JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 9
No. 2
Agustus
2023

Sistem Penilaian Jawaban Singkat Otomatis pada Ujian Online Berbasis Komputer Menggunakan Algoritma Cosine Similarity

Dedy Kurniadi[#1], Rahmat Gernowo[#2], Bayu Surarso[#3], Adi Wibowo[*4], Budi Warsito[**5]

[#]*Doktor Sistem Informasi, Universitas Diponegoro*
Jl. Imam Bardjo SH No.5, Semarang
[1]dedykurniadi@students.undip.ac.id
[2]rahmatgernowo@lecturer.undip.ac.id
[3]bayus@lecturer.undip.ac.id

[*]*Informatika, Universitas Diponegoro*
Jl. Prof. Sudarto, Semarang
[4]bowo.adi@live.undip.ac.id

[**]*Statistika, Universitas Diponegoro*
Jl. Prof. Sudarto, Semarang
[5]budiwarsito@live.undip.ac.id

***Abstrak*— Penggunaan teknologi di bidang pendidikan sekarang ini sedang trending ke arah penilaian secara otomatis, namun penilaian secara otomatis ini memiliki permasalahan yaitu belum bisa mengkoreksi jawaban teks singkat secara otomatis, selain itu pada saat ini juga belum tersedia platform yang bisa mengkoreksi jawaban singkat secara otomatis, penilaian jawaban teks singkat ini membutuhkan waktu koreksi yang lama dan hasil penilaian yang tidak konsisten jika koreksi dilakukan oleh manusia, pada penelitian ini diusulkan sistem yang mampu mengkoreksi ujian peserta didik pada bagian jawaban singkat secara otomatis atau disebut dengan *Automated Short Answer* Grading (ASAG) dengan menggunakan metode *cosine similarity*, tahapan yang dilakukan adalah melakukan ekstraksi pada dua variabel inputan yaitu teks pada jawaban peserta didik dan teks pada kunci jawaban yang dilakukan dengan ekstraksi teks *casefolding*, *tokenizing*, *stopword removal*, setelah tahapan tersebut dilakukan kemudian dihitung nilai *similarity* antara kunci jawaban ujian dengan jawaban peserta didik apakah jawaban peserta didik sama dengan kunci jawaban atau tidak, dengan menggunakan skor yang dinilai otomatis menggunakan sistem, dihasilkan *similarity* antara jawaban peserta didik dengan kunci jawaban rata-rata sebesar 85,4%, untuk menguji korelasi koreksi jawaban peserta didik dengan sistem dan koreksi yang dilakukan oleh manusia maka dilakukan uji korelasi antara hasil penilian yang dilakukan oleh sistem dengan hasil penilaian yang dilakukan oleh manusia (instruktur) dengan menggunakan kendall's w value menghasilkan nilai w antara instruktur 1 dengan sistem sebesar 0,885 dan instruktur 2 dengan sistem sebesar 0,883 dengan nilai chi square sebesar 135,4 dan 133,8 dengan p sebesar 0,0001, hasil tersebut menunjukkan ASAG memiliki korelasi yang tinggi dan sistem ASAG ini bisa melakukan penilaian secara otomatis.**

***Kata kunci*— ASAG, Cosine Similarity, Education, Casefolding, Tokenizing, Stopword Removal**

I. PENDAHULUAN

Dunia Pendidikan telah mengalami perkembangan yang sangat signifikan dalam proses melakukan penilaian hasil ujian peserta didik, proses ini mengalami perubahan dari waktu ke waktu mulai dari penilaian dengan cara tradisional menggunakan kertas sampai dengan era digital saat ini yaitu menilai hasil ujian tersebut dengan menggunakan perangkat komputer [1] [2] penilaian terhadap peserta didik merupakan hal yang sangat penting, hal ini dilakukan untuk mengevaluasi pemahaman dan kemampuan peserta didik dalam penguasaan materi yang telah dipelajari dalam pendidikannya, untuk melakukan proses penilaian ujian dari peserta didik bisa dilakukan dengan berbagai metodel [3], salah satu diantaranya adalah dengan menggunakan metode jawaban singkat (*short answer*), metode jawaban singkat merupakan pilihan terbaik untuk bisa menggali lebih dalam kemampuan peserta didik dalam hal pemahaman dan pengetahuan dasar serta penguasaan materi yang telah dipelajarinya [4].

Penggunaan metode yang digunakan sekarang ini masih lebih banyak menggunakan metode pilihan ganda, namun metode ini tidak bisa menggali kemampuan peserta didik lebih dalam [5] karena dengan menggunakan metode pilihan ganda peserta didik sudah diberikan opsi jawaban yang mana dalam opsi jawaban tersebut sudah terdapat jawaban yang benar dengan kata lain peserta didik memiliki prosesntase minimal sebesar 25% jika jumlah jawaban sebanyak 4 (empat) item atau sebesar 20% jika

jumlah jawaban sebanyak 5 (lima) item, dengan jumlah prosentase tersebut peserta didik bisa mengisikan jawaban yang tidak akurat tetapi mempunyai prosentase keberhasilan menjawab dengan tepat sebesar 25% atau 20% sehingga untuk mengukur tingkat kemampuan peserta didik dalam penguasaan kemampuan [6] yang telah dipelajarinya menjadi bias dan tidak menentu serta tidak ada tingkat akurasi pasti untuk mendapatkan hasil evaluasi dari kemampuan penguasaan materi peserta didik tersebut, Penilaian hasil ujian menggunakan metode jawaban singkat (*short answer)* [7] belum diterapkan didalam penggunakaan computer based test (CBT) dan pengujiannya masih dilakukan secara manual yaitu dengan menggunakan kertas baik itu dalam kertas fisik ataupun dengan kertas digital (*softfile*) kemudian peserta didik menuliskan jawaban dari pertanyaan tersebut pada kertas yang tersedia atau menggunakan kolom jawaban pada komputer, dan cara mengkoreksi hasil ujian masih dibaca secara manual oleh guru / dosen (manusia) [8] untuk melakukan penilaian ujian dari peserta didik tersebut [9] [10] hal ini mungkin akan membutuhkan waktu koreksi yang sangat lama dan ada kemungkinan yang besar hasil evaluasi tidak konsisten antara hasil ujian peserta didik yang satu dengan yang lainnya jika ada banyak jumlah item set jawaban yang harus dikoreksi [11]. Dengan perkembangan teknologi saat ini, hal tersebut bisa diatasi dengan membuat perangkat *software* baru untuk membantu instruktur, guru dan dosen untuk mengelola ujian peserta didik dengan lebih baik dengan membuat smart computer based test agar hasil ujian peserta didik bisa dikoreksi secara otomatis oleh computer berdasarkan kunci jawaban dari instruktur, guru atau dosen, sistem *Automated Short Answer Grading* (ASAG) telah dikembangkan untuk bisa menilai dengan cara membandingkan jawaban dengan satu atau lebih dari kunci jawaban yang benar [12] yang sudah diset dalam dataset pertanyaan dan jawaban dalam sistem untuk bisa memberikan penilaian secara otomatis pada ujian peserta didik [4]. Perancangan system ASAG ini menggunakan Bahasa Indonesia dan disebut dengan Indonesian *Automated Short Answer Grading* (ID-ASAG), desain sistem ID-ASAG perangkat software komputer pertama yang dirancang khusus dalam Bahasa Indonesia yang sesuai dengan aturan dan tata Bahasa Indonesia. Dengan menggunakan penilaian otomatis maka hal tersebut bisa memberi beberapa keuntungan seperti waktu koreksi dan pemberian *feedback* hasil ujian yang lebih cepat, hasil koreksi terhadap ujian peserta didik bisa diukur dan konsisten [13].

II. DATASET

Dataset pada penelitian ini menggunakan data open yang disebut dengan Indonesian *Query Answering Dataset for Online Essay Test System*, didalam dataset tersebut terdapat 4 topik dataset pertanyaan dan jawaban ditunjukkan pada Tabel 1.

Dari topik 4 topik tersebut kemudian masing-masing ada sebanyak 52 jawaban dari siswa yang sudah tersimpan didalam system dan menjadi data mentah yang akan diolah menggunakan metode cosine similarity, 52 jawaban tersebut semuanya dilakukan ekstraksi text dengan tahapan casefolding, tokenizing dan stopwords dari keseluruhan jawaban dan pada data di penelitian ini ditunjukkan contoh data sebagai dataset yang diuji dari masing-masing topik yang ada, dataset pertanyaan dan jawaban yang ada pada topik komputer dijadikan sebagai contoh, dataset tersebut diambil kunci jawabnnya dan kemudian akan dilakukan similarity terhadap jawaban dari siswa yang sudah diekstrak dan dimasukkan kedalam database, data contoh kunci jawaban dutunjukkan pada Tabel 2.

TABEL I
DATASET TOPIK PERTANYAAN

No	Topik	Jumlah Pertanyaan
1	Olahraga	10
2	Lifestyle	15
3	Politik	15
4	Komputer	10

Dari topik 4 topik tersebut kemudian masing-masing ada sebanyak 52 jawaban dari siswa yang sudah tersimpan didalam system dan menjadi data mentah yang akan diolah menggunakan metode cosine similarity, 52 jawaban tersebut semuanya dilakukan ekstraksi text dengan tahapan casefolding, tokenizing dan stopwords dari keseluruhan jawaban dan pada data di penelitian ini ditunjukkan contoh data sebagai dataset yang diuji dari masing-masing topik yang ada, dataset pertanyaan dan jawaban yang ada pada topik komputer dijadikan sebagai contoh, dataset tersebut diambil kunci jawabnnya dan kemudian akan dilakukan similarity terhadap jawaban dari siswa yang sudah diekstrak dan dimasukkan kedalam database, data contoh kunci jawaban dutunjukkan pada Tabel 2.

TABEL II
DATA PERTANYAAN DAN KUNCI JAWABAN

No	Pertanyaan	Kunci Jawaban
1	Apa yang dimaksud dengan komputer? (Jawab dalam 1-3 kalimat)	Komputer adalah rangkaian mesin elektronik yang dapat bekerja sama. Sistem ini digunakan untuk memudahkan pekerjaan manusia. Komputer bekerja otomatis berdasarkan urutan instruksi atau program yang diberikan.
2	Apa yang dimaksud dengan volatile memory?	Volatile memory adalah memory yang datanya dapat ditulis dan dihapus, tetapi hilang saat kehilangan power (kondisi off atau mati lampu).
3	Apa kepanjangan dari LCD, CPU dan GPS?	LCD (Liquid Crystal Display), CPU (Central Processing Unit), GPS (Global Positioning System)
4	Bagaimana cara mengatasi sampah elektronik (e-waste)? (Sebutkan minimal 3)	'- Mendaur ulang sampah elektronik menjadi barang yang berguna dan memiliki nilai jual - Memisahkan barang elektronik sesuai

No	Pertanyaan	Kunci Jawaban
		komposisi bahannya - tidak membuang sembarangan sehingga karena dapat mencemari lingkungan - Dikembalikan kepada produsen elektronik

III. METODE PENELITIAN

Metode yang digunakan pada penelitian ini adalah menggunakan algoritma *cosine similarity* untuk menghitung tingkat kesamaan antara jawaban siswa dengan kunci jawaban pada dataset yang sudah dimasukkan kedalam database kemudian dilakukan query sistem jawaban dengan kunci jawaban yang ada pada dataset kemudian dilakukan *casefolding*, *tokenizing* dan *stopword* untuk menyamakan variable inputan jawaban siswa dengan kunci jawaban yang ada di dalam database, selain menggunakan *cosine similarity* pada penelitian ini juga menggunakan *Euclidean distance* sebagai validasi lanjutan terhadap hasil jawaban siswa.

A. *Cosine Similarity*

Cosine similarity merupakan algoritma yang bisa digunakan untuk mengukur tingkat kesamaan pada teks dengan cara membandingkan teks inputan dengan teks pembanding yang sudah ada didalam dataset, untuk mencari nilai *similarity* dengan menggunakan algoritma *cosine similarity* perlu beberapa tahapan, pertama dengan melakukan penilaian skalar antara query dengan dokumen pembanding kemudian dijumlahkan, setelah selesai penjumlahan kemudian melakukan perkalian panjang dokumen [14] dengan panjang query yang sudah dikuadratkan, setelah itu di hitung akar pangkat dua. Selanjutnya hasil perkalian skalar tersebut di bagi dengan hasil perkalian panjang dokumen dan query [15], [16] persamaan untuk melakukan perhitungan cosine similarity ditunjukkan pada persamaan (1).

$$cosSim(_{d_j,q_k}) = \frac{\sum_{i=1}^{n}(td_{ij}\ x\ tq_{ik})}{\sqrt{\sum_{i=1}^{n} td_{ij^2}\ x\ \sum_{i=0}^{n} tq_{ik^2}}} \quad (1)$$

dimana :

$cosSim(_{d_j,q_k})$= tingkat kesamaan dokumen
$tdij$ = term ke-i di vector dokumen ke-j
$tqik$ = term ke-i di vector dokumen ke-k
n = jumlah term yang unik dalam dataset

Pada tahapan untuk menemukan tingkat kesamaan nilai query dengan dokumen pembanding dilakukan term-context matrix. Case Folding proses dalam text preprocessing yang dilakukan untuk menyeragamkan karakter yang ada pada data, proses case folding adalah merubah semua karakter huruf menjadi huruf kecil. Tokenizing [17] proses pemotongan string masukan berdasarkan kata-kata yang menyusunnya menjadi kalimat dipecah menjadi kata per kata, pemotongan dilakukan pada white space. Stopword proses menghilangkan kata penghubung. Stemming proses mengambalikan kata kedalam bentuk aslinya [18].

B. *Euclidean Distance*

Euclidean distance adalah metrik jarak ukur antara dua vektor dengan menghitung akar kuadrat dari jumlah selisih kuadrat antara keduanya. Metode ini mengkomparasi kalimat menggunakan jarak untuk menemukan dua data yang paling sama (similar), pada Euclidean distance yang dianggap mirip adalah semakin kecil jarak dari dua data tersebut maka kedua data tersebut dianggap data yang mirip atau sama [10]. Proses perhitungan Euclidean distance ditunjukkan pada persamaan (2).

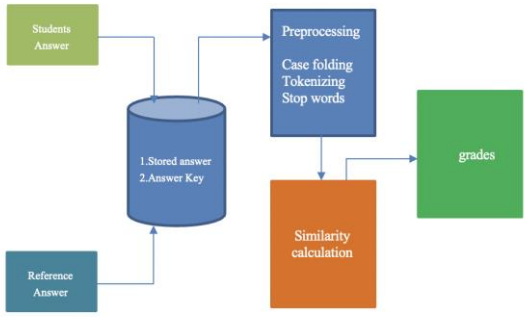
$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2)$$

dimana :

d = jarak
x_1 = koordinat latitude 1
x_2 = koordinat latitude 2
y_1 = koordinat longitude 1
y_1 = koordinat longitude 2

C. *Model Perancangan ASAG*

ID-ASAG yang dikembangkan ini merupakan system yang digunakan untuk bisa melakukan otomatisasi penilaian terhadap jawaban peserta didik, Ketika peserta didik menjawab pertanyaan dari set question yang ada dalam system maka akan dicocokkan nilai similaritynya dengan reference answer (kunci jawaban) yang ada didalam dataset, desain dan mekanisme secara detail dijelaskan pada Gambar 1.



Gambar. 1 Perancangan Skema mekanisme alur sistem ID-ASAG

Pada skema mekanisme system ID-ASAG yang ditunjukkan pada Gambar 1, menunjukkan terdapat dua inputan yaitu inputan students answer dan reference answer yang dijadikan sebagai answer key, kemudian kedua data tersebut dimasukkan kedalam database dan diberi label

data sebagai stored answer dan answer key, selanjutnya data tersebut dilakukan tahapan preprocessing yang didalamnya ada tahapan casefolding, tokenizing dan stopword untuk mendapatkan nilai vector dari kedua data tersebut, nilai vector kemudian dicari similaritynya menggunakan Euclidean distance dan cosine similarity dari hasil tersebut kemudian ditetapkan nilai atau skor dari jawaban yang dimasukkan tersebut [13], [19].

IV. HASIL PENELITIAN

Penelitian ID-ASAG ini mengkombinasikan dua metode yaitu euclidean distance dan cosine similarity untuk membandingkan antara jawaban siswa yang sudah disubmit dengan kunci jawaban dari dataset yang sudah disimpan didalam database dan dilakukan proses query untuk menemukan nilai kesamaan antara teks inputan dengan teks pembanding dan akan menghasilkan nilai similarity dengan jawaban dari kunci jawaban pertanyaan yang ada dalam dataset, dalam penelitian dilakukan pada 10 dataset pertanyaan dan pada 52 jawaban dari siswa yang mensubmit jawaban hasil pembandingkan teks inputan dan kunci jawaban menggunakan euclidean distance dan cosine similarity dari dataset ditunjukkan pada Tabel 3.

TABEL III
DATA ENSWERKEY DISTANCE SIMILARITY

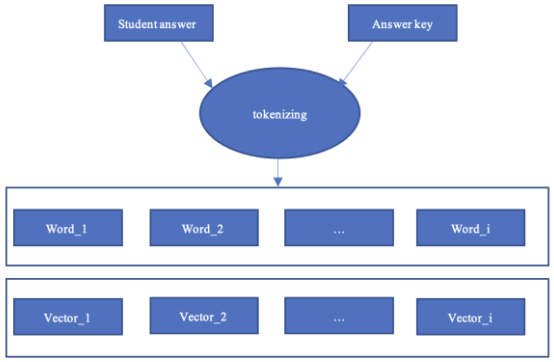
Question set	Answer Key Count	Euclidean	Cosine
1	208	4,795	0,582
2	276	2,236	0,856
3	221	3,241	0,743
4	286	4,154	0,603
…	…	…	…
10	236	2,165	0,897
Average	-	4,326	0,854

Data pertanyaan dan jawaban sudah disimpan didalam dataset, instruktur membuat *answer key* sebagai referensi jawaban yang betul terlebih dahulu kedalam dataset, kemudian dataset dilakukan query untuk menemukan jawaban dari student yang hasilnya dikomparasi dengan *answer key.* Pada penelitian ini terdapat dua kunci jawaban dari instruktur dan dimasukkan kedalam sistem untuk menguji validitas dan reliabilitas dari skor test yang sudah dilakukan oleh peserta didik, jumlah pertanyaan dalam dataset ini adalah sebanyak 10 pertanyaan dengan topik pertanyaan teknologi, kunci jawaban dan jawaban dari peserta didik harus sama-sama dimasukkan kedalam dataset pertanyaan jawaban untuk dilakukan prepocessing menggunakan casefolding seperti pada Tabel 4.

TABEL IV
DATA CASEFOLDING

Index	sentence	casefolding
Kunci jawaban	LCD (Liquid Crystal Display), CPU (Central Processing Unit), GPS (Global Positioning System)	lcd (liquid crystal display), cpu (central processing unit), gps (global positioning system)
Jawaban Peserta Didik	Liquid Crystal Display, Central Processing Unit, Global Positioning System.	liquid crystal display, central processing unit, global positioning system

Hasil dari casefolding merubah semua karakter menjadi huruf kecil seperti yang ditunjukkan pada table 3, Langkah selanjutnya adalah dengan melakukan tokenizing yaitu memecah kalimat menjadi kata per kata selain merubah menjadi kata per kata dalam proses tokenizing juga melakukan punctuation seperti simbol dan tanda baca yang tidak penting dihilangkan dan juga menghilangkan whitespace, cara melakukan tokenizing ditunjukkan pada gambar 2.



Gambar. 2 Stopword Removal dan vector tokenizing

```
# Tokenizing
from nltk.tokenize import RegexpTokenizer
tokenizer      = RegexpTokenizer(r'\w+')
hasil_token    = tokenizer.tokenize(hasil_cf)
print (hasil_token)

# OUTPUT:
# ['lcd', 'liquid', 'crystal', 'dis', 'play', 'cpu', 'central',
'pro', 'cessing', 'unit', 'gps', 'global', 'positioning', 'system']
```

Gambar. 3 Coding tokenizing

TABEL V
DATA TOKENIZING

Index	Tokenizing
Student answer	'liquid', 'crystal', 'display', 'central', 'processing', 'unit', 'global', 'positioning', 'system'
Answer key	'lcd', 'liquid', 'crystal', 'dis', 'play', 'cpu', 'central', 'pro', 'cessing', 'unit', 'gps', 'global', 'positioning', 'system'

Tokenizing atau disebut juga tahap *Lexical Analysis* adalah proses pemotongan teks menjadi bagian-bagian yang lebih kecil, yang disebut token. Pada proses ini juga dilakukan penghilangan angka, tanda baca dan karakter lain yang dianggap tidak memiliki pengaruh terhadap pemrosesan teks, berikut contoh dari proses tokenizing dalam penelitian ini ditunjukkan pada Gambar 3. Hasil dari tokenizing pada penelitian ini ditunjukkan pada Tabel 5, tabel data tokenizing.

Seperti yang ditunjukkan pada Tabel 5, kalimat dipecah menjadi kata per kata dan semua karakter yang tidak penting dihilangkan sehingga didapatkan data kata yang penting yang akan digunakan untuk mengukur tingkat similaritas dari student answer dan answer key. Data yang sudah dilakukan tokenizing kemudian dilakukan stemming untuk mencari kata dasar dari data yang sudah di tokenizing. Data stemming ditunjukkan pada Tabel 6.

TABEL VI
DATA HASIL STEMMING

Index	stemming
Student answer	'liquid', 'crystal', 'display', 'central', 'processing', 'unit', 'global', 'positioning', 'system'
Answer key	'lcd', 'liquid', 'crystal', 'display', 'cpu', 'central', 'process', 'unit', 'gps', 'global', 'position', 'system'

Kata-kata yang sudah dilakukan stemming berubah menjadi kata dasar dari kata tersebut contohnya adalah processing Ketika dilakukan stemming berubah menjadi process yang merupakan kata dasar dari processing, selanjutnya kita melakukan term frequent yaitu proses untuk mencari frekuensi kemunculan kata antara student answer dengan answer key persamaan yang digunakan untuk mencari term frequent ditunjukkan pada persamaan (3).

$$w(d,t) = TF(d,t) \quad (3)$$

Dimana
w(d,t) = weight
tf(d,t) = frekuensi kemunculan term pada dokumen

Hasil dari proses term frequent dari dataset tersebut diproses dengan memecah kata per kata untuk dibentuk kedalam masing-masing set kata dan dihitung frekuensi kemunculan kata per kata tersebut dengan answer key dan query masukkan dari student answer, hasil dari term frequent ditunjukkan pada Gambar 4.

	central	cessing	cpu	crystal	dis	display	global	gps	lcd	liquid	play	positioning	pro	processing	system	unit
d1	1		1	1		1	1	0	1	1	1	1		1	1	0	1	1
q	1		0	0		1	0	1	1	0	0	1	0	1	0	1	1	1

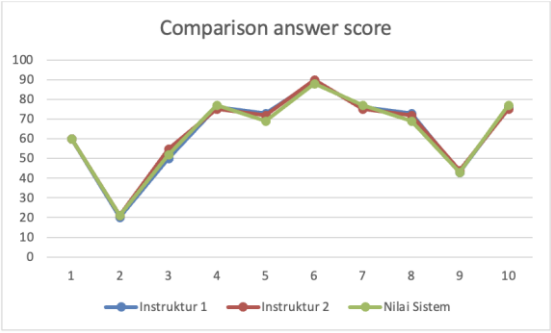
Gambar. 4 Hasil term frequent dari jawaban dan kunci jawaban

Dari proses tersebut kemudian dibentuk konten *similarity* dari *student answer* dibandingkan dengan *answer key*, pada penelitian ini data yang digunakan adalah set pertanyaan dengan topik komputer yang sudah disimpan didalam database, kemudian proses ekstraksi pada jawaban siswa dengan kunci jawaban yang ada pada topik komputer diuji tingkat kesamaannya dengan menggunakan *answer key system* yang nantinya perbandingan jawaban siswa dan kunci jawaban akan divalidasi, untuk validasi penelitian ini menggunakan dua orang instruktur (guru / dosen) untuk mengkoreksi jawaban dari awaban siswa tersebut untuk mendapatkan nilai korelasi antara hasil koreksi sistem dengan hasil koreksi dari instruktur, data yang digunakan adalah 10 set pertanyaan dengan jumlah peserta yang menjawab set pertanyaan tersebut sejumlah 10 peserta didik, hasil dari koreksi system dan koreksi instruktur ditunjukkan pada Tabel 7.

TABEL VII
DATA HASIL STEMMING

Student	System scoring	Instructor 1 scoring	Instrukctor 2 scoring
1	60	60	60
2	20	21	21
3	50	55	52
…	…	…	…
10	76	75	77

Penelitian ini mencoba mengkomaparasi hasil koreksi antara system dengan instruktur untuk memastikan bahwa hasil dari skoring system ID-ASAG tidak jauh berbeda dengan hasil koreksi manula yang dilakukan oleh instruktur (guru / dosen), arithmetic mean score pada komparasi tersebut ditunjukkan pada Gambar 5.



Gambar. 5 Komparasi hasil koreksi antara system dengan guru / dosen

Kemudian untuk menguji korelasi dan keselarasan antara penilaian system dengan penilaian yang dilakukan oleh instruktur pada penelitian ini digunakan metode Kendall's W value, merupakan metode uji nonparametrik yang digunakan untuk menguji korelasi dan keselarasan terhadap penilaian yang diberikan oleh sekelompok subjek terhadap atribut-atribut yang dianggap penting, untuk melakukan uji kendall's W value digunakan persamaan yang ditunjukkan pada persamaan 4.

$$W = \frac{12 \sum_{i=0}^{m} (R_i - \bar{R})^2}{b^2 (m^3 - m)} \quad (4)$$

dimana :
W = nilai statistic kendall's w
R_i = jumlah rangking pada atribut ke-i= 1,2, ..., m
$\bar{R}$ = ranking rata-rata
m = jumlah atribut diteliti
b. = jumlah responden / elemen dalam sample

Pada penelitian ini digunakan kendall's w value dengan menggunakan 3 rater yaitu system, instructor 1 dan instruktur 2 untuk menemukan nilai korelasi antara ketiganya, sebagai validasi dan korelasi antara hasil koreksi yang dilakukan oleh manusia (instruktur : guru dan dosen) dengan hasil koreksi yang dilakukan oleh sistem, hasil dari komparasi tersebut ditunjukkan pada Tabel 8.

TABEL VIII
HASIL KORELASI KENDAL W VALUE

Raters Comparison	**W**	**Chi Square**	**df**	***p***
Guru / dosen 1, system ID-ASAG	0,885	135,4	8	0,0001
Guru / dosen 2, system ID-ASAG	0,883	133,8	8	0,0001
Guru / dosen 1, guru / dosen 2, system ID-ASAG	0,884	134,7	8	0,0001

Kemudian dicari korelasinya menggunakan person corellation antar rater pada dataset untuk menemukan tingkat korelasi yang ada pada ID-ASAG, ditunjukkan pada tabel 9.

TABEL IX
NILAI KORELASI ANTARA GURU/DOSEN DENGSN SISTEM

#		Guru / dosen 1 score	Guru / dosen 2 score	System ID_ASAG Score
Guru / dosen 1 score	Pearson corellation sig.(2-tailed)	1	0,997	0,996
Guru / dosen 2 score	Pearson corellation sig.(2-tailed)	0,997	1	0,993
System ID_ASAG Score	Pearson corellation sig.(2-tailed)	0,996	0,993	1

Dari hasil korelasi pada tabel 9 diketahui bahwa terdapat korelasi antara penilaian manual yang dilakukan oleh guru / dosen dengan penilaian yang dilakukan oleh system, hasil menunjukkan tingkat korelasi yang tinggi artinya penilaian otomatis yang dilakukan oleh system bisa memberikan skor yang hampir sama dengan tingkat akurasi yang tinggi dengan hasil penilaian yang dilakukan oleh instruktur.

V. KESIMPULAN

Penelitian ini menghadirkan model baru dalam hal penilaian short answer secara otomatis, penelitian ini juga melakukan komparasi penilaian antara system dan manusia dengan metode kendall's W Value dengan hasil nilai korelasi antara instruktur 1 dengan system sebesar 0,885 dan nilai korelasi instruktur 2 dengan system sebesar 0,883 dengan nilai chi square 135,4 dan 133,8 dengan df sebanyak 8 menghasilkan nilai p sebesar 0,0001, kemudian penelitian ini juga melakukan uji pearson corellation dengan 2 tailed dengan hasil antara instruktur 1 dengan system sebesar 0,996 dan instruktur 2 dengan system sebesar 0,993 dengan hasil tersebut penelitian ID-ASAG menggunakan gabungan metode Euclidean distance dan cosine similarity bisa melakukan penilaian secara otomatis terhadap student answer dan memberikan skor atau nilai yang akurat.

REFERENSI

[1] A. Magooda, M. A. Zahran, M. Rashwan, H. Raafat, and M. B. Fayek, "Vector Based Techniques for Short Answer Grading." [Online]. Available: http://nlp.stanford.edu/software/corenlp.shtml

[2] Y. Huang and J. Wilson, "Using automated feedback to develop writing proficiency," *Comput Compos*, vol. 62, Dec. 2021, doi: 10.1016/j.compcom.2021.102675.

[3] R. Correnti, L. C. Matsumura, E. L. Wang, D. Litman, and H. Zhang, "Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment," *Computers and Education Open*, vol. 3, p. 100084, Dec. 2022, doi: 10.1016/j.caeo.2022.100084.

[4] K. H. Sung, E. H. Noh, and K. H. Chon, "Multivariate generalizability analysis of automated scoring for short answer items of social studies in large-scale assessment," *Asia Pacific Education Review*, vol. 18, no. 3, pp. 425–437, Sep. 2017, doi: 10.1007/s12564-017-9498-1.

[5] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," *Research Methods in Applied Linguistics*, vol. 2, no. 2, p. 100050, Aug. 2023, doi: 10.1016/j.rmal.2023.100050.

[6] J. Jacobs, K. Scornavacco, C. Harty, A. Suresh, V. Lai, and T. Sumner, "Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change," *Teach Teach Educ*, vol. 112, Apr. 2022, doi: 10.1016/j.tate.2022.103631.

[7] Y. Vo, H. Rickels, C. Welch, and S. Dunbar, "Human scoring versus automated scoring for english learners in a statewide evidence-based writing assessment," *Assessing Writing*, vol. 56, p. 100719, Apr. 2023, doi: 10.1016/j.asw.2023.100719.

[8] M. Beseiso, O. A. Alzubi, and H. Rashaideh, "A novel automated essay scoring approach for reliable higher educational assessments," *J Comput High Educ*, vol. 33, no. 3, pp. 727–746, Dec. 2021, doi: 10.1007/s12528-021-09283-1.

[9] Y. Zhang, C. Lin, and M. Chi, "Going deeper: Automatic short-answer grading by combining student and question models," *User Model User-adapt Interact*, vol. 30, no. 1, pp. 51–80, Mar. 2020, doi: 10.1007/s11257-019-09251-6.

[10] B. Das, M. Majumder, A. A. Sekh, and S. Phadikar, "Automatic question generation and answer assessment for subjective examination," *Cogn Syst Res*, vol. 72, pp. 14–22, Mar. 2022, doi: 10.1016/j.cogsys.2021.11.002.

[11] O. Nael, Y. ELmanyalawy, and N. Sharaf, "AraScore: A deep learning-based system for Arabic short answer scoring," *Array*, vol. 13, Mar. 2022, doi: 10.1016/j.array.2021.100109.

[12] A. Elnaka, O. Nael, H. Afifi, and N. Sharaf, "AraScore: Investigating Response-Based Arabic Short Answer Scoring," in *Procedia CIRP*, Elsevier B.V., 2021, pp. 282–291. doi: 10.1016/j.procs.2021.05.091.

[13] G. Liang, B. W. On, D. Jeong, H. C. Kim, and G. S. Choi, "Automated essay scoring: A siamese bidirectional LSTM neural network architecture," *Symmetry (Basel)*, vol. 10, no. 12, Dec. 2018, doi: 10.3390/sym10120682.

[14] R. A. Sottilare, R. S. Baker, A. C. Graesser, and J. C. Lester, "Special Issue on the Generalized Intelligent Framework for Tutoring (GIFT): Creating a Stable and Flexible Platform for

Innovations in AIED Research," *International Journal of Artificial Intelligence in Education*, vol. 28, no. 2. Springer New York LLC, pp. 139–151, Jun. 01, 2018. doi: 10.1007/s40593-017-0149-9.

[15] N. Birla, M. Kumar Jain, and A. Panwar, "Automated assessment of subjective assignments: A hybrid approach," *Expert Syst Appl*, vol. 203, Oct. 2022, doi: 10.1016/j.eswa.2022.117315.

[16] L. Zhang, Y. Huang, X. Yang, S. Yu, and F. Zhuang, "An automatic short-answer grading model for semi-open-ended questions," *Interactive Learning Environments*, vol. 30, no. 1, pp. 177–190, 2019, doi: 10.1080/10494820.2019.1648300.

[17] M. Beseiso and S. Alzahrani, "An Empirical Analysis of BERT Embedding for Automated Essay Scoring," 2020. [Online]. Available: www.ijacsa.thesai.org

[18] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artif Intell Rev*, vol. 55, no. 3, pp. 2495–2527, Mar. 2022, doi: 10.1007/s10462-021-10068-2.

[19] O. L. Liu, J. A. Rios, M. Heilman, L. Gerard, and M. C. Linn, "Validation of automated scoring of science assessments," *J Res Sci Teach*, vol. 53, no. 2, pp. 215–233, Feb. 2016, doi: 10.1002/tea.21299.