Penilaian Esai Pendek Otomatis dengan Pencocokan Kata Kunci Frasa Nomina

Short Essay Autoscoring using Noun Phrase - Keyword Matching

Nurul Chamidah[1], Mayanda Mega Santoni[2], Helena Nurramdhani Irmanda[3], Ria Astriratma[4]
[1,2,3,4]Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta
E-mail: [1]nurul.chamidah@upnvj.ac.id, [2]megasantoni@upnvj.ac.id,
[3]helenairmanda@upnvj.ac.id, [4]astriratma@upnvj.ac.id

Abstrak

Pembelajaran daring menjadi suatu kebutuhan dalam pengajaran baik dalam memberikan materi maupun ujian. Ujian dalam bentuk soal objektif kurang dapat mengukur kemampuan pemahaman seseorang dan soal esai dianggap lebih baik untuk mengevaluasi hasil pembelajaran. Namun, jawaban berbentuk esai memerlukan waktu yang lebih banyak untuk dilakukan penilaian serta hasil penilaiannya dapat inkonsisten. Maka dari itu, diperlukan suatu sistem penilaian esai otomatis yang dapat menilai esai dengan lebih cepat dan konsisten. Penelitian ini dilakukan untuk menganalisis performa penilain esai otomatis dengan mengekstrak kata kunci dari frasa nomina dalam jawaban berbentuk esai pendek. Penilaian esai dilakukan dengan mencocokkan kata kunci yang diekstrak dari jawaban uji dan jawaban referensi. Jawaban uji dan referensi diproses dengan *case folding*, *Part of Speech (POS) Tagging*, ekstraksi frasa nomina, dan *stemming*. Kata kunci unik jawaban uji dan jawaban referensi yang diperoleh dari proses tersebut selanjutnya dicocokkan dan kemudian dinilai berdasarkan kecocokan tersebut. Hasil evaluasi penelitian ini menunjukkan *Mean Absolute Error* (MAE) dari nilai yang diperoleh dengan mencocokkan kata kunci dengan nilai uji yang diberikan manusia sebesar 18% dan *Pearson Correlation* sebesar 0.83 yang menunjukkan korelasi antara nilai sistem dan nilai uji sangat baik.

Kata kunci: Penilaian Esai Otomatis, Pencocokan Kata Kunci, Frasa Nomina, Esai Pendek

Abstract

Online learning has become a necessity in teaching both in providing material and in examinations. Objective form exam questions are less able to measure a person's understanding ability and essay questions are considered better for evaluating learning outcomes. However, answers in essay's form require more time to be assessed and the results of the assessment can be inconsistent. Therefore, we need an automatic essay scoring system that can assess essays more quickly and consistently. This research was conducted to analyze the performance of automatic essay scoring by extracting keywords from noun phrases in short essay answers. Essay assessment is done by matching keywords extracted from test answers and reference answers. Test answers and references are processed by case folding, Part of Speech (POS) Tagging, noun phrase extraction, and stemming. The unique keywords of the test answer and the reference answer obtained from the previous process are then matched and assessed based on the match. The results of the evaluation of this study indicate the Mean Absolute Error (MAE) of the scores obtained by matching keywords with the real test score given by humans of 18% and Pearson Correlation of 0.83 which shows the correlation between system scores and real test scores is very good.

Keywords: Automatic Essay Scoring, Keywords Matching, Noun Phrase, Short Essay

1. PENDAHULUAN

Pembelajaran daring yang berkembang pada akhir-akhir ini, menjadi suatu kebutuhan dalam pendidikan dalam masa pandemi. Media pembelajaran online juga telah berkembang

dengan pesat dengan Learning Management System (LMS) dari yang gratis seperti google classroom hingga yang memerlukan infrastruktur seperti menggunakan moodle. Dalam pembelajaran daring, pengajar dapat memberikan materi baik berupa teks, video, maupun gambar. Selain memberikan materi, pengajar juga melaksanakan evaluasi seperti ujian, kuis, serta tugas.

Pada umumnya, ujian dapat berbentuk soal objektif maupun soal esai. Soal objektif biasanya berbentuk pilihan ganda, benar salah, atau pencocokan cenderung lebih mudah diterapkan karena pada umumnya LMS telah memiliki sistem untuk menilai jawaban dari soal objektif ini dimana pengajar memasukkan kunci jawaban ke dalam sistem. Namun, bentuk soal objektif memiliki kekurangan yakni sulit untuk mengetahui tingkat pemahaman seseorang yang memerlukan penjelasan [1].

Berbeda dengan objektif, soal esai yang dianggap lebih tepat digunakan untuk mengukur kedalaman pengetahuan dalam pembelajaran. Soal berbentuk esai lebih sulit untuk dinilai karena harus dilakukan secara manual. Oleh karena itu, jawaban dari bentuk soal esai ini memerlukan waktu yang lebih lama dari pada menilai soal yang berbentuk objektif. Selain masalah waktu, dapat terjadi masalah konsistensi pada evaluasi jawaban berbentuk esai, baik dengan penilai yang sama maupun antar penilai [2]. Pada penilai yang sama, penilaian dapat inkonsisten jika dilakukan penilaian pada waktu yang berbeda, sedangkan antar penilai mungkin terdapat perbedaan pendapat sehingga jawaban yang sama dapat memiliki penilaian yang berbeda. Maka dari itu, diperlukan suatu penilai esai otomatis agar evaluasi pada jawaban berbentuk esai lebih cepat dan konsisten.

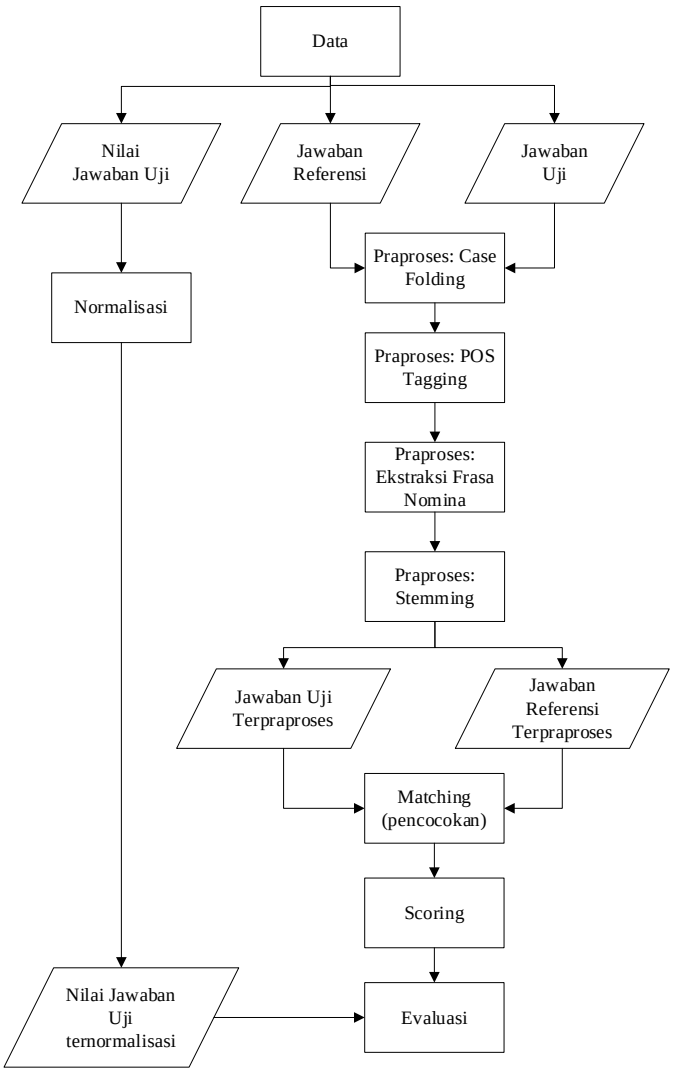
Penelitian dalam penilaian esai otomatis berbahasa Indonesia telah banyak dilakukan tapi metode yang paling cocok belum ditemukan sehingga penelitian pada ranah ini masih sangat diperlukan. Metode pengembangan sistem penilaian esai otomatis ini banyak dilakukan dengan mencocokkan kata kunci antara jawaban uji (jawaban yang akan dinilai) dan jawaban referensi (kunci jawaban). Penelitian [2][3] menggunakan *Latent Semantic Analysis* (LSA) dengan memetakan kata kunci jawaban uji dengan jawaban referensi dalam bentuk matriks. Penelitian [4] menggunakan *Vector Space Model*, yakni dengan melakukan pembobotan kata menggunakan teknik *Term Frequency-Inverse Document Frequency* (TF-IDF) kemudian menghitung similaritas antara kata kunci pada jawaban uji dengan kata kunci pada jawaban referensi dengan *cosine similarity*. Perhitungan similaritas selain dengan *cosine similarity*, penelitian [5] menghitung similaritas berdasarkan kedekatan antara jawaban referensi dengan jawaban uji menggunakan *K-Nearest Neighbour* (KNN). Sedangkan penelitian [6] menggunakan *Manhattan distance* dan *dice similarity* untuk menghitung kesesuaian jawaban uji dan referensi.

Penelitian [7] dilakukan dengan mencocokkan kata kunci antara jawaban uji dengan kata kunci jawaban referensi dengan mengambil seluruh kata yang dipraproses dengan *case folding*, *stopword removal*, pengambilan kata dasar atau *stemming* dengan algoritma Nazief-Adriani, dan tokenisasi yang kemudian hasilnya dievaluasi dengan menghitung *error* dengan *Mean Absolute Error* (MAE) serta korelasi dengan *Pearson Correlation*. Pada penelitian ini, kami menambahkan penggunaan teknik *Natural Language Processing* (NLP) dengan mengambil kata kunci hanya yang berasal dari frasa nomina yang dihasilkan dari *Part of Speech* (POS) *Tagging* dengan memanfaatkan aturan Bahasa Indonesia.

2. METODE PENELITIAN

Metodologi penelitian yang digunakan dalam penelitian ini dapat dilihat pada Gambar 1. Data yang digunakan dalam penelitian ini terdiri dari teks jawaban uji dan jawaban referensi (kunci jawaban), dan data numerik berupa nilai uji yang merupakan penilaian setiap jawaban oleh evaluator serta nilai maksimum dari soal yang akan digunakan untuk menormalisasi nilai uji. Data teks dari jawaban uji dan jawaban referensi dipraproses dengan *case folding*, *POS Tagging*, ekstraksi frasa nomina, serta *stemming*. Hasil praproses berupa kata kunci yang berasal dari frasa nomina pada jawaban uji dan jawaban referensi kemudian dicocokkan. Hasil pencocokan ini menjadi dasar untuk melakukan *scoring* (penilaian) dan kemudian hasil penilaian sistem ini dievaluasi dengan membandingkan skor atau nilai sistem dengan skor hasil penilaian manusia

atau nilai jawaban uji untuk mengetahui performa sistem penilaian esai otomatis.



Gambar 1 Metodologi Penelitian

2.1. Dataset

Dataset diperoleh dari ujian mata kuliah pengantar basis data dengan bentuk esai pendek dimana soal ini bukan berupa karangan bebas sehingga memiliki kunci jawaban atau jawaban referensi. Soal esai pendek ini berjumlah 4 yang masing-masing memiliki satu jawaban referensi dan nilai soal dimana nilai soal merupakan nilai maksimum jika jawaban benar sepenuhnya. Soal dikerjakan oleh 36 orang mahasiswa sehingga total jawaban yang akan dievaluasi dalam penelitian ini sebanyak 144 jawaban

. Komponen dataset ini terdiri dari 5 bagian, yakni soal, nilai maksimum dari soal, jawaban referensi (kunci jawaban), jawaban uji (jawaban mahasiswa), serta nilai uji (nilai jawaban uji), contoh dataset dapat dilihat pada Tabel 1. Selanjutnya, jawaban uji dan jawaban referensi akan dipraproses dengan *case folding*, *POS tagging*, ekstraksi frasa nomina, *stemming*, serta tokenisasi. Sedangkan nilai soal digunakan untuk menormalisasi nilai uji.

Soal (Nilai Soal = 10)	Model data merupakan suatu cara untuk menjelaskan bagaimana pemakai dapat melihat data secara logic. Bandingkan model data berbasis objek – Entity Relationship Model dengan model data berbasis record - Relational model, jelaskan perbedaannya!
Jawaban Referensi	Model data berbasis objek – Entity Relationship Model digunakan untuk menjelaskan hubungan antar data dalam basis data berdasarkan persepsi bahwa dunia nyata terdiri dari objek yang memiliki hubungan antar objek. Sedangkan pemodelan data berbasis record Relational Model digunakan untuk menjelaskan hubungan logic antardata dalam basis data dengan visualisai menggunakan tabel-tabel yang menunjukkan atribut tertentu.
Jawaban Uji	Model data berbasis record. Pada model ini menjelaskan data pada tingkat konsepsi dan. view, memakai seluruh struktur lo. jik basis data dan menyediakan uraian tingkat tinggi dari. implementasi. Terdiri dari sejumlah fixed format record dengan berbagai tipe. Pada model. ini teradapat 3 (tiga) macam tipe,yaitu :. . 1.Model data relational.. 2.model data nerwork, dan. 3.model data hirarki.. . Model data entity relationship (E-R). adalah berdasarkan pada persepsi dari dunia nyata yang terdiri. dari sekumpulan objek. objek dasar yang desebut entity, dan relationship antara objek. objek. tersebut. Pembentuk Model E-R (Entity Relationship) pada dasarnya terdiri dari 2 komponen, yaitu. Entiti (Entity) / entitas dan Relasi (Relation) atau hubungan.
Nilai Uji	7

Gambar 2 Tabel Contoh Dataset

2.2. Case Folding

Praproses pertama dilakukan *case folding* untuk mengubah semua huruf dalam dokumen menjadi huruf kecil serta menghilangkan karakter-karakter selain huruf abjad [8]. Contoh hasil praproses *case folding* pada jawaban uji dan jawaban referensi ditunjukkan pada Tabel 2 dari contoh *dataset* pada Tabel 1. Dapat dibandingkan antara Tabel 1 dan Tabel 2, *case folding* ini mengubah huruf besar pertama pada kata 'Model', 'Entity', 'Relationship', 'Sedangkan', 'Relational', dan sebagainya menjadi huruf kecil. Selain itu, tanda baca seperti strip, kurung, titik, koma, dan lainnya juga dihilangkan baik pada jawaban uji maupun jawaban referensi.

Jawaban referensi	Jawaban Uji
model data berbasis objek entity relationship model digunakan untuk menjelaskan hubungan antar data dalam basis data berdasarkan persepsi bahwa dunia nyata terdiri dari objek yang memiliki hubungan antar objek sedangkan pemodelan data berbasis record relational model digunakan untuk menjelaskan hubungan logic antardata dalam basis data dengan visualisai menggunakan tabel tabel yang menunjukkan atribut tertentu	model data berbasis record pada model ini menjelaskan data pada tingkat konsepsi dan view memakai seluruh struktur lo jik basis data dan menyediakan uraian tingkat tinggi dari implementasi terdiri dari sejumlah fixed format record dengan berbagai tipe pada model ini teradapat tiga macam tipe yaitu model data relational model data nerwork dan model data hirarki model data entity relationship adalah berdasarkan pada persepsi dari dunia nyata yang terdiri dari sekumpulan objek objek dasar yang desebut entity dan relationship antara objek objek tersebut pembentuk model entity relationship pada dasarnya terdiri dari komponen yaitu entiti entity entitas dan relasi relation atau hubungan

Gambar 3 Tabel Contoh Hasil *Case folding* pada Jawaban Referensi dan Jawaban Uji

2.3. POS Tagging

Setelah jawaban di praproses dengan *case folding* serta menghilangkan karakter non-abjad, selanjutkan jawaban dilabeli dengan *Part of Speech* (POS) *Tag*. Proses pelabelan menggunakan *POS Tagger* dengan memanfaatkan Stanford NLP Bahasa Indonesia (yohanes gultom, 2018)(Chen, manning 2014). Dari hasil *POS Tagging*, diambil suatu frasa nomina, yakni frasa yang memiliki kelas berupa kata benda baik sebagai predikat maupun objek (wulandari, 2018). Tabel 3 menunjukkan POS tag yang terbentuk dari jawaban referensi dan jawaban uji.

Tabel 1 POS Tag Jawaban Uji dan Jawaban Referensi.

Jawaban referensi	Jawaban Uji	
(S (NP model/NOUN data/NOUN) berbasis/VERB (NP objek/NOUN entity/NOUN relationship/NOUN model/NOUN) digunakan/VERB untuk/ADP menjelaskan/VERB (NP hubungan/NOUN) antar/ADP (NP data/NOUN) dalam/ADP	(S (NP model/NOUN data/NOUN) berbasis/VERB (NP record/NOUN) pada/ADP (NP model/NOUN) ini/DET menjelaskan/VERB (NP data/NOUN) pada/ADP (NP tingkat/NOUN konsepsi/NOUN) dan/CCONJ	(NP macam/NOUN tipe/NOUN) yaitu/ADV (NP model/NOUN data/NOUN relational/ADJ) (NP model/NOUN data/NOUN nerwork/NOUN) dan/CCONJ (NP model/NOUN data/NOUN hirarki/NOUN model/NOUN

(NP basis/NOUN data/NOUN) berdasarkan/VERB (NP persepsi/NOUN) bahwa/SCONJ (NP dunia/NOUN nyata/ADJ) terdiri/VERB dari/ADP (NP objek/NOUN) yang/PRON memiliki/VERB (NP hubungan/NOUN) antar/ADP (NP objek/NOUN) sedangkan/SCONJ (NP pemodelan/NOUN data/NOUN) berbasis/VERB (NP record/NOUN relational/ADJ) (NP model/NOUN) digunakan/VERB untuk/ADP menjelaskan/VERB (NP hubungan/NOUN logic/NOUN antardata/NOUN) dalam/ADP (NP basis/NOUN data/NOUN) dengan/ADP visualisai/ADJ menggunakan/VERB (NP tabel/NOUN tabel/NOUN) yang/PRON menunjukkan/VERB (NP atribut/NOUN) tertentu/DET)	(NP view/NOUN) memakai/VERB seluruh/DET (NP struktur/NOUN lo/NOUN jik/NOUN basis/NOUN data/NOUN) dan/CCONJ menyediakan/VERB (NP uraian/NOUN tingkat/NOUN tinggi/ADJ) dari/ADP (NP implementasi/NOUN) terdiri/VERB dari/ADP sejumlah/DET (NP komponen/NOUN) yaitu/ADV (NP entiti/NOUN entity/NOUN entitas/NOUN) dan/CCONJ (NP relasi/NOUN relation/NOUN) atau/CCONJ (NP hubungan/NOUN)) (NP fixed/NOUN format/PROPN record/PROPN) dengan/ADP berbagai/DET (NP tipe/NOUN) pada/ADP (NP model/NOUN) ini/DET teradapat/VERB tiga/NUM	data/NOUN entity/NOUN relationship/NOUN) adalah/AUX berdasarkan/VERB pada/ADP (NP persepsi/NOUN) dari/ADP (NP dunia/NOUN nyata/ADJ) yang/PRON terdiri/VERB dari/ADP sekumpulan/DET (NP objek/NOUN objek/NOUN dasar/ADJ) yang/PRON desebut/VERB (NP entity/NOUN) dan/CCONJ (NP relationship/NOUN) antara/ADP (NP objek/NOUN objek/NOUN) tersebut/DET (NP pembentuk/NOUN model/NOUN entity/NOUN relationship/NOUN) pada/ADP (NP dasarnya/NOUN) terdiri/VERB dari/ADP

2.4. Ekstraksi Frasa Nomina

Hasil POS Tagging berupa frasa nomina ditandai dengan NP (Noun Phrase). Dapat dilihat pada Tabel 3, kata dalam kelas NP dapat terdiri satu atau lebih kata yang merupakan frasa. Semua kata dalam frasa NP kemudian di ekstrak untuk dijadikan kata kunci, contoh pada (NP model/NOUN data/NOUN) ekstraksi frasa nominanya adalah ‘model data’. Hasil ekstraksi frasa nomina yang ditandai dengan NP dapat dilihat pada Tabel 4. Dapat dilihat pada tabel tersebut, kata-kata dalam kurung siku merupakan satu frasa nomina (NP).

Tabel 2 Frasa Nomina (NP) Jawaban Uji dan Jawaban Referensi

Jawaban referensi	Jawaban Uji
['model', 'data'], ['objek', 'entity', 'relationship', 'model'], ['hubungan'], ['data'], ['basis', 'data'], ['persepsi'], ['dunia', 'nyata'], ['objek'], ['hubungan'], ['objek'], ['pemodelan', 'data'], ['record', 'relational'], ['model'], ['hubungan', 'logic', 'antardata'], ['basis', 'data'], ['tabel', 'tabel'], ['atribut']	['model', 'data'], ['record'], ['model'], ['data'], ['tingkat', 'konsepsi'], ['view'], ['struktur', 'lo', 'jik', 'basis', 'data'], ['uraian', 'tingkat', 'tinggi'], ['implementasi'], ['fixed', 'format', 'record'], ['tipe'], ['model'], ['macam', 'tipe'], ['model', 'data', 'relational'], ['model', 'data', 'nerwork'], ['model', 'data', 'hirarki', 'model', 'data', 'entity', 'relationship'], ['persepsi'], ['dunia', 'nyata'], ['objek', 'objek', 'dasar'], ['entity'], ['relationship'], ['objek', 'objek'], ['pembentuk', 'model', 'entity', 'relationship'], ['dasarnya'], ['komponen'], ['entiti', 'entity', 'entitas'], ['relasi', 'relation'], ['hubungan']

2.5. Stemming

Stemming adalah teknik mendapatkan kata dasar atau *stem* dari suatu kata dengan menghilangkan imbuhannya, baik awalan (prefiks), akhiran (sufiks), maupun awalan dan akhiran [9]. Penelitian menggunakan algoritma Nazief-Adriani untuk mengambil kata dasar atau *stem*. *Stemming* ini dilakukan untuk menghilangkan variasi kata karena imbuhan yang membentuk kata bentuk pasif, bentuk aktif, pembendaan dan sebagainya dan hanya mengambil kata dasarnya saja.

Tabel 3 Hasil *Stemming*

Jawaban referensi	Jawaban Uji
'model', 'data', 'objek', 'entity', 'relationship', 'model', 'hubung', 'data', 'basis', 'data', 'persepsi',	'model', 'data', 'record', 'model', 'data', 'tingkat', 'konsepsi', 'view', 'struktur', 'lo', 'jik', 'basis', 'data', 'urai', 'tingkat', 'tinggi', 'implementasi', 'fixed', 'format',

'dunia', 'nyata', 'objek', 'hubung', 'objek', 'model', 'data', 'record', 'relational', 'model', 'hubung', 'logic', 'antardata', 'basis', 'data', 'tabel', 'tabel', 'atribut'	'record', 'tipe', 'model', 'macam', 'tipe', 'model', 'data', 'relational', 'model', 'data', 'nerwork', 'model', 'data', 'hirarki', 'model', 'data', 'entity', 'relationship', 'persepsi', 'dunia', 'nyata', 'objek', 'objek', 'dasar', 'entity', 'relationship', 'objek', 'objek', 'bentuk', 'model', 'entity', 'relationship', 'dasar', 'komponen', 'entit', 'entity', 'entitas', 'relasi', 'relation', 'hubung'

Tabel 5 menunjukkan hasil stemming dari kata yang diekstrak dari frasa nomina sebelumnya. Contoh stemming ini seperti pada kata ‘hubungan’ menjadi ‘hubung’, ‘pemodelan’ menjadi ‘model’, dan lain-lain. Selanjutnya, hasil stemming antara jawaban uji dan jawaban referensi akan dibandingkan. Namun, sebelum kedua teks dibandingkan, dilakukan penghilangan duplikasi kata. Tabel 6 menunjukkan kata unik hasil penghilangan duplikasi, misal pada jawaban referensi Tabel 5 terdapat kata ‘data’ sebanyak empat kali, setelah dilakukan penghilangan duplikasi hanya terdapat satu kata ‘data’.

Tabel 4 Kata Unik Jawaban Uji dan Jawaban Referensi

Jawaban referensi	Jawaban Uji
'relational', 'tabel', 'logic', 'dunia', 'model', 'basis', 'antardata', 'entity', 'hubung', 'objek', 'data', 'relationship', 'record', 'nyata', 'persepsi', 'atribut'	'entit', 'nerwork', 'fixed', 'view', 'entity', 'data', 'nyata', 'format', 'hubung', 'relasi', 'macam', 'implementasi', 'objek', 'entitas', 'lo', 'hirarki', 'tingkat', 'persepsi', 'tinggi', 'model', 'urai', 'dasar', 'relation', 'record', 'bentuk', 'struktur', 'relational', 'dunia', 'basis', 'konsepsi', 'komponen', 'relationship', 'jik', 'tipe'

2.6. Matching

Pencocokan atau proses *matching* dilakukan untuk melihat kecocokan antara jawaban uji terhadap jawaban referensi dimana kata kunci atau *keyword* didapatkan dengan mengambil kata unik dari frasa nomina yang telah di *stemming* pada proses sebelumnya. Contoh *matching* ini dapat dilihat pada Tabel 7. Dapat diketahui pada Tabel 7 tersebut, bahwa jawaban uji dan jawaban referensi memiliki irisan sebanyak 12 kata, yakni kata: 'entity', 'data', 'nyata', 'hubung', 'objek', 'persepsi', 'model', 'record', 'relational', 'dunia', 'basis', 'relationship'.

Tabel 5 Pencocokan Kata Kunci

Jawaban referensi	Jawaban Uji
'**relational**', 'tabel', 'logic', '**dunia**', '**model**', '**basis**', 'antardata', '**entity**', '**hubung**', '**objek**', '**data**', '**relationship**', '**record**', '**nyata**', '**persepsi**', 'atribut'	'entit', 'nerwork', 'fixed', 'view', '**entity**', '**data**', '**nyata**', 'format', '**hubung**', 'relasi', 'macam', 'implementasi', '**objek**', 'entitas', 'lo', 'hirarki', 'tingkat', '**persepsi**', 'tinggi', '**model**', 'urai', 'dasar', 'relation', '**record**', 'bentuk', 'struktur', '**relational**', '**dunia**', '**basis**', 'konsepsi', 'komponen', '**relationship**', 'jik', 'tipe'

2.7. Scoring

Scoring atau penilaian dihitung berdasarkan kecocokan antara jawaban uji dan jawaban referensi. Sebagai contoh hasil matching sebelumnya, pada Tabel 7 dapat dilihat jumlah kata unik pada jawaban referensi (N) adalah 16 dan kata unik pada jawaban uji sebanyak 34 dengan kata yang beririsan dengan jawaban referensi (n) sebanyak 12. Sehingga, penilaian untuk jawaban uji tersebut adalah n/N atau 12/16=0.75 Pada *scoring* ini, jawaban uji akan bernilai 0 jika tidak ada kata yang cocok dengan jawaban referensi, dan 1 jika semua kata pada jawaban referensi ada dalam jawaban uji.

2.8. Normalisasi data nilai

Nilai maksimal dari suatu soal dapat berbeda dengan soal lainnya, maka dari itu nilai uji dari setiap soal dinormalisasi dengan membandingkannya dengan nilai maksimum soal. Dengan melakukan normalisasi, skala nilai uji dari seluruh soal menjadi sama yakni antara 0 hingga 1. Normalisasi nilai ini dilakukan dengan menerapkan normalisasi *min-max* [10]. Teknik normalisasi *min-max* dipilih dalam penelitian ini karena nilai dari suatu jawaban memiliki *range* yang pasti, yakni nilai minimum 0 jika salah total, dan nilai maksimum berupa nilai jawaban uji. Karena nilai minimum soal adalah 0 dan maksimumnya adalah 1, maka normalisasi *min-max* untuk nilai uji dapat dilihat pada rumus 1.

$$n_{baru} = \frac{n_{uji}}{n_{soal}} \quad (1)$$

Dengan n_{baru} merupakan nilai hasil normalisasi yang akan dicari, n_{uji} adalah nilai dari jawaban uji yang diberikan oleh evaluator, dan n_{soal} merupakan nilai soal atau nilai maksimum jika jawaban uji merupakan jawaban yang tepat. Nilai n_{baru} hasil normalisasi ini selanjutnya digunakan untuk mengevaluasi performa sistem. Dapat dilihat pada contoh Tabel 1, nilai soal $n_{soal} = 10$, dan nilai uji $n_{uji} = 7$, maka nilai baru n_{baru} hasil normalisasi adalah 7/10 = 0.7.

2.9. Evaluasi performa

Performa sistem dievaluasi dengan menghitung nilai *Mean Absolute Error* (MAE) dapat dilihat pada rumus 2 [11]. Yakni, kesalahan antara nilai yang diperoleh dari *scoring* oleh sistem dengan nilai uji yang diberikan oleh evaluator atau pengajar.

$$MAE = \frac{\sum|n_{sistem} - n_{uji}|}{Jum} \quad (2)$$

Dengan n_{sistem} merupakan nilai dari hasil *scoring* oleh sistem, n_{uji} berupa nilai jawaban uji yang merupakan hasil penilaian dari evaluator, dan Jum merupakan total jawaban yang dievaluasi.

Selain dengan MAE, evaluasi performa juga dilakukan dengan mencari korelasi. Korelasi ini digunakan untuk mencari tingkat kesepakatan atau kesesuaian antara nilai jawaban uji yang diberikan oleh evaluator manusia dengan nilai yang diperoleh dari sistem penilaian esai otomatis. Nilai korelasi dalam penelitian ini menggunakan *Pearson Correlation* yang dapat dilihat pada rumus (3), yakni perbandingan antara kovarian dengan perkalian standar deviasi dari nilai uji dan nilai hasil *scoring* oleh sistem.

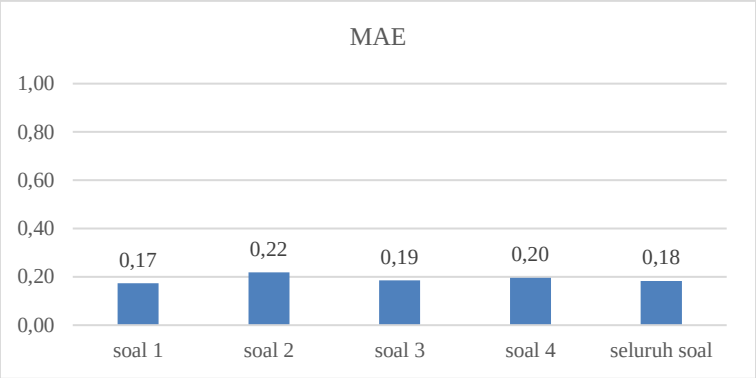
$$corr(n_{sistem}, n_{uji}) = \frac{cov(n_{sistem},\ n_{uji})}{stdev(n_{sistem}).stdev(n_{uji})} \quad (3)$$

Dengan $corr$ adalah nilai korelasi antara nilai hasil *scoring* oleh sistem dengan nilai uji dengan rentang nilai antara 0 hingga 1, cov adalah kovarian antara nilai hasil sistem penilaian otomatis dengan nilai jawaban uji, dan $stdev$ berupa standar deviasi untuk masing-masing nilai sistem dan nilai jawaban uji. Kriteria korelasi untuk menentukan kesuksesan sistem adalah kurang jika korelasi < 0.4, baik jika korelasi bernilai antara 0.4 hingga 0.75, dan sangat baik jika nilai korelasi > 0.75 [12].

3. HASIL DAN PEMBAHASAN

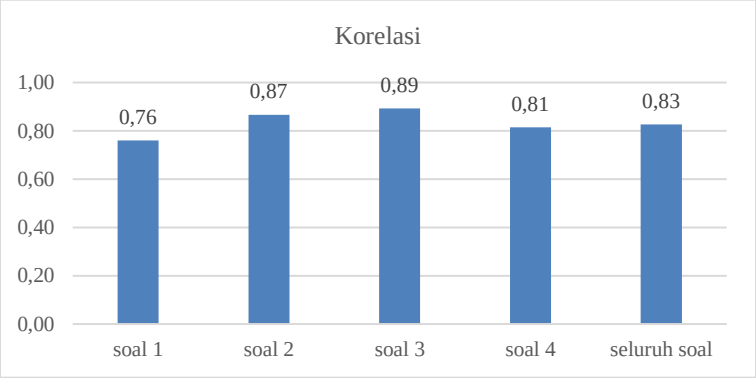
Penelitian dilakukan terhadap data yang terdiri dari 4 soal, 4 jawaban referensi dengan setiap soal dijawab oleh 36 orang. Sehingga terdapat 36 jawaban uji untuk setiap soal dengan nilai jawaban uji dan nilai soal diketahui. Pemrosesan pada teks jawaban referensi dan jawaban uji dilakukan dengan *case folding*, *POS Tagging*, ekstraksi frasa nomina, serta *stemming*. Kata unik hasil stemming dari jawaban uji selanjutnya dibandingkan dengan kata unik pada jawaban referensi untuk mendapatkan kata yang beririsan dan kemudian dilakukan *scoring* atau penilaian.

Hasil penilaian dari sistem penilai ini kemudian dievaluasi dengan menghitung *Mean Absolute Error* (MAE), yakni membandingkan hasil penjumlahan dari selisih nilai sistem dan nilai uji dengan total jawaban yang dievaluasi. Nilai uji sebelum digunakan dalam evaluasi MAE, terlebih dahulu dinormalisasi agar nilai uji dan nilai sistem berada dalam rentang nilai yang sama.



Gambar 4 Nilai Mean Absolute Error (MAE)

Hasil evaluasi dengan MAE dapat dilihat pada Gambar 2. Sebagaimana ditunjukkan pada gambar tersebut, MAE dari hasil penilaian dengan teknik pencocokan kata kunci yang diekstrak dari frasa nomina pada soal 1 adalah 17%, soal 2 sebesar 22%, soal 3 sebesar 19%, dan soal 4 sebesar 20%. Sedangkan MAE untuk keseluruhan soal dengan total 154 jawaban uji, diperoleh MAE sebesar 18%.



Gambar 5 Korelasi

Pearson Correlation antara hasil *scoring* dengan nilai uji yang telah dinormalisasi dapat dilihat pada Gambar 3. Gambar 3 tersebut menunjukkan bahwa soal 1,2, 3, dan 4 memiliki korelasi dengan kategori sangat baik dimana nilai hasil *scoring* dan nilai uji memiliki korelasi lebih dari 0.75 dan untuk keseluruhan hasil penilaian juga dalam kategori sangat baik dengan korelasi 0.83. Hal ini menunjukkan bahwa teknik penilaian esai pendek yang digunakan dalam penelitian ini sudah memiliki korelasi yang sangat baik dengan penilaian esai yang dilakukan manusia.

Kesalahan dalam penilaian disebabkan oleh perbedaan penggunaan istilah, kesalahan penulisan, maupun penggunaan Bahasa asing. Misal dalam Tabel 7, pada jawaban referensi terdapat kata ‘logic’ sedangkan pada jawaban uji terjadi kesalahan penulisan ‘lo.jik’ dan di proses menjadi ‘lo’ dan ‘jik’, sehingga dianggap sebagai 2 kata yang berbeda. Pada kasus lain, terdapat jawaban referensi yang tidak beririsan dengan jawaban uji seperti kata ‘pakai’ dimana kata awalnya adalah ‘pemakai’ sedangkan pada jawaban uji menggunakan istilah ‘user’ yang merupakan istilah bahasa asing yang dapat diterima oleh evaluator manusia tapi sistem tidak dapat mengenali.

4. KESIMPULAN DAN SARAN

Penilaian esai otomatis dalam penelitian ini mengambil kata kunci yang diekstrak dari frasa nomina. Hasil ekstraksi kata kunci jawaban uji dicocokkan dengan ekstraksi kata kunci jawaban referensi dan dilakukan penilaian. Dari hasil eksperimen diperoleh kesimpulan bahwa

teknik penilaian esai pendek menggunakan pencocokan kata dengan ekstraksi kata kunci dari frasa nomina ini menghasilkan korelasi yang sangat baik dengan penilaian yang dilakukan oleh evaluator manusia. Meskipun demikian, kesalahan sistem masih cukup besar dengan MAE keseluruhan mencapai 18% yang menunjukkan masih terdapat perbedaan sebesar 18% antara nilai dari manusia dengan nilai dari sistem.

Pengembangan penilaian esai otomatis masih sangat diperlukan. Hal-hal yang mungkin bisa dikembangkan untuk mengatasi variasi Bahasa asing misalnya menggunakan mesin penerjemah untuk istilah asing, memperbanyak jawaban referensi dengan menyediakan alternatif jawaban, menggunakan pendekatan machine learning bila memiliki data yang cukup banyak dapat digunakan untuk pelatihan model dengan deep learning.

DAFTAR PUSTAKA

[1] H. Rababah and A. T. Al-Taani, "An automated scoring approach for Arabic short answers essay questions," in *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, Oct. 2017, pp. 697–702. doi: 10.1109/ICITECH.2017.8079930.

[2] R. Adhitia and A. Purwarianti, "Penilaian Esai Jawaban Bahasa Indonesia Menggunakan Metode SVM - LSA Dengan Fitur Generik," *Jurnal Sistem Informasi*, vol. 5, no. 1, p. 33, Jul. 2012, doi: 10.21609/jsi.v5i1.260.

[3] R. B. Aji, Z. A. Baisal, and Y. Firdaus, "Automatic Essay Grading System Menggunakan Metode Latent Semantic Analysis E-78 E-79," *Seminar Nasional Aplikasi Teknologi Informasi*, vol. 2011, no. Snati, pp. 1–9, 2011.

[4] J. Zeniarja, A. Salam, and I. Achsanu, "Sistem Koreksi Jawaban Esai Otomatis (E-Valuation) dengan Vector Space Model pada Computer Based Test (CBT)," *Seri Prosiding Seminar Nasional Dinamika Informatika*, vol. 4, no. 1, Apr. 2020.

[5] M. Jamaluddin, N. Yuniarti, A. Rahmani, and J. Hutahaean, "Aplikasi Penilaian Otomatis Ujian Esai Berbahasa Indonesia Menggunakan Algoritma K-Nearest Neighbor (Studi kasus MAN Cimahi)," *Prosiding Industrial Research Workshop and National Seminar*, vol. 10, no. August 2019, pp. 314–324, Aug. 2020, doi: 10.35313/irwns.v10i1.1404.

[6] F. Rahutomo, Y. P. Putra, and M. H. Ali, "Implementasi Manhattan Distance dan Dice Similarity pada Ujian Esai Daring Berbahasa Indonesia," *Seminar Informatika Aplikatif Polinema*, pp. 171–174, 2019.

[7] N. Chamidah and M. M. Santoni, "Pencocokan Berbasis Kata Kunci pada Penilaian Esai Pendek Otomatis Berbahasa Indonesia," *Techno.Com*, vol. 20, no. 1, pp. 19–27, Feb. 2021, doi: 10.33633/tc.v20i1.4115.

[8] F. Pratama, "Rancang Bangun Aplikasi Peringkas Tkes Otomatis Artikel Berbahasa Indonesia Mengunakan Metode Term Frequency Inverse Document Frequency (TF-IDF) dan K-mean Clustering," *Fakultas Sains dan Teknologi*, Apr. 2014.

[9] D. Wahyudi, T. Susyanto, and D. Nugroho, "Implementasi dan Analisis Algoritma Stemming Nazief & Adriani dan Porter pada Dokumen Berbahasa Indonesia," *Jurnal Ilmiah SINUS*, vol. 15, no. 2, 2017, doi: 10.30646/sinus.v15i2.305.

[10] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012. doi: 10.1016/B978-0-12-381479-1.00001-0.

[11] G. Brassington, "Mean absolute error and root mean square error: which is the better metric for assessing model performance?," *Geophysical Research Abstracts*, vol. 19, pp. 2017–3574, 2017.

[12] T. F. de C. Marshall and J. L. Fleiss, "Statistical Methods for Rates and Proportions.," *The Statistician*, vol. 25, no. 1, p. 70, 1976, doi: 10.2307/2988144.