PERBANDINGAN METODE *COSINE SIMILARITY* DAN *JACCARD SIMILARITY* UNTUK PENILAIAN OTOMATIS JAWABAN PENDEK

Uswatun Hasanah[*1], Dwi Ayu Mutiara[2]
[1,2]Program Studi Teknologi Informasi, Universitas Amikom Purwokerto, Purwokerto
e-mail: [*1]uswatun_hasanah@amikompurwokerto.ac.id, [2]dwiayumutiara270@gmail.com

Abstrak

Sistem penilaian otomatis pada jawaban pendek dapat dipertimbangkan sebagai alternatif dalam proses penilaian ujian siswa. Berbeda dengan model pilihan ganda, model penilaian jawaban pendek lebih sulit untuk dihitung dengan metode komputasi, karena penilaian jawaban pendek membutuhkan teknik-teknik pengolahan bahasa alami. Beberapa metode komputasi telah digunakan dan dikembangkan oleh peneliti sebelumnya. Salah satu teknik yang paling dasar untuk digunakan adalah metode penilaian berbasis leksikal yang menilai jawaban berdasarkan kemiripan susunan karakternya. Penelitian ini menggunakan metode Cosine Similarity dan Jaccard Similarity yang dapat digunakan untuk mengukur kemiripan jawaban siswa dengan jawaban guru berdasarkan kata-kata penyusunnya. Teknik pra-pemrosesan teks juga digunakan untuk membandingkan hasil pada masing-masing metode. Data yang digunakan pada penelitian ini merupakan soal dan jawaban berbahasa Indonesia. Hasil menunjukkan bahwa metode Cosine Similarity yang dilengkapi dengan pra-pemrosesan teks mampu meraih korelasi tertinggi (0.62) pada pengukuran korelasi Pearson.

Kata kunci—penilaian otomatis, *cosine similarity*, *jaccard similarity*, jawaban pendek, pengolahan bahasa alami

Abstract

An automatic grading system on short answers can be considered as an alternative in the student examination grading process. Unlike the multiple choice model, the short answer assessment model is more difficult to calculate with computational methods, because the assessment of short answers requires natural language processing techniques. Several computational methods have been used and developed by previous researchers. One of the most basic techniques to use is a lexical-based assessment method that evaluates answers based on the similarity of character arrangements. This study uses the Cosine Similarity and Jaccard Similarity methods that can be used to measure the similarity of students' answers to the teacher's answers based on their constituent words. Text pre-processing techniques are also used to compare results for each method. The data used in this study are Indonesian questions and answers. The results show that the Cosine Similarity method which is equipped with text pre-processing is able to achieve the highest correlation (0.62) on the Pearson's Correlation measurement.

Keywords—automatic grading, cosine similarity, jaccard similarity, short answer, natural language processing

1. PENDAHULUAN

Indikator keberhasilan suatu pembelajaran dapat direpresentasikan dengan berbagai evaluasi pembelajaran dan model penilaian. Model penilaian yang paling banyak digunakan oleh guru/dosen adalah pilihan ganda dan esai. Model pilihan ganda terdiri dari serangkaian pertanyaan yang mana pada setiap pertanyaan terdapat satu jawaban benar dan beberapa jawaban yang salah. Dalam penggunaan e-learning, komputasi pada model penilaian pilihan ganda lebih mudah dilakukan daripada pada model esai [1]. Hal ini dikarenakan penilaian jawaban esai membutuhkan teknik pengolahan bahasa alami terlebih dahulu. Walaupun lebih mudah dihitung dengan menggunakan metode komputasi, model penilaian pilihan ganda memiliki realibilitas yang rendah, karena siswa dapat menebak jawaban dari alternatif jawaban yang disediakan [2]. Berbeda dengan pilihan ganda, esai membutuhkan jawaban dalam format bahasa alami yang mengharuskan siswa untuk memaparkan pemahamannya pada materi yang diujikan. Peneliti beranggapan bahwa esai dapat menjadi alat yang efektif untuk menilai pencapaian hasil pembelajaran, serta dapat dijadikan sebagai alat untuk mengamati tingkat kemahiran berpikir siswa, seperti sintesa dan analisa [3].

Secara umum ada banyak manfaat yang dapat diperoleh melalui penilaian otomatis jawaban pendek, diantaranya adalah membuat waktu koreksi lebih cepat, lebih objektif, dan meminimalisir faktor kesalahan manusia. Namun, masalahnya ada pada kualitas sistem penilaian otomatis. Bukan hal yang mudah untuk membuat mesin melakukan penilaian seperti yang dilakukan oleh penilai manusia (*human rater*). Berbagai teknik dan metode telah dikembangkan oleh peneliti-peneliti sebelumnya agar skor yang dihasilkan oleh mesin mampu menyamai skor yang dihasilkan oleh *human rater*.

Penelitian ini bertujuan untuk membandingkan nilai kemiripan antara dua teks, yaitu jawaban mahasiswa sebagai jawaban yang akan dinilai dan jawaban dosen sebagai kunci jawaban (*key answer*). Metode yang akan digunakan adalah *Cosine Similarity* dan *Jaccard Similarity* yang mampu menilai kemiripan di antara dua teks secara leksikal.

2. TINJAUAN PUSTAKA

2.1 Sistem Penilaian untuk Esai

Salah satu cabang ilmu dari data mining adalah text mining. Perbedaan yang cukup mendasar dari data mining dan *text mining* adalah pada bentuk datanya. Data mining umumnya memiliki data terstruktur, sementara *text mining* berisi data yang tidak terstruktur. Pada text mining, teknik-teknik natural language processing digunakan untuk memaksimalkan information retrieval (IR). Beberapa penelitian seperti peringkasan dokumen, pendeteksian plagiarisme, sentiment analysis, dan penilaian otomatis pada jawaban esai menggunakan teknik pengolahan bahasa alami.

Penelitian tentang penilaian bahasa alami menggunakan metode komputasi diawali oleh Page pada tahun 1966 [4]. Sejak saat itu, penelitian dalam bidang ini terus berkembang menjadi area yang lebih luas. Teknik penilaian otomatis pada bahasa alami dibagi menjadi dua jenis berdasarkan tipe pertanyaannya, yaitu jawaban pendek dan esai [1]. Jawaban pendek merupakan kalimat yang terdiri dari satu frasa atau memiliki panjang tiga sampai empat kalimat [5]. Jawaban pendek panjangnya tidak lebih dari 100 kata [6][7]. Sedangkan esai dapat terdiri dari dua atau lebih paragraf sampai dengan beberapa halaman [1]. Jawaban pendek lebih berfokus pada konten sementara fokus esai panjang terletak pada style penulisan [8].

Beberapa penelitian mengenai sistem penilaian otomatis untuk jawaban pendek telah banyak dilakukan oleh peneliti sebelumnya. Burrows [1] membagi bidang penelitian ini menjadi lima era, yaitu *Concept Mapping, Information Extraction, Corpus-Based Methods, Machine Learning* dan *Evaluation Era*. Sementara Roy [9] membagi teknik penilaian otomatis jawaban pendek menjadi lima teknik, yaitu *Natural Language Processing* (NLP), *Information Extraction and Pattern Matching, Machine Learning, Document Similarity*, dan *Clustering*. Keunggulan dan

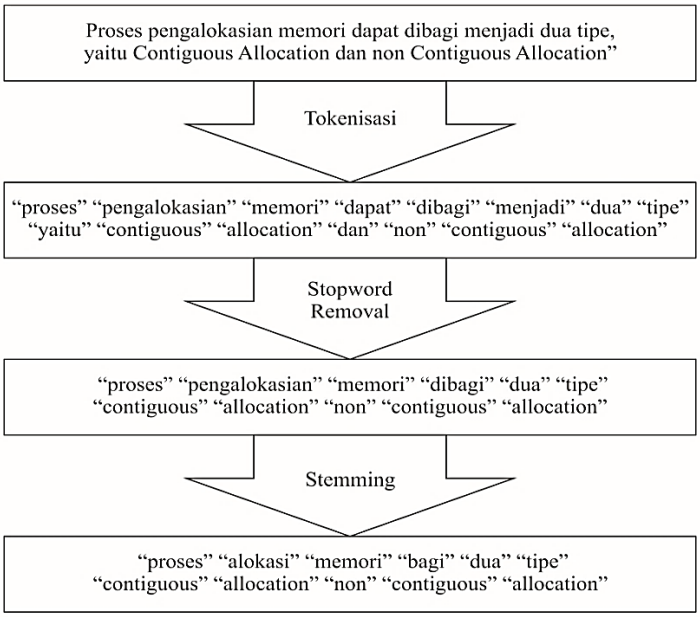
kelemahan dari masing-masing teknik tersebut bergantung pada jenis materi pembelajaran, tipe soal, serta interval nilai yang ditentukan oleh guru/dosen.

2.2 Pra-pemrosesan Teks

Pada sistem penilaian esai otomatis, jawaban kunci dan jawaban siswa perlu melalui proses pra-pemrosesan teks. Pra-pemrosesan teks yang digunakan untuk bahasa Indonesia dapat terdiri dari:

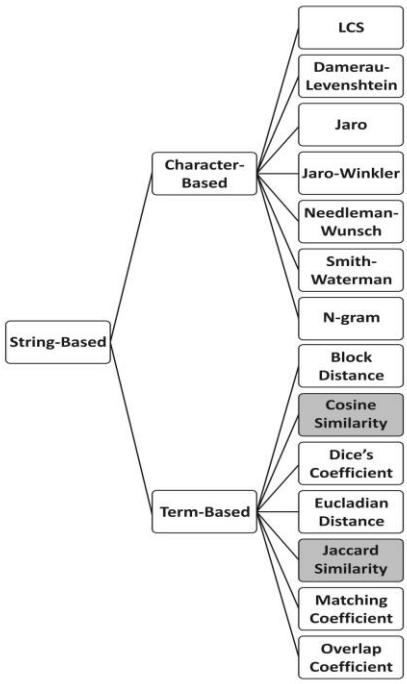
1) *Tokenization*: Proses tokenisasi berfungsi untuk mengubah kalimat menjadi token-token atau string. Pada proses ini, seluruh tanda baca dihilangkan dan semua huruf besar diubah menjadi huruf kecil.
2) *Stopword Removal*: Stopword merupakan kata-kata yang tidak terlalu penting dalam kalimat namun memiliki frekuensi kemunculan yang tinggi. Contoh *stopword* dalam bahasa Indonesia adalah "yang", "adalah", "oleh", dan "akan".
3) *Stemming*: merupakan metode untuk meningkatkan performa *information retrieval* dengan cara mentransformasi kata-kata di dalam suatu dokumen teks ke dalam bentuk kata dasarnya [10]. Proses *stemming* mengeliminasi imbuhan pada kata-kata sehingga dapat merepresentasikan semantik kata yang sama walaupun morfologinya berbeda. Contoh *library stemmer* dalam bahasa Indonesia adalah *Stemmer* Sastrawi yang mengadopsi algoritma pada penelitian Nazief & Adriani [11], Asian [12], Arifin [13], dan Tahitoe [14].
4) *Casefolding*: merupakan teknik untuk mengubah huruf/karakter menjadi ukuran yang seragam. Dalam kasus ini, baik jawaban dosen maupun jawaban mahasiswa diproses ke dalam bentuk *lower case* atau huruf kecil.
5) *Punctuation Removal*: teknik *punctuation removal* atau penghapusan tanda baca digunakan untuk membuat seluruh kalimat bebas dari simbol-simbol dan tanda baca. Teknik ini mengurangi kesalahan pengukuran kemiripan kalimat akibat tanda baca, yang mana dalam *dataset* matakuliah riset operasi tanda baca tidak diperhitungkan sebagai poin penilaian.

Secara umum, contoh pra-pemrosesan teks di atas dapat direpresentasikan pada Gambar 1 berikut ini:



Gambar 1 Contoh pra-pemrosesan teks

2.3 Cosine Similarity dan Jaccard Similarity



Gambar 2 Teknik pengukuran kemiripan kalimat berbasis string

Mengukur kesamaan/kemiripan antara kata-kata, kalimat, paragraf dan dokumen merupakan komponen penting dalam berbagai pekerjaan seperti pencarian informasi, klastering dokumen, disambiguasi makna, penilaian esai otomatis, penilaian jawaban pendek, mesin penerjemah dan peringkasan teks [15]. Kata-kata dapat disebut sama/mirip melalui dua cara, yaitu secara leksikal dan secara semantik. Kata-kata dikatakan mirip secara leksikal apabila memiliki urutan karakter yang mirip. Sedangkan kata-kata dikatakan mirip secara semantik apabila memiliki arti yang sama, berlawanan satu sama lain, digunakan dengan cara yang sama, digunakan pada konteks yang sama dan kata tersebut merupakan salah satu jenis dari kata lainnya. Dalam penelitiannya, Gomaa [15] mengklasifikasikan kemiripan kata-kata secara leksikal sebagai *String-Based Algorithm* yang ditunjukkan oleh Gambar 2.

Cosine similarity (CS) dan *Jaccard Similarity* (JS) merupakan metode berbasis *term* yang dapat digunakan untuk mengukur kemiripan antara dua dokumen atau teks. Pada penelitian ini, *cosine similarity* dan jaccard similarity digunakan untuk membandingkan nilai kemiripan antara jawaban mahasiswa dan jawaban dosen (*key answer*). Kedua metode ini menghitung jumlah term yang sama pada masing-masing kalimat dan membandingkannya dengan jumlah seluruh term pada kedua kalimat.

Nilai *Cosine Similarity* (CS) dari jawaban dosen (X) dan jawaban mahasiswa (Y) dapat ditulis dalam persamaan (1) berikut:

$$CS(X,Y) = \frac{|X \cap Y|}{|X|^{\frac{1}{2}}.|Y|^{\frac{1}{2}}} \quad (1)$$

Sedangkan nilai *Jaccard Similarity* (JS) dari jawaban dosen (X) dan jawaban mahasiswa (Y) dapat ditulis dalam persamaan (2) berikut:

$$JS(X,Y) = \frac{|X \cap Y|}{|X|+|Y|-|X \cap Y|} \quad (2)$$

Misalkan terdapat suatu dokumen jawaban dosen (X) dan jawaban mahasiswa (Y) sebagai berikut:

Pertanyaan : Apa yang dimaksud dengan MFT?
Jawaban Dosen (*X*) : Memori dibagi menjadi beberapa blok dengan ukuran tertentu yang seragam.
Jawaban Siswa (*Y*) : Membagi memori menjadi beberapa blok dengan ukuran tertentu yang seragam.

Dengan demikian dapat diketahui bahwa irisan pada $X \cap Y$ = "memori, menjadi, beberapa, blok, dengan, ukuran, tertentu, yang, seragam". Sehingga didapatkan nilai $X \cap Y = 8$, sedangkan nilai $X = 10$, dan nilai $Y = 10$. Nilai *X* dan *Y* adalah jumlah *term* pada kalimat, bukan jumlah keseluruhan kata. Misal ada dua kata "memori" pada masing-masing kalimat, maka term "memori" dihitung 1. Dengan menggunakan persamaan (1) dan (2) maka nilai *Cosine Similarity* kedua kalimat tersebut adalah 9/(sqrt(10)*sqrt(10))=0,9 (90%) dan untuk nilai *Jaccard Similarity* adalah 9/(10+10-9)=0,82 (82%).

3. METODE PENELITIAN

3.1 Bahan Penelitian

Bahan yang digunakan dalam penelitian ini adalah dokumen soal dan jawaban pendek dari ujian tengah semester matakuliah sistem operasi. *Dataset* terdiri 1 pertanyaan yang mana masing-masing pertanyaan dijawab oleh 74 mahasiswa. Berikut adalah pertanyaan yang diajukan oleh dosen:

Pertanyaan: Apa yang dimaksud dengan MFT?

Jawaban:

- Memori dibagi menjadi beberapa blok dengan ukuran tertentu yang seragam.
- Setiap partisi berisi tepat 1 proses.
- Digunakan oleh IBM OS/360 yang disebut Multiprogramming with a Fixed number of Task (MFT).

Tabel II menggambarkan sebagian jawaban yang diberikan oleh mahasiswa. Skor ditentukan oleh *human rater*, yang mana dalam penelitian ini adalah dosen yang memberikan pertanyaan. Nilai skor 2 diberikan untuk jawaban dengan label "benar", nilai skor 1 untuk jawaban dengan label "benar sebagian", dan nilai skor 0 untuk jawaban "salah". Dalam beberapa jawaban dengan label "benar sebagian", *human rater* memberikan nilai 1.5 atau 0.5.

Tabel 1 Potongan dataset

Pertanyaan: Apa yang dimaksud dengan MFT?		
No.	Jawaban	Skor
1.	Skema yang membuat memori dibagi menjadi beberapa blok dengan ukuran seragam (fixed size).	2
2.	Pengalokasian memori berurutan dengan besar memori tetap (fixed).	2
3.	Pengalokasian memori yang tetap dan tidak bisa dirubah-rubah lagi ukurannya sehingga pembagiannya sama besar.	2
4.	Model yang table sudah dibagi-bagi besarnya, sehingga proses hanya tinggal mencari yang besarnya sesuai dengan yang dibutuhkan juga yang tersedia.	2
5.	Alokasi tetap untuk besaran-besaran pada partisi yang telah ditentukan	2

Dataset di atas adalah *dataset* yang belum mengalami perubahan. *Dataset* yang digunakan dalam penelitian ini diperoleh dari ujian tertulis yang ditulis ulang dan telah memperoleh perbaikan pada beberapa bagian. Bagian-bagian yang diperbaiki antara lain:

- Jawaban yang menggunakan kosakata bahasa Inggris. Perbaikan yang dilakukan adalah mentransformasikan kosakata tersebut ke dalam bahasa Indonesia. Contoh: "table" menjadi "tabel", "system" menjadi "sistem", "memory" menjadi "memori". Pengecualian diberikan kepada kosakata bahasa Inggris yang menjadi esensi utama dari jawaban (*keyword)*.
- Jawaban yang menggunakan singkatan dalam bahasa Indonesia. Contoh: "yg" menjadi "yang", "scr" menjadi "secara".
- Jawaban yang memiliki kesalahan penulisan. Contoh: "pedekatan" menjadi "pendekatan".
- Jawaban yang memiliki kesalahan dalam penggunaan imbuhan. Contoh: "dimemori" menjadi "di memori"

Langkah tersebut dilakukan karena dalam penelitian ini belum menggunakan *spell corrector* otomatis.

3.2 Alat Penelitian

Alat penelitian yang digunakan dalam penelitian ini adalah sebagai berikut:

- Laptop dengan spesifikasi Intel Core i3, 2.1GHz, RAM 2GB, HDD 500GB.
- Software yang digunakan adalah Microsoft Windows 8.1, bahasa pemrograman Python 3.5, dan Microsot Excel 2013.

3.3 Evaluasi Metode

Model penilaian otomatis ini ditujukan untuk mengukur kemiripan nilai di antara dua teks, yaitu jawaban dosen dan jawaban siswa. Nilai kemiripan diperoleh dengan rumus *Cosine Similarity* dan *Jaccard Similarity* yang telah dibahas pada bagian sebelumnya. Selanjutnya, nilai yang dihasilkan oleh kedua metode dihitung korelasinya menggunakan *Pearson's Correlation* untuk mengukur kedekatan antara jawaban hasil komputasi dan penilaian manual yang dihasilkan dosen. Tingkat *agreement* yang sangat baik direpresentasikan dengan nilai korelasi 0.75 atau lebih [16]. Persamaan (3) berikut merupakan cara mencari korelasi *Pearson* antara nilai yang dihasilkan oleh guru (A) dan nilai yang dihasilkan oleh komputer (B):

$$Correlation(A,B) = \frac{covariance(A,B)}{standardDev(A) \times standardDev(B)} \quad (3)$$

4. HASIL DAN PEMBAHASAN

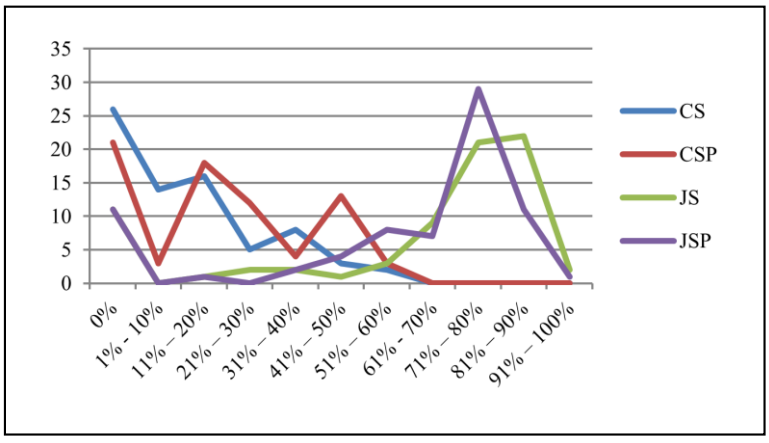
Metode *Cosine Similarity* dan *Jaccard similarity* yang digunakan pada penelitian ini hanya menilai kemiripan dua teks jawaban berdasarkan susunan leksikalnya. Oleh karena itu, diperlukan penyesuaian pada *dataset*, mengingat dalam penelitian ini *dataset* yang digunakan menggunakan soal dan jawaban dari ujian tertulis. Beberapa penyesuaian yang dilakukan adalah mengubah karakter dalam teks sesuai dengan kaidah bahasa Indonesia yang baik dan benar.

Tabel III menunjukkan distribusi nilai kemiripan dari jawaban siswa dan jawaban dosen. Pada percobaan 1 dilakukan penghitungan nilai *Cosine Similarity* (CS) dan *Jaccard Similarity* (JS) tanpa melalui pra-pemrosesan teks apapun. Sedangkan pada percobaan 2 dilakukan penghitungan nilai kemiripan dengan menggunakan pra-pemrosesan teks pada masing-masing metode yaitu *Cosine Similarity* (JSP) dan *Jaccard Similarity* (JSP). Hasilnya, pada *Cosine Similarity* nilai kemiripan meningkat pada percobaan 2, namun baik percobaan 1 dan percobaan 2 belum mampu mencapai nilai kemiripan di atas 60%. Sedangkan *Jaccard Similarity* menghasilkan nilai kemiripan yang tinggi pada kedua percobaan. Namun, masing-masing nilai kedua metode ini masih harus dihitung nilai korelasinya. HR menunjukkan penilaian manual yang diberikan oleh *human rater*.

Tabel 2 Distribusi Nilai Kemiripan

No.	Nilai Kemiripan	Jumlah Data				
		HR	***CS***	***JS***	***CSP***	***JSP***
1.	91 – 100%	34	-	2	-	1
2.	81% – 90%	-	-	22	-	11
3.	71% – 80%	1	-	21	-	29
4.	61% – 70%	-	-	9	-	7
5.	51% – 60%	-	2	3	3	8
6.	41% – 50%	2	3	1	15	4
7.	31% – 40%	-	8	2	4	2
8.	21% – 30%	2	5	2	13	-
9.	11% – 20%	-	16	1	17	1
10.	1% – 10%	-	14	-	3	-
11.	0%	35	26	11	21	11
Total		74	74	74	74	74

Nilai kemiripan 0% muncul karena dua alasan: (1) jawaban mahasiswa dianggap tidak mirip sama sekali oleh penilaian otomatis, (2) jawaban tidak diisi oleh mahasiswa (*blank answer*). Tabel IV menunjukkan distribusi nilai kemiripan 0%. Pada *Cosine Similarity*, nilai kemiripan 0% terlihat lebih mendekati kepada nilai 0 yang dihasilkan oleh *human rater*. Sedangkan pada pada *Jaccard similarity* memiliki kelemahan bahwa beberapa jawaban mendapatkan kredit yang tinggi meskipun pada kenyataannya memiliki nilai yang rendah pada penilaian oleh *human rater*.



Gambar 3 Grafik distribusi nilai kemiripan

Bagaimanapun, baik *metode Cosine Simiarity* dan *Jaccard Similarity* tidak ada yang mendekati nilai *human rater* pada nilai kemiripan 91-100%. Rendahnya nilai kemiripan ini dapat dipertimbangkan dalam sudut pandang bahwa *Cosine Similarity* dan *Jaccard Similarity* adalah metode yang memperhatikan urutan *term* dalam teks, sehingga kalimat dalam jawaban mahasiswa harus memunculkan jawaban yang sama persis dalam *keyword* jawaban dosen. Selain itu, *dataset* berisikan pertanyaan yang mengharuskan munculnya pengetahuan eksternal mahasiswa, sehingga diperlukan elemen semantik dalam proses penilaian otomatis ini. Penggunaan pra-pemrosesan teks pada penelitian ini mampu meningkatkan nilai kemiripan jawaban. Hal ini dikarenakan pra-pemrosesan teks menyaring kata-kata dalam bentuk yang lebih sederhana dan membuang kata-kata yang tidak terlalu penting yang dapat mengurangi nilai kemiripan teks. Distribusi nilai

kemiripan *Cosine Similarity* dan *Jaccard Similarity* ditunjukan pada Gambar 3. Sedangkan berikut ini merupakan contoh perhitungan korelasi antara skor *human rater* dan *Cosine Similarity* (CS):

$$\left.\begin{array}{l} covariance(X,Y) = 0{,}036547708 \\ standardDev(X) = 0{,}485323779 \\ standardDev(Y) = 0{,}148323564 \end{array}\right\} \rightarrow Correlation(X,Y) = 0{,}507713166$$

Selanjutnya, hasil perbandingan metode *Cosine Similarity* dan *Jaccard Similarity* ditunjukkan pada tabel 3.

Tabel 3 Hasil Korelasi *Pearson*

CS	***JS***	***CSP***	***JSP***
0,51	0,40	**0,62**	0,43

Berdasarkan hasil perhitungan nilai korelasi, metode *Cosine Similarity* dengan menerapkan teknik pra-pemrosesan teks menghasilkan nilai tertinggi yaitu 0,62. Namun nilai ini bukanlah suatu hasil yang memuaskan. Penelitian selanjutnya diharapkan mampu mengolah *dataset* dengan jumlah yang lebih banyak sebagai bahan perbandingan, sehingga efektivitas kedua metode dapat lebih diukur dengan lebih baik.

5. KESIMPULAN

Penggunaan metode *Cosine Similarity* dan *Jaccard Similarity* pada penilaian otomatis untuk jawaban pendek belum mampu mencapai nilai yang memuaskan. Hal ini dikarenakan kedua metode hanya menilai kemiripan berdasarkan susunan leksikalnya. Sementara itu, jawaban mahasiswa juga sangat bervariasi dan menggunakan kata-kata yang jauh berbeda dari jawaban kunci, walaupun pada dasarnya memiliki makna semantik yang sama. Oleh karena itu, pada penelitian selanjutnya diperlukan metode lain yang mampu menangani makna semantik pada jawaban. Bagaimanapun, metode *Cosine Similarity* dan *Jaccard Similarity* masih dapat dipertimbangkan untuk menilai jawaban pendek secara otomatis, dengan batasan bahwa pertanyaan yang digunakan mengharuskan jawaban dalam format *keyword* sehingga tidak memunculkan kata-kata lain yang mampu menurunkan nilai kemiripan. Sebagai bahan pertimbangan dalam mengukur efektivitas kedua metode ini, tugas selanjutnya adalah dengan melibatkan *dataset* yang lebih banyak dan pertanyaan yang lebih bervariasi.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Amikom Purwokerto yang telah memberi dukungan pendanaan pada penelitian ini.

DAFTAR PUSTAKA

[1] S. Burrows, I. Gurevych, and B. Stein, “The eras and trends of automatic short answer grading,” *Int. J. Artif. Intell. Educ.*, vol. 25, no. 1, pp. 60–117, 2015.

[2] M. K. Singley and H. L. Taft, “Open-ended approaches to science assessment using computers,” *J. Sci. Educ. Technol.*, vol. 4, no. 1, pp. 7–20, 1995.

[3] L. M. Rudner and T. Liang, “Automated Essay Scoring Using Bayes’ Theorem,” *J. Technol. Learn. Assess.*, vol. 1, no. 2, pp. 1–21, 2002.

[4] E. B. Page, "Grading Essays by Computer: Progress Report," *Invit. Conf. Test. Probl. 29 October, 1966*, vol. 47, no. 5, pp. 87–100, 1966.

[5] R. Siddiqi, C. J. Harrison, and R. Siddiqi, "Improving teaching and learning through automated short-answer marking," *IEEE Trans. Learn. Technol.*, vol. 3, no. 3, pp. 237–249, 2010.

[6] J. Sukkarieh and S. Stoyanchev, "Automating Model Building in c-rater," *Proc. 2009 Work. ...*, no. August, pp. 61–69, 2009.

[7] S. Jordan, "Student engagement with assessment and feedback: Some lessons from short-answer free-text e-assessment questions," *Comput. Educ.*, vol. 58, no. 2, pp. 818–834, 2012.

[8] C. Gütl, "e-Examiner : Towards a Fully-Automatic Knowledge Assessment Tool applicable in Adaptive E-Learning Systems," *Int. Conf. Interact. Mob. Comput. Aided Learn.*, vol. 1, no. 10, pp. 1–10, 2007.

[9] S. Roy, Y. Narahari, and O. D. Deskhmukh, "A perspective on computer assisted assessment techniques for short free-text answers.," in *International Computer Assisted Assessment Conference*, 2015, vol. 571, pp. 96–109.

[10] S. A. Abdurrasyid and Suyanto, "Implementasi dan Optimasi Algoritma Nazief dan Adriani untuk Stemming Dokumen Bahasa Indonesia," 2012.

[11] M. Adriani, J. Asian, B. Nazief, and H. E. Williams, "Stemming Indonesian : A Confix-Stripping Approach," *ACM Trans. Asian Lang. Inf. Process.*, vol. 6, no. 4, pp. 1–33, 2007.

[12] J. Asian, "Effective Techniques for Indonesian Text Retrieval," 2007.

[13] A. Z. Arifin, I. P. A. K. Mahendra, and H. T. Ciptaningtyas, "Enhanced Confix Stripping Stemmer and Ants Algorithm For Classifying News Document In Indonesian Language," in *Proceeding of International Conference on Information & Communication Technology and Systems (ICTS)*, 2009, pp. 149–158.

[14] A. D. Tahitoe and D. Purwitasari, "Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia dengan Metode Corpus Based Stemming," Surabaya, 2010.

[15] W. H.Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13–18, 2013.

[16] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.