khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Efektivitas Algoritma Semantik dengan Keterkaitan Kata dalam Mengukur Kemiripan Teks Bahasa Indonesia

Husni Thamrin[1*], Atiqa Sabardilla[2]
[1]Program Studi Informatika
Universitas Muhammadiyah Surakarta
Surakarta
*husni.thamrin@ums.ac.id
[2]Program Studi Bahasa dan Sastra Indonesia dan Daerah
Universitas Muhammadiyah Surakarta
Surakarta
atiqa.sabardila@ums.ac.id

Abstrak
Algoritma similaritas terhadap teks telah diterapkan pada berbagai aplikasi seperti deteksi plagiasi, pengelompokan dokumen, klasifikasi teks berita, mesin penjawab otomatis dan aplikasi penerjemahan bahasa. Beberapa aplikasi telah menunjukkan hasil yang baik. Sayangnya, upaya menerapkan algoritma similaritas semantik belum cukup berhasil terhadap teks bahasa Indonesia karena minimnya koleksi basis pengetahuan bahasa Indonesia, misalnya terkait keberadaan tesaurus atau *word net*. Penelitian ini berfokus pada upaya menghimpun hiponim dan meronim pada bahasa Indonesia, membangun korpus pasangan kalimat yang direview oleh penutur bahasa untuk menilai tingkat similaritas, dan mencermati efektivitas algoritma similaritas semantik dalam mengukur kemiripan kalimat bahasa Indonesia yang ada dalam korpus. Kemiripan kata diperoleh dari keterkaitan kata dalam bentuk sinonim, hiponim dan meronim sebagai basis pengetahuan. Penelitian ini menunjukkan bahwa penggunaan basis pengetahuan tersebut meningkatkan skor similaritas kalimat yang mengandung kata-kata yang berkaitan secara leksikal. Pada penelitian ini dihitung korelasi antara skor similaritas hasil perhitungan algoritma dengan skor kemiripan kalimat sebagaimana dipersepsikan oleh penutur bahasa. Tiga macam algoritma perhitungan telah diujicoba. Perhitungan similaritas menggunakan persentase jumlah kemunculan kata yang sama memberikan angka korelasi sebesar 0,7128. Angka korelasi untuk perhitungan similaritas menggunakan fungsi kosinus adalah sebesar 0,7408. Sedangkan perhitungan similaritas menggunakan algoritma semantik yang memperhatikan keterkaitan kata memberikan tingkat korelasi tertinggi sebesar 0,7508.

Kata kunci: similaritas, kemiripan teks, sinonim, hiponim, bahasa Indonesia

1. PENDAHULUAN

Similaritas antara dua teks atau kalimat merupakan angka yang menggambarkan kedekatan makna antara kedua teks atau kalimat. Perhitungan similaritas digunakan dalam berbagai keperluan, misalnya untuk melakukan pencarian informasi di internet, pencarian dokumen di harddisk, klasifikasi dokumen dalam arsip, deteksi plagiasi, dan kegiatan menganalisis informasi di dunia maya (*data analysis*) [1].

Proses pencarian informasi menerapkan algoritma similaritas untuk mengukur kemiripan makna kata atau frase yang dicari dengan teks yang ada dalam halaman yang ditelusuri. Pencarian informasi tidak cukup dilakukan dengan membandingkan kata atau frase yang dicari dengan kata atau frase yang ada dalam dokumen. Pencarian yang efektif memerlukan analisis mengenai kata yang dibandingkan dengan fitur dokumen. Sebagai contoh, hasil analisis makna kata dan frase yang dicari oleh *user* dan disertai penentuan tema dokumen dapat memperbaiki hasil pencarian. Fitur lain seperti *backlink* (banyaknya *link* ke sebuah website) yang dipadukan dengan skor similaritas akan menghasilkan daftar hasil pencarian yang mempunyai kemungkinan tinggi mengandung informasi yang dicari *user* [2].

Analisis similaritas dapat digunakan membantu proses klasifikasi dengan menentukan tag atau kata kunci yang paling tepat untuk sebuah dokumen. Pengklasifikasian kumpulan dokumen diperlukan pada sebuah perpustakaan digital untuk mengelompokkan

dokumen dengan subjek yang sama [3], [4]. Algoritma similaritas juga diterapkan dalam proses deteksi plagiasi, yaitu dengan membandingkan dua dokumen atau lebih dan menentukan tingkat kemiripan dari paragraf-paragraf yang ada dalam dokumen [5]. Adapun dalam kegiatan analisis data, algoritma similaritas digunakan untuk mendefinisikan kata yang dicermati beserta kata sejenis untuk dihitung frekuensi kemunculannya dalam berita di dunia maya atau dalam obrolan di situs media sosial.

Similaritas dua buah kalimat dapat ditentukan dengan algoritma similaritas semantik, yaitu algoritma yang memperhatikan makna kata yang menyusun kalimat. Penentuan similaritas secara semantik lebih akurat daripada perhitungan similaritas berdasarkan jumlah kata yang tepat sama [6]. Namun, penerapan algoritma similaritas semantik untuk teks bahasa Indonesia belum banyak dilakukan karena berbagai kendala di antaranya karena belum adanya jejaring kata bahasa Indonesia dan belum ada himpunan data uji yang standar (*standard test bed*) untuk pengujian algoritma [7].

Pada penelitian ini dibuat aplikasi yang menerapkan algoritma similaritas semantik berbasis jejaring kata untuk mengukur similaritas kalimat bahasa Indonesia, menyusun korpus pasangan kalimat bahasa Indonesia dan menguji efektivitas algoritma dalam mengukur kemiripan kalimat Bahasa Indonesia yang ada di dalam korpus.

Penulis mencermati berbagai penelitian terkait dengan analisis similaritas. Penelitian [8], misalnya, membandingkan berbagai algoritma similaritas terhadap kalimat dari majalah berbahasa Inggris dan mendapati pentingnya similaritas yang memperhatikan sinonim, sebagai salah satu elemen jejaring kata. Sedangkan [3] meneliti penerapan algoritma similaritas pada proses pengelompokan dokumen dan mendapati bahwa metode *related article* menghasilkan pengelompokan yang paling terkonsentrasi.

Pada [9] diteliti penerapan jarak Levenshtein sebagai landasan dalam menilai kemiripan jawaban siswa dengan kunci jawaban dan mendapati adanya korelasi antara jarak Levenshtein dengan skor jawaban guru jika jawaban siswa tidak membentuk kata yang dikenal kamus. Namun jika jawaban siswa membentuk kata baru, diperlukan perhitungan similaritas semantik untuk perhitungan skor.

Pada [10], [11] dicermati penerapan komponen jejaring kata dalam perhitungan similaritas dua teks untuk melakukan pengelompokan teks dan klasifikasi teks. Komponen jejaring kata yang digunakan adalah sinonim dan hiponim. Kedua peneliti menyatakan bahwa penggunaan algoritma similaritas berbasis jejaring kata tidak terbukti memperbaiki hasil pengelompokan teks (text clustering) kalimat singkat. Kedua peneliti mendapati bahwa algoritma yang serupa dapat memperbaiki secara signifikan kinerja proses klasifikasi teks singkat (*short text classification*).

Perhitungan similaritas semantik dengan memperhatikan keterkaitan kata diperkirakan meningkatkan korelasi antara nilai hasil perhitungan dengan persepsi penutur bahasa Indonesia terkait kemiripan makna kalimat bahasa Indonesia. Studi oleh [12] menggunakan WordNet untuk menentukan keterkaitan kata. Upaya dilakukan dengan memodifikasi vektor semantik pada fungsi kosinus dengan mengisi vektor dengan jarak relatif kata dalam WordNet. Dengan cara ini mereka mengklaim dapat memperbaiki kinerja algoritma dibandingkan metode yang diterapkan oleh banyak peneliti lain seperti pada [13]–[15].

Sementara itu, [16] menggunakan Google tri-gram sebagai skor keterkaitan kata dan dengan algoritma tertentu dapat menghitung similaritas dua kalimat dengan tingkat korelasi mendekati 0,9 terhadap penilaian pakar (*expert judgement*).

2. METODE

Penelitian ini menggunakan basis data pengetahuan (*knowledge base*) dari komponen jejaring kata, dan tidak murni menggunakan jejaring kata karena jejaring kata bahasa Indonesia belum tersedia. Komponen yang dimaksud adalah kamus dengan sinonim dan hiponim. Kamus yang digunakan mempunyai sekitar 72000 kata/frase dengan lebih dari 150000 definisi. Setiap kata atau frase mempunyai paling sedikit satu definisi dan sebagian mempunyai contoh kalimat. Sebagian kata atau frase mempunyai komponen keterkaitan kata, seperti sinonim, antonin, kata turunan, hiponim dan meronim.

Similaritas dua kalimat dihitung berdasarkan similaritas antar kata penyusun kalimat. Kami menggunakan hasil penelitian sebelumnya sebagai angka similaritas antar kata [17] seperti yang tersaji pada Tabel 1. Skor diperoleh dari hasil survei terhadap lebih dari 120 responden tentang penilaian terhadap kemiripan kata-kata yang berbeda. Angka pada tabel dapat dinormalisasi jika diinginkan sehingga kata yang tidak terkait mempunyai skor similaritas nol.

Tabel 1. Skor keterkaitan antar kata sebagai dasar perhitungan similaritas antar kalimat

Keterkaitan kata	Skor similaritas antar kata
Kata yang sama	1
Sinonim	0.8
Hiponim dan Hipernim	0.7
Holonim dan Meronim	0.6
Tidak terkait	0.3

Perhitungan similaritas antar kalimat dilakukan dalam penelitian ini menggunakan cara seperti yang dilakukan oleh [12] namun dengan modifikasi. Modifikasi yang dilakukan adalah menggunakan keterkaitan kata seperti Tabel 1, sedangkan [12] menggunakan jarak relatif kata pada jejaring kata bahasa Inggris.

Data uji yang digunakan adalah 114 pasangan kalimat yang telah direview oleh tingkat kemiripannya oleh penutur bahasa Indonesia. Salah satu langkah dalam penelitian ini adalah mengumpulkan pasangan kalimat yang dicari dari ribuan artikel di internet. Pasangan kalimat direview oleh dua orang penutur Bahasa Indonesia (yaitu mahasiswa atau dosen) untuk dinilai apakah kedua kalimat tersebut mempunyai kemiripan makna atau merupakan parafrase. Penilaian dilakukan dengan memberikan skor 0 – 3. Skor 0 berarti kedua pasangan kalimat tidak mirip sama sekali, skor 1 kurang mirip, skor 2 cukup mirip dan skor 3 sangat mirip atau bermakna sama. Jika kedua penutur memberi nilai yang sama, maka keputusan keduanya digunakan sebagai nilai kemiripan. Jika kedua penutur memberi nilai yang berbeda, dan perbedaannya tidak lebih dari 1 poin, maka pasangan kalimat tersebut direview oleh penutur ketiga sebagai penentu nilai kemiripan. Jika para penutur memberi nilai yang berbeda dengan lebih dari 1 poin,

maka hasil penilaian para penutur dianggap tidak cukup kompak atau masih terdapat beda pendapat yang besar sehingga hasil penilaian tersebut diabaikan dan tidak digunakan. Jumlah 114 pasangan kalimat berisi penilaian dengan perbedaan maksimal 1 poin, dan sudah merupakan saringan dari 143 pasangan kalimat yang direview.

Algoritma similaritas semantik kemudian diujikan terhadap pasangan kalimat tersebut. Algoritma akan menghitung kemiripan pasangan kalimat. Hasil penilaian algoritma berupa angka dari 0 sampai 1. Sebagai pembanding, dalam penelitian ini diuji pula fungsi kosinus dan perhitungan kemiripan berdasarkan persentase jumlah kata yang sama terhadap seluruh kata yang ada pada pasangan kalimat. Nilai hasil perhitungan kemudian ditarik korelasinya terhadap hasil review penutur. Korelasi positif tertinggi menunjukkan kinerja algoritma yang terbaik.

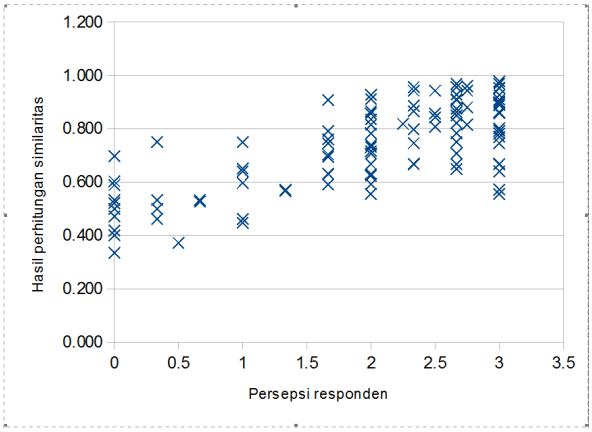
3. HASIL

Perhitungan skor similaritas dengan algoritma semantik telah dilakukan dan hasilnya, bersama skor similaritas berdasarkan persepsi penutur dan skor similaritas berdasarkan perhitungan fungsi kosinus, dapat dilihat sampelnya pada Tabel 2. Pada tabel terdapat kolom 'uid1' dan 'uid2' yang menunjukkan nomor kode kalimat dalam basis data, kolom 'Persen' yang merupakan hasil perhitungan similaritas berdasarkan jumlah kata yang sama dibanding jumlah seluruh kata pada kedua kalimat, dan kolom 'Persepsi responden' yang menunjukkan nilai rerata skor yang diberikan responden dalam skala 0 – 3. Selain itu terdapat kolom 'Perhitungan fungsi kosinus' yag berisi hasil perhitungan similaritas dengan fungsi kosinus, dan kolom 'Perhitungan algoritma semantik' yang merupakah hasil perhitungan dengan algoritma similaritas semantik berbasis jejaring kata.

Data pada Tabel 2 menunjukkan bahwa hasil perhitungan dengan fungsi kosinus memberikan nilai similaritas yang tidak banyak berbeda dengan hasil perhitungan dengan algoritma similaritas semantik berbasis jejaring kata. Dari sepuluh data yang disajikan pada Tabel 2, hanya satu yang menunjukkan nilai yang berbeda yaitu data nomor 7. Perhitungan algoritma similaritas semantik sedikit lebih baik dibanding hasil perhitungan dengan fungsi kosinus. Jika dicermati keseluruhan 114 data hasil penelitian (lihat Gambar 1 dan Gambar 2), terlihat pola yang serupa, yaitu tidak nyata terdapat perbaikan pada hasil perhitungan similaritas menggunakan algoritma similaritas semantik.

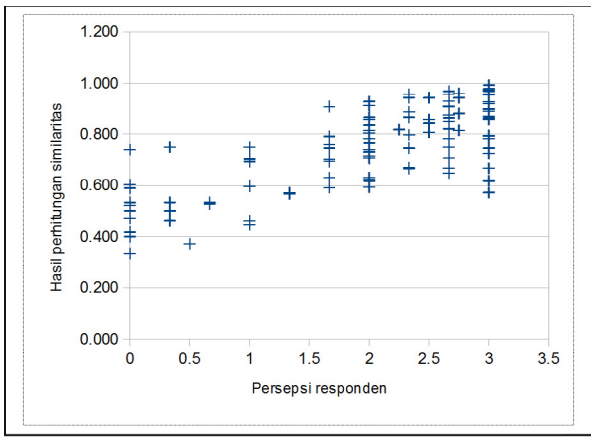
Tabel 2. Data pengamatan tentang persepsi responden terhadap kemiripan pasangan kalimat dan skor perhitungan dengan fungsi kosinus dan algoritma similaritas

No	uid1	uid2	Persen	Persepsi responden	Perhitungan fungsi kosinus	Perhitungan algoritma semantik
1	3988	1845	48	2	0,731	0,731
2	3992	1846	85	2	0,912	0,912
3	3994	1483	55	2	0,707	0,707
4	3995	1840	44	2	0,630	0,630
5	4029	2335	58	2	0,716	0,716
6	4031	2341	42	2	0,593	0,593
7	4033	2342	56	2	0,731	0,805
8	4034	2343	50	2,3333	0,669	0,669
9	4036	2339	64	2,3333	0,798	0,798
10	4118	1845	48	2	0,731	0,731



Gambar 1. Peta korelasi antara skor similaritas yang dihitung dengan fungsi kosinus dan skor kemiripan makna kalimat menurut persepsi responden

Selanjutnya dilakukan penentuan korelasi antara hasil perhitungan dan nilai similaritas berdasarkan persepsi responden (penutur bahasa Indonesia) yang diperoleh melalui survei. Persepsi responden (penutur) bagaimanapun merupakan tolok ukur kebaikan algoritma karena persepsi penutur menunjukkan bagaimana penutur memaknai kalimat dan menentukan kemiripan makna pasangan kalimat. Korelasi antara persepsi responden dengan perhitungan kemiripan berdasarkan persentase keberadaan kata yang sama bernilai 0,7128 (lihat Tabel 3). Angka korelasi meningkat menjadi 0,7408 jika similaritas dihitung menggunakan fungsi kosinus, yang berarti penggunaan fungsi kosinus meningkatkan akurasi dalam menentukan kemiripan kata jika dibandingkan dengan pemaknaan kalimat oleh penutur bahasa. Korelasi tertinggi dalam penelitian ini diperoleh jika perhitungan similaritas kalimat dilakukan menggunakan algoritma similaritas semantik berbasis jejaring kata.



Gambar 2. Peta korelasi antara skor similaritas yang dihitung dengan fungsi kosinus dan skor kemiripan makna kalimat menurut persepsi responden

Meskipuan algoritma similaritas semantik berbasis jejaring kata menghasilkan tingkat korelasi tertinggi, perbedaan nilai korelasi dibanding penggunaan fungsi kosinus tidak cukup signifikan, yaitu hanya sebesar 0,01 poin (setara dengan satu persen). Fungsi kosinus dan algoritma similaritas semantik berbasis jejaring kata sama-sama menghasilkan nilai korelasi yang berbeda cukup signifikan dibanding nilai korelasi yang dihasilkan oleh perhitungan similaritas dengan berdasarkan persentase keberadaan kata yang sama. Artinya sejauh pengamatan yang dilakukan dalam penelitian ini, penggunaan algoritma similaritas semantik berbasis jejaring kata meningkatkan skor similaritas jika dibandingkan dengan perhitungan similaritas sederhana menggunakan persentase jumlah kesamaan kata pada dua teks yang dibandingkan.

Namun, pengamatan pada penelitian ini juga menunjukkan bahwa tidak terdapat perbedaan signifikan dalam penggunaan algoritma similaritas semantik berbasis jejaring kata dibanding penggunaan fungsi kosinus.

Tabel 3. Korelasi antara persepsi responden terhadap kemiripan pasangan kalimat dengan skor similaritas hasil perhitungan

Nomor	Data yang dikorelasikan	Nilai korelasi
1	Perhitungan persen persepsi responden	0,7128
2	Perhitungan fungsi kosinus persepsi responden	0,7408
3	Perhitungan algoritma similaritas semantik persepsi responden	0,7508

4. DISKUSI

Implementasi dari algoritma similaritas semantik berbasis jejaring kata telah diupayakan untuk melakukan klustering [10] dan klasifikasi teks bahasa Indonesia [11]. Hasilnya menunjukkan bahwa proses klustering tidak menjadi lebih baik sedangkan proses klasifikasi memberikan hasil yang lebih baik secara signifikan. Jika digunakan fungsi kosinus dalam proses klasifikasi teks, diperoleh nilai kinerja F-Measure sebesar 0,4 sedangkan ketika digunakan algoritma similaritas semantik diperoleh nilai kinerja sebesar 0,595.

Penelitian terhadap dokumen berbahasa Inggris menunjukkan tingkat korelasi yang lebih tinggi dapat diperoleh dengan berbagai metode. Penelitian [18] menggunakan word net dan statistik korpus menghasilkan korelasi sebesar 0,816 dibandingkan penilaian similaritas oleh manusia (human judgement). Sedangkan penelitian [1] menggunakan faktor similaritas string (bukan hanya similaritas kata) menghasilkan korelasi sebesar 0,853. Selanjutnya [16] dengan memanfaatkan keterkaitan kata pada korpus Google Tri Grams dapat menghitung similaritas antar kalimat dengan korelasi sebesar 0,916 terhadap penilaian manusia.

Berbagai penelitian menunjukkan bahwa perhitungan similaritas kalimat berbahasa Inggris sejauh ini menunjukkan hasil yang lebih baik dalam melakukan perhitungan similaritas dibanding perhitungan similaritas pada kalimat atau dokumen berbahasa Indonesia. Faktor yang mempengaruhi hasil ini adalah keberadaan basis data pengetahuan (*knowledge base*). Bahasa Inggris telah memiliki tesaurus yang matang dan selalu *update*, dan juga telah memiliki jejaring kata. Sementara itu, Google Tri Grams untuk bahasa Indonesia belum pernah dipublikasikan oleh Google sedangkan versi bahasa Inggris dalam jumlah terbatas (satu juta entri) dapat diperoleh secara cuma-cuma sedangkan versi lengkap dapat diperoleh dengan biaya tertentu (yang artinya dijual). Oleh karena itu langkah yang diperlukan ke depan adalah membuat basis data pengetahuan kata bahasa Indonesia yang lebih lengkap. Lebih jauh lagi, untuk mendapatkan korpus berbahasa Indonesia sebaiknya diupayakan mewujudkan sistem sendiri untuk melakukan pencarian (*search engine*) maupun untuk melakukan aktivitas komunikasi dalam media sosial dengan meninggalkan ketergantungan kepada *search engine* maupun aplikasi media sosial yang berbasis di luar negeri. Upaya ini memungkinkan perusahaan dalam negeri dan peneliti dalam negeri mendapatkan data penelitian yang memadai terkait penggunaan teks bahasa Indonesia oleh penutur bahasa Indonesia.

5. KESIMPULAN

Algoritma similaritas semantik dapat dikembangkan agar memperhatikan keterkaitan antar kata (lexical relationship) seperti sinonim, hiponim, dan meronim. Keterkaitan antar kata menjadi bagian dari basis pengetahuan (knowledge base) dan penggunaan skor keterkaitan antar kata dapat memperbaiki skor similaritas antar kalimat/teks.

Penutur bahasa Indonesia mempunyai persepsi terhadap kemiripan dua kalimat. Hasil survei terhadap persepsi penutur menghasilkan data numerik tentang persepsi penutur terhadap kemiripan pasangan kalimat dalam korpus. Korelasi antara persepsi penutur dengan algoritma menunjukkan bahwa algoritma kosinus mempunyai korelasi 0,7408. Jika algoritma itu memperhatikan keterkaitan makna leksikal, didapat korelasi sebesar 0,7508, yang berarti terdapat perbaikan korelasi meskipun kurang signifikan.

DAFTAR PUSTAKA

[1] A. Islam and D. Inkpen, "Semantic text similarity using corpus-based word similarity and string similarity," *ACM Trans. Knowl. Discov. from Data*, vol. 2, no. 2, p. 10, 2008.

[2] E. D. Ochoa, "An Analysis of the Application of Selected Search Engine Optimization (SEO) Techniques and Their Effectiveness on Google's Search Ranking Algorithm," California State University, Northbridge, 2012.

[3] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, and K. Börner, "Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches," *PLoS One*, vol. 6, no. 3, p. e18029, 2011.

[4] Z. Sun, M. Errami, T. Long, C. Renard, N. Choradia, and H. Garner, "Systematic characterizations of text similarity in full text biomedical publications," *PLoS One*, vol. 5, no. 9, p. e12704, 2010.

[5] J. Malcolm and P. C. R. Lane, "Efficient search for plagiarism on the web," *Kuwait*, 2008.

[6] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in *AAAI*, 2006, vol. 6, pp. 775–780.

[7] J. Asian, H. E. Williams, and S. M. M. Tahaghoghi, "A Testbed for Indonesian Text Retrieval Jelita Asian," in *Proceedings of the 9th Australasian Document Computing Symposium*, 2004, no. June, pp. 2–5.

[8] J. Bao, C. Lyon, P. C. R. Lane, W. Ji, and J. Malcolm, "Comparing different text similarity methods," 2007.

[9] H. Thamrin, "Pengembangan Sistem Penilaian Otomatis Terhadap Jawaban Soal Pendek dan Terbuka dalam Evaluasi Belajar Online Berbahasa Indonesia," 2013.

[10] H. Thamrin and A. Sabardila, "Using Dictionary as a Knowledge Base for Clustering Short Texts in Bahasa Indonesia," in *International Conference on Data and Software Engineering*, 2014.

[11] H. Thamrin and A. Sabardila, "Utilizing Lexical Relationship in Term-Based Similarity Measure Improves Indonesian Short Text Classification," *ARPN J. Eng. Appl. Sci.*, 2015.

[12] H. Liu and P. Wang, "Assessing sentence similarity using wordnet based word similarity," *J. Softw.*, vol. 8, no. 6, pp. 1451–1458, 2013.

[13] M. E. Lesk, "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," in *Proceedings of SIGDOC Conference*, 1986.

[14] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification," in *WordNet, An Electronic Lexical Database*, The MIT Press, 1998.

[15] Z. Wu and M. Palmer, "Verb Semantics and Lexical Selection," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 1994.

[16] I. Islam, E. Milios, and V. Keselj, "Text Similarity Using Google Tri-Grams," in *25th Canadian Conference on Advances in Artificial Intelligence*, 2012, pp. 312–317.

[17] H. Thamrin and J. Wantoro, "An Attempt to Create an Automatic Scoring Tool of Short Text Answer in Bahasa Indonesia," in *Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2014)*, 2014.

[18] Y. Li, D. McLean, Z. Bandar, J. D. O'shea, K. Crockett, and others, "Sentence similarity based on semantic nets and corpus statistics," *Knowl. Data Eng. IEEE Trans.*, vol. 18, no. 8, pp. 1138–1150, 2006.