confusion Matrix*. Kemudian dari Tabel *Confusion Matrix* akan dihitung nilai akurasi, *precision*, *recall* dan *f1score.* Adapun rumus Confusion Matrix dituliskan pada Tabel 3 :

Tabel 3. Tabel Confusion Matrix

		Actual Values	
		Positif	**Negatif**
Predicted Value	Positive	TP	FP
	Negative	FN	TN

Keterangan :
TP : *True Positive* merupakan data positif yang diprediksi sebagai data positif.
TN : *True Negative* merupakan data negatif yang diprediksi sebagai data negative
FP : *False Positive* merupakan data positif yang diprediksi sebagai data negatif.
FN : *False Negative* merupakan data positif yang diprediksi sebagai data negatif.

$$Rumus\ Akurasi = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (4.7)$$

$$Rumus\ Precision = \frac{TP}{TP + FP} * 100\% \quad (4.8)$$

$$Rumus\ Recall = \frac{TP}{TP + FN} * 100\% \quad (4.9)$$

$$Rumus\ F1\ Score = 2\frac{Precision\ x\ Recall}{Precision + Recall} * 100\% \quad (4.9)$$

a) Evaluasi hasil klasifikasi algoritma *K-Nearest Neighbor* berdasarkan pengukuran *Manhattan distance*

Hasil klasifikasi algoritma *K-Nearest Neighbor* dengan pengukuran jarak terdekat menggunakan metode *Manhattan distance* dalam penelitian ini dapat dilihat pada Tabel 4.14.

Dari hasil klasifikasi algoritma KNN dengan pengukuran jarak terdekat menggunakan metode *Manhattan distance* pada Tabel 4.14 dapat dilihat bahwa dari 60 pasangan data ada 22 data yang termasuk dalam *true positif*, 31 data termasuk true negative, 7 data termasuk false positif, dan tidak ada data yang termasuk *false negative*. Maka berdasarkan Tabel 4.14 dapat dihitung nilai akurasi, presisi, *recall* dan *f1 measure.*

$$Akurasi = \frac{31 + 22}{31 + 22 + 7 + 0} * 100\% = 88\%$$
$$Presisi = \frac{22}{22 + 0} * 100\% \ = 100\%$$
$$Recall = \frac{22}{7 + 22} * 100\% \ = 70\%$$
$$f - measure = 2\frac{100\ x\ 70}{100 + 70} \ = 82\%$$

b) Evaluasi hasil klasifikasi algoritma *K-Nearest Neighbor* berdasarkan pengukuran *Euclidean distance*

Hasil klasifikasi algoritma *K-Nearest Neighbor* dengan pengukuran jarak terdekat menggunakan metode *Euclidean distance* pada Tabel 4.15 dapat dilihat bahwa dari 60 pasangan data ada 22 data yang termasuk dalam *true positif*, 31 data termasuk *true negative*, 7 data termasuk *false positif*, dan tidak ada data yang termasuk *false negative*. Maka dapat dihitung nilai akurasi, presisi, *recall* dan *f1 measure*.

$$Akurasi = \frac{31 + 22}{31 + 22 + 7 + 0} * 100\% = 88\%$$
$$Presisi = \frac{22}{22 + 0} * 100\% \ = 100\%$$
$$Recall = \frac{22}{7 + 22} * 100\% \ = 70\%$$
$$f - measure = 2\frac{100\ x\ 70}{100 + 70} \ = 82\%$$

c) Evaluasi hasil klasifikasi algoritma *K-Nearest Neighbor* berdasarkan pengukuran *Minkowsky distance*

Hasil klasifikasi algoritma KNN dengan pengukuran jarak terdekat menggunakan metode *Minkowsky distance* bahwa dari 60 pasangan data ada 22 data yang termasuk dalam *true positif*, 31 data termasuk *true negative*, 7 data termasuk *false positif*, dan tidak ada data yang termasuk *false negative*. Maka dapat dihitung nilai akurasi, presisi, *recall*, dan *f1 measure*.

$$Akurasi = \frac{31 + 22}{31 + 22 + 7 + 0} * 100\% = 88\%$$
$$Presisi = \frac{22}{22 + 0} * 100\% \ = 100\%$$
$$Recall = \frac{22}{7 + 22} * 100\% \ = 70\%$$
$$f - measure = 2\frac{100\ x\ 70}{100 + 70} \ = 82\%$$

d) Evaluasi hasil klasifikasi algoritma *K-Nearest Neighbor* berdasarkan pengukuran *Cebyshev distance*

Hasil klasifikasi algoritma KNN dengan pengukuran jarak terdekat menggunakan metode *Cebyshev distance* pada Tabel 4.17 dapat dilihat bahwa dari 60 pasangan data ada 22 data yang termasuk dalam *true positif*, 31 data termasuk *true negative*, 7 data termasuk *false positif*, dan tidak ada data yang termasuk *false negative*. Maka berdasarkan Tabel 4.17 dapat dihitung nilai akurasi, presisi, *recall* dan *f1 measure*.

$$Akurasi = \frac{31 + 22}{31 + 22 + 7 + 0} * 100\% = 88\%$$

$$Presisi = \frac{22}{22 + 0} * 100\% = 100\%$$

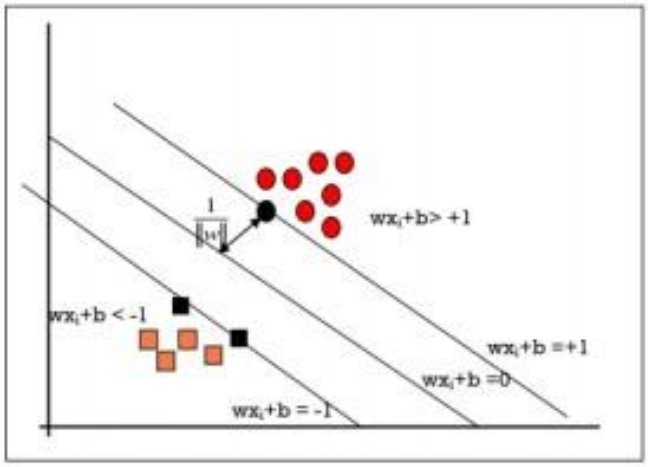
$$Recall = \frac{22}{7 + 22} * 100\% = 70\%$$

$$f - measure = 2\frac{100\ x\ 70}{100 + 70} = 82\%$$

5.1 Desain & Implementasi

Support Vector Machine (SVM) adalah metode yang mempelajari area yang memisahkan antar kategori dalam sebuah observasi. Dalam terminologi SVM, kita membahas jarak atau margin antar kategori. Setiap kategori memiliki observasi dimana nilai variabel targetnya sama (Williams, 2011). SVM juga dikenal sebagai sistem pembelajaran yang menggunakan hipotesis fungsi linear dalam ruang dimensi tinggi dan dilatih dengan algoritma berdasarkan teori optimasi dengan menerapkan learning bias yang berasal dari teori statistik. Tujuan dari metode ini adalah membangun pemisah optimum yang disebut OSH (*Optimal Separating Hyperplane*) sehingga dapat digunakan untuk klasifikasi.

Hyperplane terbaik antara kedua kelas dapat ditemukan dengan melakukan pengukuran margin *hyperplane* dan kemudian mencari titik maksimalnya. Margin adalah jarak antar *hyperplane* tersebut dengan data terdekat dari masing – masing kelas. Data yang paling dekat ini, disebut dengan *support vector* (Kasim & Sudarsono, 2019). Ilustrasi hyperplane ditunjukan pada Gambar 5.1.



Gambar 5.1 Margin Hyperplane

Seperti pada Gambar 5.1, *Support Vector Machine* bekerja menemukan *hyperplane* dengan margin yang maksimal. *Hyperplane* klasifikasi linear memisahkan kedua kelas dengan persamaan 5.1.

$$w.\ x_i + b = 0 \quad (5.1)$$

Keterangan:
w = vector bobot
x = nilai masukan atribut
b = bias

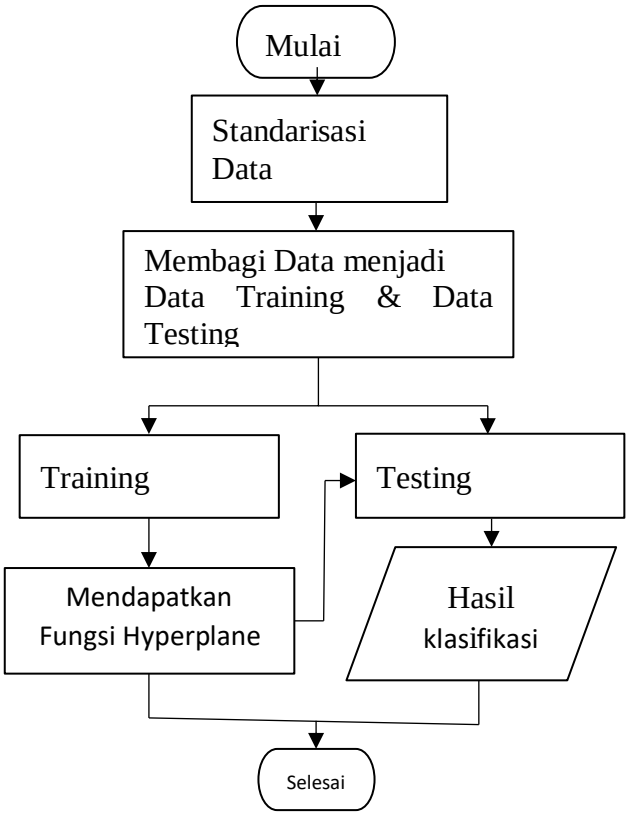
Sehingga didapatkan persamaan untuk kelas positif dan kelas negatif. Maka, suatu data x_i dapat diklasifikasikan sebagai kelas +1 jika :

$$w.\ x_i + b > 1 \quad (5.2)$$

dan dapat diklasifikasikan kedalam kelas -1 jika :

$$w.\ x_i + b \leq -1 \quad (5.3)$$

Adapun pada algoritma SVM flowchart proses klasifikasi *Support Vector Machine* pada penelitian ini ditunjukan pada Gambar 5.2.



Gambar 5.2 Flowchart *Proses Klasifikasi SVM*

5.2 Uji Coba

Tahap uji coba algoritma *support vector machine* pada penelitian ini menggunakan fungsi kernel linear untuk melakukan klasifikasi data. Kernel linear adalah fungsi kernel yang paling sederhana. Kernel linear digunakan ketika data yang dianalisis sudah terpisah secara linear. Rumus persamaan untuk fungsi kernel linear dapat dilihat pada persamaan 5.1

Berikut ini adalah cara kerja klasifikasi algoritma SVM dalam penelitian ini. Tabel 5.3 adalah sample dataset pasangan input dan output pada penelitian ini.

Table 5.3 Pasangan input *dan* output

No.	X	Y_i
1	0.97885132	1
2	-1.32669202	1
3	-0.19312049	-1
…	…	…
58	0.09695081	-1
59	1.36382549	1
60	0.22459681	1

Pada Tabel 5.3 terdapat atribut X yang akan menghasilkan nilai bobot w. Kemudian margin diminimalkan menggunakan Rumus 5.10.

$$y_i(x_i \ . \ w + b) - 1 \ \geq 0 \quad (5.10)$$

dengan syarat

$$y_i(x_i \ \cdot w + b) \geq 1, i = 1,2,3 \ldots, N \quad (5.11)$$

Sehingga dari persamaan diatas didapatkan persamaan berikut ini :

(1) 1(0.97885132 w + b) ≥1→ (0.97885132 w + b) ≥1 (5.12)
(2) 1(-1.32669202 w + b) ≥1→ (-1.32669202 w + b) ≥1 (5.13)
(3) -1(-0.19312049 w + b) ≥1→ (0.19312049 w - b) ≥1 (5.14)
(4) -1(0.09695081 w + b) ≥1→ (-0.09695081 w - b) ≥1 (5.15)
(5) 1(1.36382549 w + b) ≥1→ (1.36382549 w + b) ≥1 (5.16)
(6) 1(0.22459681w + b) ≥1→ (0.22459681 w + b) ≥1 (5.17)

Selanjutnya mencari nilai w dan b dari persamaan 5.12 dan 5.14.

(0.97885132 w + b) ≥1
(0.19312049 w - b) ≥1
——————————— +
1,17197181 w = 2
w = 2/1,17197181
w=1,72

Dari persamaan diatas dihasilkan nilai w sebesar 1,72. Kemudian Langkah selanjutnya adalah mencari nilai b dengan mensubstitusikan nilai w ke dalam persamaan 5.12.

(0.97885132 w + b) ≥1
(0.97885132 (1,72) + b) ≥1
1,66 + b = 1
b = 1-1,66
b = 0,66

Maka Persamaan hyperplane menjadi seperti berikut ini :

$$w.x + b = 0 \quad (5.18)$$
$$1{,}72\ x + 0{,}66 = 0 \quad (5.19)$$

Setelah mendapatkan garis hyperplane, maka langkah selanjutnya yaitu mengklasifikasikan data uji melalui hyperplane dengan menggunakan persamaan hyperplane pada persamaan 5.19 dengan $g(x) \coloneqq sgn(f(x))$. Hasil klasifikasi menggunakan algoritma *Support Vector Machine* menggunakan kernel linear dapat dilihat bahwa dari 60 pasangan data ada 24 data yang termasuk dalam *true positif*, 26 data termasuk *true negative*, 5 data termasuk *false positif*, dan 5 data termasuk *false negative*. Maka berdasarkan Tabel 5.4 dapat dihitung nilai akurasi, presisi, *recall* dan *f1 measure*.

$$Akurasi = \frac{26 + 24}{26 + 5 + 5 + 24} * 100\% = 83\%$$
$$Presisi = \frac{24}{24 + 5} * 100\% \ = 82\%$$

$$Recall = \frac{24}{5+24} * 100\% \ = 82\%$$

$$f - measure = 2\frac{82\,.82}{82+82} = 82\%$$

Kesimpulan

Setelah melakukan implementasi dan uji coba pada penelitian ini dapat ditemukan beberapa kesimpulan antara lain. Pemilihan nilai K paling optimal menggunakan metode elbow menunjukan hasil yang valid. Hal ini terbukti dengan hasil akurasi dari nilai K=4 menghasilkan akurasi yang paling tertinggi. Pada penelitian ini juga dilakukan pengukuran kedekatan jarak antar obyek dengan 4 metode, yaitu Manhattan Distance, Euclidean Distance, Minkowsky, Distance, dan Chebyshev Distance. Berdasarkan hasil pengujian yang telah diuraikan maka dapat disimpulkan bahwa algoritma KNN dan SVM cukup optimal untuk mengklasifikasi dataset dalam penelitian ini karena menhasilkan akurasi yang memuaskan pada penelitian ini.

BIBLIOGRAFI

Amini, F. (2022). *Deteksi Plagiarisme berbasis parafrase pada teks Bahasa Indonesia*. Universitas Islam Negeri Maulana Malik Ibrahim.

Aziz, L. A. (2015). Upaya perpustakaan dalam mengurangi plagiarisme pada karya ilmiah mahasiswa (Studi kasus di UPT Perpustakaan UNIKA Soegijapranata). *Jurnal Ilmu Perpustakaan*, *4*(3), 131–140.

Baharuddin, M. M., Azis, H., & Hasanuddin, T. (2019). Analisis Performa Metode K-Nearest Neighbor Untuk Identifikasi Jenis Kaca. *ILKOM Jurnal Ilmiah*, *11*(3), 269–274.

Clough, P., & Stevenson, M. (2011). Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, *45*(1), 5–24.

Damanik, C. M., Widjaja, F. I., Tafonao, T., Evimalinda, R., Lahagu, A., & Hartono, H. (2021). Peningkatan Kemampuan Para Dosen dalam Melakukan Tridharma sebagai Syarat Menuju Standar Pendidikan Keagamaan yang Unggul di Sekolah Tinggi Teologi Bethel Medan. *Jurnal Teologi Praksis*, *1*(2), 56–62.

Handhika, B. I., & Hendrawan, B. (2021). Implementasi Algoritma Multifactor Evaluation Process (MFEP) Untuk Penilaian Teknisi Promosi Karyawan Tetap Berbasis Web. *Syntax Idea*, *3*(1), 30–38.

Haryanto, N. C., Krisnawati, L. D., & Chrismanto, A. R. (2020). Temu kembali dokumen sumber rujukan dalam sistem daur ulang teks. *Jurnal Teknologi Dan Sistem Komputer*, *8*(2), 140–149.

Isnaini, R. L. (2019). Turn Back Plagiarism! Budaya Organisasi Anti Plagiarism. *Jurnal Akuntabilitas Manajemen Pendidikan*, *7*(2), 174–187.

Iswara, A. F. (2020). Peran Mahasiswa dalam Gerakan Open Access. *BIBLIOTIKA: Jurnal Kajian Perpustakaan Dan Informasi*, *4*(1), 64–71.

Julianto, B., Adiwijaya, A., & Mubarok, M. (2017). Identifikasi Parafrasa Bahasa Indonesia Menggunakan Naive Bayes. *EProceedings of Engineering*, *4*(3).

Kasim, A. A., & Sudarsono, M. (2019). Algoritma Support Vector Machine (SVM) untuk Klasifikasi Ekonomi Penduduk Penerima Bantuan Pemerintah di Kecamatan Simpang Raya Sulawesi Tengah. *SEMINAR NASIONAL APTIKOM (SEMNASTIK) 2019*, 568–573.

Muflikhah, L., & Mahmudy, W. F. (2021). *Machine Learning dalam Bioinformatika*. Universitas Brawijaya Press.

Surenggana, F. F., Aranta, A., & Bimantoro, F. (2022). Klasifikasi Mood Musik Menggunakan K-Nearest Neighbor dan Mel Frequency Cepstral Coefficients. *Jurnal Teknologi Informasi, Komputer, Dan Aplikasinya (JTIKA)*, *4*(2), 263–276.

Yudhana, A., Djayali, A. D., & Sunardi, S. (2017). Sistem Deteksi Plagiarisme Dokumen Karya Ilmiah dengan Algoritma Pencocokan Pola. *Jurnal Rekayasa Teknologi Informasi (JURTI)*, *1*(2), 178–187.

Yudiantoko, A. (2016). *Analisa Dampak Supply Chain Management Pada Kinerja Operasional Industri Kreatif (Studi Kasus Industri Kerajinan Batik di DI Yogyakarta)*.

Zuhdi, A. M., Utami, E., & Raharjo, S. (2019). Analisis sentiment twitter terhadap capres Indonesia 2019 dengan metode K-NN. *Jurnal Informa: Jurnal Penelitian Dan Pengabdian Masyarakat*, *5*(2), 1–7.