PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* (KNN) DALAM KLASIFIKASI PENILAIAN JAWABAN UJISAN ESAI

Muhammad Iqbal Mubarok, Purwantoro, Carudin
Informatika, Universitas Singaperbangsa Karawang
Jl. HS. Ronggo Waluyo, Telukjambe Timur, Karawang, Indonesia
Email: 1910631170214@student.unsika.ac.id

ABSTRAK

Pendidikan menjadi faktor yang mempengaruhi manusia untuk dapat berproses dan berinteraksi dengan lingkungan sekitar dan merupakan hal pokok yang perlu diperhatikan karena pendidikan dapat membentuk karakter pribadi seseorang. Evaluasi pembelajaran dilakukan sebagai salah satu tolak ukur pada siswa terhadap pemahaman dalam materi yang diberikan oleh guru dan guru akan mengetahui kemampuan dari setiap masing-masing siswanya. Ujian itu sendiri dibagi menjadi dua sifat yang berbeda, yakni ujian objektif yang jawabannya bersifat pilihan ganda dan ujian subjektif yang jawabannya bersifat jawaban uraian atau esai. Ujian esai seringkali menjadi tantangan bagi pengajar dan penilai dalam mengidentifikasi serta memberikan skor yang objektif kepada setiap jawaban mahasiswa. Dalam Upaya meningkatkan efisiensi dan keakuratan dalam penilaian, penelitian ini menggunakan metode KNN untuk mengklasifikasikan jawaban ujian esai berdasarkan kesamaan dengan jawaban-jawaban referensi yang suda ada sebelumnya. Tujuan dari penelitian ini yaitu untuk menerapkan algoritma *K-Nearest Neighbor* (KNN) dalam proses klasifikasi penilaian jawaban ujian esai dan mengetahui performa dari algortima *K-Nearest Neighbor* (KNN) yang diterapkan dalam klasifikasi penilaian jawaban ujian esai. Metodologi penelitian yang diterakan dalam penerapan algoritma KNN dalam klasifikasi penilaian ujian esai menggunakan pendekatan *Knowledge Discovert in Database* (KDD). Klasifikasi menggunakan KNN dan pembobotan TF-IDF ini menghasilkan akurasi sebesar 14,01%. Sedangkan performa lainnya yaitu nilai *precision* diperoleh rata-rata 10,4% dan *recall* dengan rata-rata 14,1%, serta *f-measure* sebesar 5,7%. Data latih yang terdiri dari berbagai jawaban referensi digunakan untuk melatih model KNN dan kemudian model ini diuji pada jawaban ujian esai yang belum pernah dilihat sebelumnya. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam mempermudah proses penilaian jawaban ujian esai secara efisien dan akurat, serta meningkatkan kualitas evaluasi akademik.

Kata kunci : *K-Nearest Neighbor, Knowledge Discovery in Database,* Ujian Esai

1. PENDAHULUAN

Pendidikan menjadi faktor yang mempengaruhi manusia untuk dapat berproses dan berinteraksi dengan lingkungan sekitar dan merupakan hal pokok yang perlu diperhatikan karena pendidikan dapat membentuk karakter pribadi seseorang. Pendidikan menjadi salah satu bekal yang penting di masa depan. Pendidikan dikenal sejak zaman sebelum negara Indonesia merdeka hingga saat ini [1]

Ki Hajar Dewantara selaku Bapak Pendidikan Nasional Indonesia, berpendapat bahwa "Pendidikan sangat penting selama pertumbuhan anak-anak, karena pendidikan menuntun segala kekuatan alam pada anak-anak untuk membantu mereka menjadi manusia dan anggota masyarakat yang baik dan bahagia." [1] Murid atau siswa bukan mesin berbentuk manusia yang dapat diatur sekehendaknya, tetapi mereka adalah generasi yang harus kita bantu dengan memberikan kepedulian terhadap setiap reaksi perubahan menuju dewasa agar dapat membentuk insan yang berpikir kritis serta memiliki sikap akhlak yang baik [1]

Provinsi Jawa Barat merupakan Provinsi dengan jumlah peserta didik terbanyak se-Indonesia pada tahun ajaran 2022/2023. Tercatat sebanyak 9.504.508 peserta didik di Jawa Barat pada tahun ajaran tersebut [2]. Dibawah ini gambar 1 merupakan data dari jumlah peserta didik.

Gambar 1. Data peserta didik nasional 2022/2023

Sekolah adalah suatu lembaga yang menyediakan kegiatan belajar mengajar melibatkan guru yang bertugas sebagai tenaga mengajar, memberikan materi pelajaran kepada peserta didik atau siswa. Tugas guru tidak hanya memberikan materi pelajaran, guru harus melakukan berupa ujian agar dapat memberikan evaluasi pembelajaran terhadap siswanya.

Evaluasi pembelajaran dilakukan sebagai salah satu tolak ukur pada siswa terhadap pemahaman dalam materi yang diberikan oleh guru dan guru akan mengetahui kemampuan dari setiap masing-masing siswanya. Ujian itu sendiri dibagi menjadi dua sifat yang berbeda, yakni ujian objektif yang jawabannya

bersifat pilihan ganda dan ujian subjektif yang jawabannya berupa jawaban uraian atau esai.

Penelitian yang terkait dengan data mining ataupun text mining khususnya pada perbandingan algoritma K-Nearest Neighbor (KNN) dengan algoritma yang lainnya sudah banyak dilakukan. Pada penelitian [3] mengenai implementasi algoritma KNN untuk melakukan klasifikasi produk dari *e-marketplace*. Sebanyak 450 data produk yang berhasil dikumpulkan dari masing-masing *marketplace* untuk diklasifikasikan, 450 data tersebut dibagi menjadi 3 pengujian yang berbeda. Klasifikasi ini menggunakan algoritma KNN. Hasil yang didapatkan dari masing-masing pengujian pada klasifikasi produk dari beberapa *e-marketplace* dengan algoritma KNN menghasilkan hasil untuk pengujian 1, dengan nilai k =1, 5, 10 adalah 78%, 97,33%, dan 92%. Pada pengujian ini nilai k yang ideal adalah 5. Pada pengujian 2 akurasinya adalah 96,67%, merupakan nilai yang sangat akurat karena melebihi 90%. Data latih sangat memengaruhi akurasi algoritma KNN. Data latih yang lebih lengkap akan meningkatkan akurasi.

Berdasarkan beberapa hasil dari penelitian terdahulu, algortima KNN memiliki kinerja klasifikasi data yang baik. Hal ini didasari dengan tingginya nilai akurasi yang didapatkan dari algoritma KNN. Dengan demikian, pada penelitian saat ini akan menggunakan algoritma KNN untuk klasifikasi ujian esai yang tentunya akan memudahkan para guru dalam mengoreksi jawaban ujian esai. Pada penelitian ini akan dilakukan implementasi dari algortima *K-Nearest Neighbor* dalam pendekatan *text mining* dengan tujuan utama untuk mengetahui performa dari algoritma tersebut.

2. TINJAUAN PUSTAKA

2.1. Data

Data adalah hasil observasi secara langsung dari representasi visual sebuah objek atau ide di dunia nyata [4]. Oleh sebab itu, diperlukannya pengolahan data dan diproses lebih lanjut dengan model untuk mendapatkan suatu informasi. Data dapat berupa teks atau *image* yang dilengkapi dengan nilai-nilai tertentu setelah diamati secara langsung peristiwa atau fenomena di dunia nyata [4].

Data dibagi menjadi data primer serta data sekunder. Data primer didapat melalui metode langsung, seperti interaksi wawancara, opini, serta hasil pengujian atau peristiwa aktual. Di sisi lain, data sekunder merujuk pada sumber informasi yang didapatkan secara tidak langsung, seperti buku, catatan, dokumen yang ada, serta arsip yang dapat ditemukan baik dalam bentuk publikasi maupun tidak dipublikasikan. Keduanya memiliki kelebihan dan kekurangan, baik data premier yang lebih valid karena langsung dari sumbernya, akan tetapi membutuhkan waktu dan biaya untuk mencari sumbernya, sedangkan data sekunder yang tidak membutuhkan waktu dan biaya dibutuhkan untuk penelitian, tetapi resiko yang akan didapat adalah data yang diambil bisa terlalu lama atau tidak sesuai yang dapat mempengaruhi penelitian [5].

2.2. *Data Mining*

Data mining menurut [6] merupakan sebuah proses yang dalam pengerjaannya menggunakan statistik, *artificial intelligene*, dan pembelajaran mesin untuk mengekstrak. *Data mining* merupakan istilah dari penemuan pengetahuan atau pemahaman pola untuk mengetahui pengetahuan yang terdapat dari kumpulan data yang sangat besar [7].

2.3. Ujian

Ujian atau tes adalah metode yang menyeluruh, sistematis, dan objektif untuk mengevaluasi pembelajaran. Hasilnya digunakan sebagai dasar untuk membuat keputusan tentang bagaimana guru memberikan instruksi dalam pengajaran. [8]. Ujian adalah suatu hal yang perlu dilakukan bagi seseorang yang sedang belajar agar mengetahui tingkat pemahaman terhadap materi yang sedang dipelajari. Ujian saat ini banyak yang menggunakan dengan cara melalui internet atau komputer yang dapat diakses dimanapun dan mempermudah para siswa [8].

2.4. *Text Mining*

Menurut [9], text mining adalah sebuah variasi dari data mining, mengekstraksi pengetahuan implisit dari data tekstual atau data yang bersifat teks yang tidak terstruktur dan dalam jumlah besar. Tujuannya untuk mendapatkan informasi dari data tekstual. Data tekstual yang dapat digunakan adalah dokumen yang berupa berita, artikel, laporan penelitian, ujian dan lain-lain. Text mining dibagi menjadi empat tipe, yaitu klasifikasi (classification), regresi (regression), pengelompokkan (clustering) dan asosiasi (association).

2.5. Klasifikasi Teks

Menurut [10] Klasifikasi teks merupakan sebuah proses menemukan model yang dapat membedakan *class* data atau konsep yang tujuannya untuk memprediksi *class* dari objek yang belum teridentifikasi. Ada dua golongan teks, yaitu *clustering text* dan klasifikasi teks. *Clustering text* berhubungan suatu struktur *unsupervised* dari sekumpulan data. Sedangkan klasifikasi teks adalah proses untuk membentuk golongan kelas dari dokumen secara *supervised*

2.6. *Knowladge Discovery in Database* (KDD)

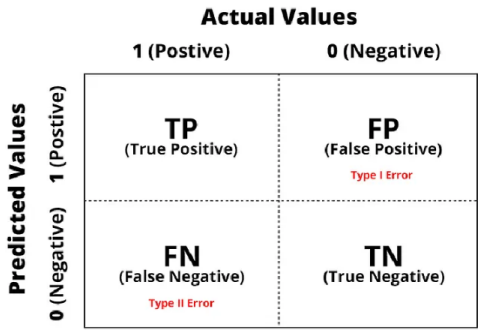
Algoritma *K-Nearest Neighbor* (KNN) adalah algoritma yang memiliki kegunaan dalam mengklasifikasikan objek berdasarkan sejumlah k data latih yang paling dekat dengan objek tersebut. Nilai k harus memenuhi syarat tidak boleh melebihi jumlah data latih, dan k harus bernilai ganjil dan lebih dari satu [11]. Untuk menghitung jarak antara dua objek *x* dan *y*

digunakan rumus jarak *Euclidean* sebagaimana tercantum pada persamaan 1.

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

2.7. *Confusion Matrix*

Confusion matrix diperlukan untuk mengevaluasi kinerja model klasifikasi karena terdiri dari banyak baris data uji *true* atau *false*. Kasus *multiple classifier* atau kelas yang jumlahnya lebih dari dua memerlukan penggunaan teknik ini [11]. Berikut adalah tabel *confusion matrix* yang dapat dilihat pada gambar 2.



Gambar 2. Tabel *confusion matrix*
(sumber: Nugroho, 2019)

Gambar 2.1 di atas adalah tabel *confusion matrix* yang dimana terdapat empat kombinasi nilai aktual dan nilai prediksi, yaitu:

1. *True Positive* (TP) adalah data positif yang diprediksi benar.
2. *True Negative* (TN) adalah data negative yang diprediksi benar.
3. *False Positive* (FP) merupakan data negatif namun diprediksi sebagai data positif. *False Negativve* (FN) merupakan data yang diprediksi negatif tetapi sebenarnya sebagai positif.

2.8. *Term Frequency Inverse Document Frequency*

Dalam penelusuran informasi dan *text mining*, pembobotan *Frequency Inverse Document Frequency* sering digunakan [13]. Dalam proses pembobotan ini terdapat dua konsep yang relevan, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF) [14]. TF mengindikasikan jumlah kemunculan kata tertentu dalam suatu dokumen, dan semakin sering kata tersebut muncul, semakin tinggi nilai TF-nya. Sementara itu, IDF menggambarkan nilai dokumentasi dari kata yang jarang muncul dalam suatu dokumen, sehingga nilai IDF akan lebih tinggi dibandingkan dengan kata yang sering muncul. Persamaan untuk menghitung TF [13] dapat dilihat pada persamaan 1.

$$TF(d,t) = f(d,t) \quad (2)$$

Keterangan:
F(d,t) kemunculan kata t dalam dokumen d
Berikut adalah persamaan untuk menghitung nilai IDF yang dapat dilihat pada perssamaan 3

$$IDF(t) = log\frac{N}{DF_t} \quad (3)$$

Keterangan:
N: banyaknya dokumen
DF_t: banyaknya dokumen DF_t yang mengandung fitur t
Nilai TF-IDF didapat dengan cara mengalihkan nilai TF dengan nilai IDF, dapat dilihat pada persamaan 4

$$TFIDF = TF(d,t).IDF(t) \quad (4)$$

Hasil dari pembobotan TF-IDF yang didapatkan selanjutnya dapat berguna untuk klasifikasi dokumen

2.9. *Supervised Learning*

Supervised Learning merujuk pada kumpulan contoh yang dilengkapi dengan label yang digunakan untuk mengidentifikasi pola distribusi karakteristik perilaku dalam berbagai aplikasi, sehingga menghasilkan model perilaku data yang dapat digunakan. [15]. *Supervised learning* terbagi menjadi dua jenis masalah, yaitu klasifikasi dan regresi. Ketika variabel *output* berupa kategori seperti merah, biru, atau penyakit dan tidak memiliki nilai berkelanjutan, disebut sebagai masalah klasifikasi. Di sisi lain, masalah regresi terjadi ketika variabel *output* memiliki nilai riil misalnya, dolar atau berat [15]

2.10. Bahasa Pemrograman Phyton

Bahasa pemrograman Phyton menurut [16] adalah bahasa pemrogram yang interpretative multiguna dengan filosofi perancangan yang berfokus pada tingkat keterbacaan kode. Phyton juga diklaim sebagai bahasa yang menggabungkan kapabilitas, kemampuan dengan sintaks kode yang sangat jelas dan juga dilengkapi dengan fungsionalitas Pustaka standar yang besar serta komprehensif.

Gambar 3. Logo phyton
(Sumber: python.org)

Phyton sendiri merupakan salah satu bahasa pemrograman dengan tingkat tinggi. Phyton ini dirancang agar memberikan kemudahan bagi para programmer melalui segi efisiensi waktu, kemudahan dalam pengembangan dan kompatibilitas dengan sistem. Phyton juga mampu digunakan untuk membuat aplikasi *standalone* (berdiri sendiri) dan pemrograman *script* (*scripting programming*).

3. METODE PENELITIAN

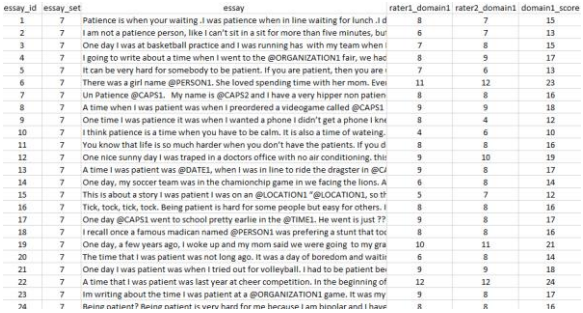
Metodologi penelitian yang di terapkan dalam penerapan algoritma KNN dalam klasifikasi penilaian ujian esai menggunakan pendekatan Knowledge Discovert in Database (KDD). Adapun tahap-tahap dari KKD yaitu sebagai berikut:

1. Data Selection
   Pada tahap seleksi data, akan dilakukan pengambilan data yang terkait dengan penelitian. Data yang akan digunakan diambil dari situs Kaggle.com yaitu Automated Essay Scoring. Selanjutnya data yang telah didapatkan kemudian dipilih essay set berapa yang akan digunakan. Pada penelitian ini yang akan digunakan yaitu essay set nomor empat.
2. Pre-Processing
   Tahap ini merupakan tahap pengolahan data. Data yang telah diambil melalui tahap seleksi tetapi data tersebut masih dalam kondisi tidak terstruktur, yang kemudian akan dilakukan pembersihan data agar dapat diolah dan mendapatkan hasil yang baik. Pada tahap ini merupakan tahap yang sangat berperan penting dalam hasil akhir pengolahan data. *Pre-processing* memiliki beberapa tahapan di dalamnya, yaitu *cleaning*, *case folding*, *tokenizing stopword removal* dan *stemming*
3. Transformation
   Tahap transformation ini akan dilakukan proses pengubahan nilai dari ketegorikal menjadi numerik atau sebagai teknik seleksi feature. Teknik yang akan digunakan pada tahap ini menggunakan Term Frequency Inverse Document Frequency (TF-IDF) yang merupakan proses pembobotan kata yang dimana akan dilakukan ekstraksi kata menjadi suatu nilai.
4. Data Mining
   Pada tahap *data mining* akan dilakukan proses pengolahan data dari proses yang sebelumnya dengan menggunakan algoritma yang telah dipilih. Algoritma yang akan digunakan pada proses ini menggunakan algoritma *K-Nearest Neighbor* (KNN). Kemudian, teknik yang digunakan dalam menghitung jarak tetangga terdekat dengan menggunakna teknik yang umum digunakan, yaitu *Euclidean distance*
5. Interpretasi/Evaluasi.
   Pada tahap akhir dari proses KDD ini terdapat dua bagian yaitu interpretasi dan evaluasi. Interpretasi merupakan pola informasi yang telah dihasilkan dari proses *data mining* yang kemudian ditampilkan dalam bentuk tabel pengetahuan (*knowledge*). Sedangkan evaluasi akan menganalisa apakah hasil dari pengolahan *data mining* memiliki pola yang menarik atau bertentangan dengan fakta yang ada pada sebelumnya. Selain itu, akan dilakukan juga Analisa mengenai prediksi yang dilakukan.

4. HASIL DAN PEMBAHASAN

4.1. Hasil

Hasil penelitian yang telah dilakukan adalah bagaimana mengklasifikasikan penilaian jawaban ujian esai menggunakan algoritma K-Nearest Neighbor (KNN). Automated Essay Scoring berisikan suatu data penilaian ujian esai siswa kelas 7-10 yang didapat dari situs Kaggle.com. Data penilaian ujian esai siswa dikumpulkan dalam kurun waktu dua bulan, yaitu sejak tanggal 10 Februari hingga 1 Mei 2012. Berikut gambar 4 data yang digunakan untuk proses klasifikasi:

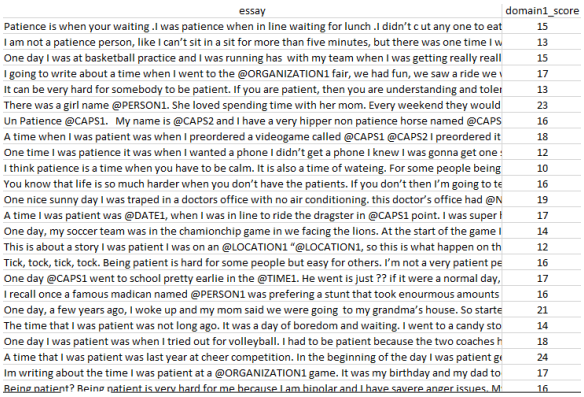


essay_id	essay_set	essay	rater1_domain1	rater2_domain1	domain1_score
1	7	Patience is when your waiting .I was patience when in line waiting for lunch .I d	8	7	15
2	7	I am not a patience person, like I can't sit in a sit for more than five minutes, bu	6	7	13
3	7	One day I was at basketball practice and I was running has with my team when I	7	8	15
4	7	I going to write about a time when I went to the @ORGANIZATION1 fair, we had	8	9	17
5	7	It can be very hard for somebody to be patient. If you are patient, then you are	7	6	13
6	7	There was a girl name @PERSON1. She loved spending time with her mom. Eve	11	12	23
7	7	Un Patience @CAPS1. My name is @CAPS2 and I have a very hipper non patien	8	8	16
8	7	A time when I was patient was when I preordered a videogame called @CAPS1	9	9	18
9	7	One time I was patience it was when I wanted a phone I didn't get a phone I kn	8	4	12
10	7	I think patience is a time when you have to be calm. It is also a time of wateing.	4	6	10
11	7	You know that life is so much harder when you don't have the patients. If you d	8	8	16
12	7	One nice sunny day I was traped in a doctors office with no air conditioning. this	9	10	19
13	7	A time I was patient was @DATE1, when I was in line to ride the dragster in @C	9	8	17
14	7	One day, my soccer team was in the chamionchip game in we facing the lions. A	6	8	14
15	7	This is about a story I was patient I was on an @LOCATION1 "@LOCATION1, so th	5	7	12
16	7	Tick, tock, tick, tock. Being patient is hard for some people but easy for others. I	8	8	16
17	7	One day @CAPS1 went to school pretty earlie in the @TIME1. He went is just ??	9	8	17
18	7	I recall once a famous madican named @PERSON1 was prefering a stunt that toc	8	8	16
19	7	One day, a few years ago, I woke up and my mom said we were going to my gra	10	11	21
20	7	The time that I was patient was not long ago. It was a day of boredom and waitin	6	8	14
21	7	One day I was patient was when I tried out for volleyball. I had to be patient bec	9	9	18
22	7	A time that I was patient was last year at cheer competition. In the beginning of	12	12	24
23	7	Im writing about the time I was patient at a @ORGANIZATION1 game. It was my	9	8	17
24	7	Being patient? Being patient is very hard for me because I am bipolar and I have	8	8	16

Gambar 4. Data *Automated Essay Scoring*

Selama periode pengumpulan data yang dilakukan, didapatkan sebanyak 12.978 data dari total 8 *essay-set.* Pada penelitian ini hanya akan menggunakan *essay-set* nomor 7 yang merupakan penilaian ujian esai siswa kelas 9. Adapun jumlah data pada *essay-set* 7 yaitu sebanyak 1569 data.

1. *Selection Data*
   Pada tahap ini akan berfokus pada seleksi data *essay set*, yang dimana *essay set* yang dipilih untuk diolah yaitu *essay set* nomor 7. Sebelum pada tahap seleksi data akan dilakukan seleksi atribut terlebih dahulu untuk pengolahan *text mining*. Secara keseluruhan, terdapat enam atribut yang ada pada data *Automated Essay Scoring.* Pada gambar 5 dibawah adalah data yang dipilih dan akan diolah setelah melalui tahap *data selection*.



essay	domain1_score
Patience is when your waiting .I was patience when in line waiting for lunch .I didn't c ut any one to eat	15
I am not a patience person, like I can't sit in a sit for more than five minutes, but there was one time I w	13
One day I was at basketball practice and I was running has with my team when I was getting really reall	15
I going to write about a time when I went to the @ORGANIZATION1 fair, we had fun, we saw a ride we	17
It can be very hard for somebody to be patient. If you are patient, then you are understanding and tole	13
There was a girl name @PERSON1. She loved spending time with her mom. Every weekend they would	23
Un Patience @CAPS1. My name is @CAPS2 and I have a very hipper non patience horse named @CAPS	16
A time when I was patient was when I preordered a videogame called @CAPS1 @CAPS2 I preordered it	18
One time I was patience it was when I wanted a phone I didn't get a phone I knew I was gonna get one	12
I think patience is a time when you have to be calm. It is also a time of wateing. For some people being	10
You know that life is so much harder when you don't have the patients. If you don't then I'm going to te	16
One nice sunny day I was traped in a doctors office with no air conditioning. this doctor's office had @N	19
A time I was patient was @DATE1, when I was in line to ride the dragster in @CAPS1 point. I was super	17
One day, my soccer team was in the chamionchip game in we facing the lions. At the start of the game	14
This is about a story I was patient I was on an @LOCATION1 "@LOCATION1, so this is what happen on th	12
Tick, tock, tick, tock. Being patient is hard for some people but easy for others. I'm not a very patient pe	16
One day @CAPS1 went to school pretty earlie in the @TIME1. He went is just ?? if it were a normal day,	17
I recall once a famous madican named @PERSON1 was prefering a stunt that took enourmous amounts	16
One day, a few years ago, I woke up and my mom said we were going to my grandma's house. So starte	21
The time that I was patient was not long ago. It was a day of boredom and waiting. I went to a candy sto	14
One day I was patient was when I tried out for volleyball. I had to be patient because the two coaches h	18
A time that I was patient was last year at cheer competition. In the beginning of the day I was patient g	24
Im writing about the time I was patient at a @ORGANIZATION1 game. It was my birthday and my dad to	17
Being patient? Being patient is very hard for me because I am bipolar and I have savere anger issues. M	16

Gambar 5. Hasil *Selection Data*

Pada gambar 5 diatas adalah hasil dari proses seleksi data yang selanjutnya akan diolah pada tahap pre-processing yang dimana terdapat atribut essay yang berisikan jawaban ujian esasi siswa dan class domain1_score yang berisi nilai dari jawaban esai tersebut.

2. *Pre-Processing*
   Data yang telah melalui tahap seleksi, selanjutnya akan melalui tahap pre-processin. Pada tahap ini terbagi menjadi beberapa proses, yaitu *data cleaning, case folding, tokenizing, stemming* dan *stopword removal.*

a. *Data Cleaning*
   Pada bagian proses *data cleaning*, pembersihan karakter yang berupa simbol atau tanda baca (*punctuation*) dan angka. Karakter tersebut akan dihapus atau digantikan oleh spasi. Pada proses penghilangan tanda baca, data sebelum dan sesuah melalui tahap ini.

b. *Case Folding*
   Pada proses *case folding*, dilakukan penyeragaman huruf dengan merubah semua bentuk huruf kapital menjadi huruf kecil. Hal ini dilakukan agar seluruh data yang akan diproses memiliki penyeragaman karakter, sehingga ketika proses selanjutnya akan lebih mudah dilakukan.

c. *Tokenizing*
   Pada proses *tokenizing* akan dilakukan proses pemotongan *string* berdasarkan kata yang tersusun sehingga memudahkan dalam proses analisis selanjutnya.

d. *Stemming*
   Pada proses *stemming* merupakan tahap akhir dari tahap *pre-processing.* Pada proses ini akan dilakukan penghilangan imbuhan (kata awal atau akhir) pada setiap kata untuk mencari kata dasar.

e. *Stopword removal*
   Pada proses *stopword removal* akan dilakukan penyaringan kata yang tidak relevan dalam klasifikasi. Proses ini dilakukan untuk meningkatkan tingkat efisiensi analisis teks dengan mengurangi jumlah kata yang diproses dan meningkatkan kualitas analisis dengan menghilangkan kata yang tidak relevan.

3. *Transformation*
   Pada tahap transformation akan dilakukan pembobotan kata dengan menggunakan *Term Frequency Inverse Document Frequency* (TF-IDF). Langkah awal yang dilakukan pada proses ini adalah mencari *Term Frequency* (TF) atau kata dengan frekuensi terbanyak. Berikut pada gambar 6, merupakan hasil dari *Term Frequency* (TF).

Kata	Frekuensi		
'was',	7219	'were',	1390
'patient',	2660	'then',	1307
'had',	2199	'there',	1224
'be',	1823	'go',	1208
'got',	1688	'patience',	1204
'get',	1574	'up',	1197
'time',	1553	'mom',	1176
'have',	1475	'wait',	1098
'said',	1423	'about',	1084
		'went',	1060
		'out',	1027

Gambar 6. Hasil *Term Frequency* (TF)

Pada gambar 6 merupakan hasil dalam pencarian 20 besar kata yang memiliki frekuensi terbanyak. Kata dengan perolehan frekuensi terbesar dimiliki oleh kata “was” dengan frekuensi 7219, kemudian disusul oleh ”patient” dan “had” dengan perolehan frekuensi 2660 dan 2199.

Proses selanjutnya setelah melalui TF yaitu proses pembobotan kata dengan Term-Frequency-Inverse Document Frequency (TF-IDF). Berikut pada gambar 7 merupakan hasil pembobotan kata TD-IDF

```
# Menampilkan tabel hasil TF-IDF
print(tfidf_df)
```

```
      abandoned  abat  abc  abel  ability  abl  able  ablle  abnacious  \
0           0.0   0.0  0.0   0.0      0.0  0.0   0.0    0.0        0.0
1           0.0   0.0  0.0   0.0      0.0  0.0   0.0    0.0        0.0
2           0.0   0.0  0.0   0.0      0.0  0.0   0.0    0.0        0.0
3           0.0   0.0  0.0   0.0      0.0  0.0   0.0    0.0        0.0
4           0.0   0.0  0.0   0.0      0.0  0.0   0.0    0.0        0.0
...         ...   ...  ...   ...      ...  ...   ...    ...        ...
1564        0.0   0.0  0.0   0.0      0.0  0.0   0.0    0.0        0.0
1565        0.0   0.0  0.0   0.0      0.0  0.0   0.0    0.0        0.0
1566        0.0   0.0  0.0   0.0      0.0  0.0   0.0    0.0        0.0
1567        0.0   0.0  0.0   0.0      0.0  0.0   0.0    0.0        0.0
1568        0.0   0.0  0.0   0.0      0.0  0.0   0.0    0.0        0.0
```

Gambar 7. Hasil *Term-Frequency-Inverse Document Frequency* (TF-IDF)

4. *Data Mining*
   Setelah melalui dua tahap yaitu tahap pre-processing dan tahap transformation, tahap selanjutnya adalah data mining. Pada tahap data mining merupakan tahap klasifikasi data penilaian ujian esai menggunakan algoritma *K-Nearest Neighbor* (KNN) untuk masuk dalam proses pengklasiikasian, yaitu data latih (*training*) dan data uji (*testing*). Dalam pembagian data tersebut diterapkan teknik *percentage split* yang dimana terdapat empat skenario pembagian data yang berbeda, yaitu 10:90, 80:20, 70:30 dan 60:40. Berikut adalah pembagian data pada tabel 1.

Tabel 1. Pembagian Data

Split	**Data Latih**	**Data Uji**
60:40	941	628
70:30	1098	471
80:20	1255	314
90:10	1412	157
Total Data: 1569		

Dalam melakukan implementasi skenario pembagian data yang telah ditetapkan, akan dilakukan proses klasifikasi menggunakan algoritma *K-nearest Neighbor.* Berikut hasil dari klasifikasi dapat dilihat pada tabel 2.

Tabel 2. Hasil Klasifikasi

Split	***K***	***Accuracy***	***Precision***	***Recall***	***F-Measure***
60:40	29	12,261%	3,243%	12,261%	4,239%
70:39	29	12,314%	2,979%	12,314%	4,449%
80:20	29	12,101%	8,159%	12,101%	12,101%
90:10	**29**	**14,012%**	10,434%	14,012%	5,702

Berdasarkan tabel 2, terdapat empat skenario pembagian data training dan data testing yang

dimana untuk detail pembagiannya telah dijelaskan pada tabel 1. Dari hasil yang ditunjukkan pada tabel 2, bahwa rasio 90:10 merupakan hasil yang cukup baik dibandingkan dengan perobaan yang lainnya. Maka dari itu dalam klasifikasi penilaian jawaban ujian esai dengan menggunakan algoritma *K-Nearest Neighbor* (KNN) dilakukan dengan pembagian data training yakni 1.412 dan data testing 157.

5. Evaluasi
   Evaluasi untuk mengukur kinerja dari algoritma klasifikasi K-Nearest Neighbor (KNN) dilakukan dengan menggunakan confusion matrix. Berdasarkan dengan perhitungan kinerja klasifikasi sebelumnya, dapat dilihat seperti pada gambar 8.

confusion matrix:

	precision	recall	f1-score	support
4	0.00	0.00	0.00	1
5	0.00	0.00	0.00	1
6	0.25	0.50	0.33	2
7	0.00	0.00	0.00	4
8	0.33	0.07	0.12	14
9	0.00	0.00	0.00	4
10	0.00	0.00	0.00	13
11	0.07	0.07	0.07	15
12	0.10	0.08	0.09	24
13	0.17	0.06	0.08	18
14	0.00	0.00	0.00	16
15	0.11	0.08	0.10	12
16	0.18	0.55	0.27	40
17	0.13	0.16	0.14	38
18	0.23	0.13	0.17	23
19	0.00	0.00	0.00	17
20	0.14	0.06	0.08	18
21	0.00	0.00	0.00	10
22	0.00	0.00	0.00	11
23	1.00	0.07	0.12	15
24	0.50	0.17	0.25	18
accuracy			0.14	314
macro avg	0.15	0.09	0.09	314
weighted avg	0.18	0.14	0.11	314

Gambar 8. Evaluasi Klasifikasi Menggunakan *Confusion Matrix*

Berdasarkan gambar 8, dapat diketahui bahwa hasil klasifikasi dengan menggunakan algoritma K-Nearest Neighbor diperoleh dengan model nilai akurasi sebesar 13,69%.

4.2. Pembahasan

Data *automated essay scoring* diambil dari situs Kaggle.com dengan pilihan *essay set* nomor 7 dengan total 1569 data. Penelitian ini berfokus pada klasifikasi penilaian jawaban ujian esai dengan menggunakan algoritma *K-Nearest Neighbor* (KNN) serta *Knowledge Discovery in Database* (KDD) sebagai metodologinya yang terdiri dari lima proses yaitu, *data selection, pre-processing, transformation, data mining* dan *evaluation.*

Proses pengambilan data diambil dari situs Kaggle.com yaitu *automated essay scoring* dengan jumlah 12.978 data berdasarkan kelas yang berbeda dan *essay set* yang berbeda. *Essay set* nomor 7 yang menjadi pilihan untuk dilakukan pengklasifikasian data, dengan jumlah total 1.569 data esai.

Data yang terpilih akan diproses dengan *preprocessing* yang terdiri dari beberapa tahapan, yaitu:

1. *Cleaning*
   Pembersihan data dari *noise-noise* kata ataupun simbol "!,.&-:;".
2. *Case Folding*
   Penyeragaman huruf kapital menjadi huruf kecil
3. *Tokenizing*
   Memecah kalimat menjadi potongan kata-kata
4. *Stopword Removal*
   Menghapus kata-kata yang tidak penting
5. *Stemming*
   Menyederhanakan kata ke bentuk kata dasar atau akar

Selanjutnya setelah melalui proses *pre-processing* dilakukan proses *transformation* menggunakan *Term Frequency – Inverse Document Frequency* (TF-IDF). Proses *data mining* menjadi proses selanjutnya setelah *transformation,* dengan *K-Nearest Neighbor* sebagai algoritma pada *data mining.*

Dan yang terakhir yaitu proses *evaluation* dengan menggunakan *confusion matrix* untuk mengukur performa dari nilai-nilai seperti *accuracy, precision, recall* dan *f-measure*. Dalam penelitian ini dilakukan pengujian dengan menggunakan empat skenario pembagian data yang berbeda yaitu, 60:40, 70:30, 80:20 dan 90:10. Hasil pengujian tersebut menunjukkan bahwa pada skenario terakhir memperoleh nilai akurasi tertinggi sebesar 77%.

Jawaban ujian esai yang diberikan oleh siswa sangat beragam. Kata "i", "was" menjadi kata yang paling sering muncul pada jawaban ujian esai. Berikut mmerupakan visualisasi kemunculan kata yang sering muncul dalam ujian esai yang dapat dilihat pada gambar

Gambar 4. *Worldcloud* jawaban penilaian ujian esai

Berdasarkan gambar 4. 9 menunjukkan *wordcloud* yang merupakan kata paling sering muncul pada data ujian esai. Kata yang sering muncul diantaranya yaitu, "*was", "patient", "had", "be", "got"* dan sebagainya. Pada visualisasi frekuensi kata melalui *wordcloud* dapat disimpulkan bahwa semakin besar ukuran kata pada *wordcloud,*

maka semakin tinggi juga frekuensi atau jumlah kemunculan dari kata tersebut.

Klasifikasi penilaian jawaban ujian esai menggunakan algoritma *K-Nearest Neighbor* (KNN) dengan k = 29 didapatkan 0 kelas 2, 3, 4, 5, 6, 7, 10, 11, 12 yang diprediksi benar, 114 kelas 8 yang diprediksi benar dan 3 kelas 9 yang diprediksi benar. Dengan demikian, jika akumulasikan terdapat 117 data yang berhassil diprediksi benar dari 157 *data testing*. Klasifikasi menggunakan KNN dan pembobotan TF-IDF ini menghasilkan akurasi sebesar 14,01%. Sedangkan performa lainnya yaitu nilai *precision* diperoleh rata-rata 10,4% dan *recall* dengan rata-rata 14,1%, serta *f-measure* sebesar 5,7%.

5. KESIMPULAN DAN SARAN

Berdasarkan penelitian yang telah dilakukan dapat disimpulkan bahwa beberapa hal, yakni sebagai berikut: Klasifikasi penilaian jawaban ujian esai dilakukan dengan membagi data yang telah melewati proses preprocessisng dan transformation dengan rasio 90% data training yaitu 1412 data dan 30% data testing yaitu 157 data. Hasil dari klasifikasi menggunakan algoritma *K-Nearest Neighbor* (KNN) dengan k = 29 didapatkan bahwa terdapat 0 kelas 2, 3, 4, 5, 6, 7, 10, 11, 12 yang diprediksi benar, 114 kelas 8 yang diprediksi benar dan 3 kelas 9 yang diprediksi benar. Jadi terdapat sejumlah 117 data yang berhasil di prediksi dengan benar dari 157 data testing yang ada. Evaluasi performa dari penerapan algoritma K-Nearest Neighbor (KNN) dilakukan dengan menggunakan confusion matrix. Dari hasil tersebut didapatkan nilai akurasi sebesar 14,01%, precision dengan rata-rata 10,4%, recall dengan rata-rata 10,4% dan f-measure sebesar 5,7%. Kata “was”, “patient”, “had”, “be”, “got” merupakan kata dengan frekuensi kemunculan terbanyak pada data ujian esai. Dari dari kelima kata tersebut, kata “was” merupakan kata yang paling sering muncul.

Berdasarkan hassil penelitian ini, terdapat beberapa saran yang dapat dikembangkan lagi pada penelitian selanjutnya diantaranya, yaitu: Klasifikasi penilaian jawaban ujian esai dapat dilakukan dengan menggunakan metode atau algoritma lainnya. Penelitian ini mengalami jumlah data yang tidak seimbang atau berat sebelah antar perkelas atau perlabelnya, sehingga turut mengaruhi dari hasil klasifikasi. Maka dari itu, lebih baik apabila diterapkan teknik pembagian data lainnya seperti teknik holdout. Sumber data penilaian jawaban ujian esai dapat diperluas seperti dengan mengambil data langsung dari sekolah, perguruan tinggi dan lain sebagainya.

DAFTAR PUSTAKA

[1] D. Pristiwanti, B. Badariah, S. Hidayat, and R. S. Dewi, “Pengertian Pendidikan,” 2022. [Online]. Available: http://repo.iain-

[2] Dapodikdasmen, “Data Peserta Didik Nasional - Dapodikdasmen,” 2023.

[3] D. Sebastian, “Implementasi Algoritma K-Nearest Neighbor untuk Melakukan Klasifikasi Produk dari beberapa E-marketplace,” vol. 5, pp. 2443–2229, 2019, doi: 10.28932/jutisi.v5i1.913.

[4] S. Ati, Nurdien, Kistanto, and A. Taufik, “Pengantar Konsep Informasi, Data, dan Pengetahuan,” 2014.

[5] E. Setiawan, “PEMAHAMAN MASYARAKAT TENTANG PENERAPAN AKUNTANSI PADA USAHA MIKRO KECIL DAN MENENGAH (UMKM),” 2021.

[6] H. Saragih, E. Buulolo, and F. Tinus Waruwu, “IMPLEMENTASI DATA MINING PENYESUAIAN JENIS LENSA TERHADAP KEBUTUHAN PASIEN DENGAN MENGGUNAKAN ALGORITMA C4.5,” 2017, [Online]. Available: http://ejurnal.stmik-budidarma.ac.id/index.php/komik

[7] R. Setiawan, “PENERAPAN DATA MINING MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING UNTUK MENENTUKAN STRATEGI PROMOSI MAHASISWA BARU (Studi Kasus: Politeknik LP3I Jakarta),” 2016.

[8] F. E. Kurniawati and W. Mega Pradnya, “Implementasi Algoritma Winnowing Pada Sistem Penilaian Otomatis Jawaban Esai Pada Ujian Online Berbasis Web,” *Jurnal Teknik Komputer AMIK BSI*, vol. VI, no. 2, 2020, doi: 10.31294/jtk.v4i2.

[9] Jo. T, “Intorduction in: Text Mining In Studies in Big Data,” pp. 3–17, 2019.

[10] D. Susandi, U. Sholahudin, J. Raya, C. Serang - Drangong, and K. Serang, “Pemanfaatan Vector Space Model pada Penerapan Algoritma Nazief Adriani, KNN dan Fungsi Similarity Cosine untuk Pembobotan IDF dan WIDF pada Prototipe Sistem Klasifikasi Teks Bahasa Indonesia,” 2016.

[11] M. Rivki and A. M. Bachtiar, “IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN FOLLOWER TWITTER YANG MENGGUNAKAN BAHASA INDONESIA,” *Jurnal Sistem Informasi*, vol. 13, no. 1, p. 31, May 2017, doi: 10.21609/jsi.v13i1.500.