PENERAPAN *KNOWLEDGE-BASED TEXT SIMILARITY* UNTUK MENINGKATKAN AKURASI KLASIFIKASI KESESUAIAN PENULIS PADA DATA BIBLIOGRAFI

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

OLEH :

NAMA: ANNISA KARIMA R. HARAHAP

NIM: 09011181722029

JURUSAN SISTEM KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA
2021

HALAMAN PENGESAHAN

PENERAPAN *KNOWLEDGE-BASED TEXT SIMILARITY* UNTUK MENINGKATKAN AKURASI KLASIFIKASI KESESUAIAN PENULIS PADA DATA BIBLIOGRAFI

TUGAS AKHIR

Program Studi Sistem Komputer

Jenjang S1

Oleh

ANNISA KARIMA R. HARAHAP

09011181722029

Indralaya, Juli 2021

Mengetahui,

Ketua Jurusan Sistem Komputer	Pembimbing Tugas Akhir
Dr. Ir. H. Sukemi, M.T.	Firdaus, M.Kom.
NIP. 196612032006041001	NIP. 197801212008121003

HALAMAN PERSETUJUAN

Telah diuji dan lulus pada:

Hari : Rabu

Tanggal : 28 Juli 2021

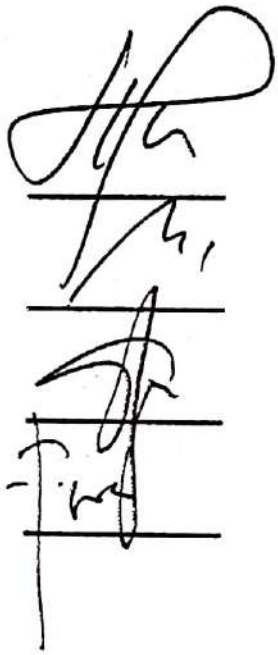
Tim Penguji :

1. Ketua : **Huda Ubaya, M.T.**

2. Sekretaris : **Adi Hermansyah, M.T.**

3. Penguji : **Prof. Dr. Ir. Siti Nurmaini, M.T.**

4. Pembimbing : **Firdaus, M.Kom.**



Mengetahui,

Ketua Jurusan Sistem Komputer

Dr. Ir. H. Sukemi M.T.

NIP. 196612032006041001

HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini:

Nama : Annisa Karima R. Harahap

NIM : 09011181722029

Judul : Penerapan *Knowledge-based Text Similarity* Untuk Meningkatkan Akurasi Klasifikasi Kesesuaian Penulis pada Data Bibliografi

Hasil Pengecekan Software *iThenticate/Turnitin* : 3%

Menyatakan bahwa laporan tugas akhir saya merupakan hasil karya sendiri dan bukan hasil penjiplakan atau plagiat. Apabila ditemukan unsur penjiplakan atau plagiat dalam laporan tugas akhir ini, maka saya bersedia menerima sanksi akademik dari universitas Sriwijaya.

Demikian, pernyataan ini saya buat dalam keadaan sadar dan tidak dipaksakan.

Indralaya, Juli 2021

Annisa Karima R. Harahap

09011181722029

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh, puji dan syukur penulis panjatkan kepada Allah SWT yang telah meberikan nikmat iman, nikmat kesehatan, nikmat taufik, karunia serta rahmat-Nya sehingga penulis dapat menyelesaikan Tugas Akhir yang berjudul "Penerapan *Knowledge-based Text Similarity* untuk Meningkatkan Akurasi Klasifikasi Kesesuaian Penulis pada Data Bibliografi".

Pada penyusunan Tugas Akhir ini, tentunya tidak terlepas dari bantuan, bimbingan serta dukungan dari berbagai pihak. Untuk itu, pada kesempatan yang baik ini, izinkan saya untuk mengucapkan rasa terima kasih kepada:

1. Kedua orang tua saya, Papa dan Mama. Mereka selalu mendukung saya secara penuh dan memberikan kasih sayang yang tiada tara.

2. Saudara-saudari saya, Ayuk Virna, Bang Rahmat, serta Vaniya yang selalu memberikan dukungan, selalu ada dan menghibur dikala sedang tidak merasa baik.

3. Bapak Jaidan Jauhari, S.Pd. M.T selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.

4. Bapak Dr. Ir. H. Sukemi, M.T. selaku Ketua Jurusan Sistem Komputer Fakultas Ilmu Komputer Universitas Sriwijaya.

5. Bapak Dr. Erwin, S.Si, M.Si, selaku Dosen Pembimbing Akademik di Jurusan Sistem Komputer.

6. Bapak Firdaus, S.T., M.Kom, selaku Pembimbing Tugas Akhir yang selalu mengarahkan dan memberi saran terkait penyusunan Tugas Akhir ini serta memberikan motivasi dan ilmu yang pastinya akan berguna untuk penulis.

7. Ibu Prof. Dr. Ir Siti Nurmaini, M.T. selaku Head of Intelligent System Research Group (ISysRG) yang telah memberi kesempatan besar untuk menjadi bagian dari team *research group* ini.

8. Kak Naufal Rachmatullah, S.Kom., M.T., Mbak Ade Iriani Safitri, M.Kom. dan Mbak Annisa Darmawahyuni, M.Kom.

9. Rekan seperjuangan di *group ISysRG* yang selalu membantu dalam pengerjaan Tugas Akhir ini.

10. Team *Teks Processing*. Qiliq selaku *leader*, serta Suci, Azis, Irvan Wais dan Jorgi yang selalu membantu penulis.

11. Teman dekat saya di jurusan ini, AAPS (Alna, Putri, Suci) yang selalu ada dan turut menemani dalam menyelesaikan Tugas Akhir ini.

12. Teman SMP saya Dinda dan Alfin yang selalu mendegarkan keluh kesah saya dalam proses pembuatan Tugas Akhir.

13. Kakak tingkat, teman-teman seperjuangan Sistem Komputer angkatan 2017, serta semua pihak yang tidak dapat penulis sebut satu-persatu.

Dalam pembuatan Tugas Akhir ini terlampau jauh dari kata sempurna. Untuk itu penulis selalu menerima kritik dan saran guna membangun Tugas Akhir ini menjadi lebih baik dan dapat menjadi refrensi untuk penelitian selanjutnya

Wassalamu'alaikum Warrahmatullahi Wabarakatuh.

Indralaya, Juli 2021

Penulis

Annisa Karima R. Harahap
09011181722029

IMPLEMENTATION KNOWLEDGE-BASED TEXT SIMILARITY TO INCREASING ACCURACION OF AUTHOR MATCHING CLASSIFICATION ON DATA BIBLIOGRAPHY

ANNISA KARIMA R. HARAHAP (09011181722029)

Computer Engineering Department, Computer Science Faculty, Sriwijaya University

Email: annisakarima72@gmail.com

ABSTRACT

In research using data in the form of text, a problem theme can be drawn called Author Name Disambiguation (AND). The theme can include a case of synonymy and polysemy that occurs when classifying authors in bibliographic data. This research will then use Deep Neural Network (DNN) as a classifier used to classify authors. Before the classification process is carried out, the data must enter the pre-processing stage. For feature extraction, the level of similarity will be calculated after combining data using cosine similarity for the author name, author list, and venue attributes as well as absolute reduction for the year attribute. Specifically for the title attribute, the knowledge-based text similarity method will be used to calculate the level of similarity which is carried out using two approaches and wordnet as the ontology. Furthermore, the results will be evaluated using performance measurement as a reference for its success. Accuracy obtained reached 99% for both knowledge-based text similarity approaches, namely path and wu palmer.

Keywords **:** *Author Name Disambiguation, Synonymy, Polysemy, Bibliographic Data, Deep Neural Network* (DNN), *Knowledge-Based Text Similarity, Cosine Similarity*.

PENERAPAN *KNOWLEDGE-BASED TEXT SIMILARITY* UNTUK MENINGKATKAN AKURASI KLASIFIKASI KESESUAIAN PENULIS PADA DATA BIBLIOGRAFI

ANNISA KARIMA R. HARAHAP (09011181722029)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya

Email: annisakarima72@gmail.com

ABSTRAK

Dalam penelitian dengan menggunakan data yang berbentuk teks dapat ditarik sebuah tema permasalahan yang disebut dengan *Author Name Disambiguation* (AND). Tema tersebut dapat mencakup sebuah kasus yaitu sinonim dan polisemi yang terjadi ketika mengklasifikasi penulis yang ada pada data bibliografi. Penelitian ini kemudian akan menggunakan *Deep Neural Network* (DNN) sebagai *classifier* yang digunakan untuk mengklasifikasi penulis. Sebelum dilakukan proses klasifikasi, data harus masuk ke dalam tahap pra-pemrosesan. Untuk ekstraksi fitur akan dilakukan perhitungan tingkat kesamaan setelah dilakukan kombinasi data dengan menggunakan *cosine similarity* untuk atribut *author name*, *author list*, dan *venue* serta pengurangan *absolute* untuk atribut *year*. Terkhusus atribut *title* akan digunakan metode *knowledge-based text similarity* untuk menghitung tingkat kesamaannya dimana dilakukan dengan dua pendekatan serta wordnet sebagai ontologinya. Selanjutnya hasil akan dievaluasi dengan menggunakan *performance measurement* sebagai acuannya keberhasilannya. Akurasi yang diperoleh mencapai 99% untuk kedua pendekatan *knowledge-based text similarity* yaitu path dan wu palmer.

Kata Kunci : *Author Name Disambiguation,* Sinonim, Polisemi, Data Bibliografi, *Deep Neural Network* (DNN), *Knowledge-Based Text Similarity*, *Cosine Similarity*.

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

BAB I
PENDAHULUAN

1.1. Latar Belakang

Seorang penulis yang telah melakukan penelitian dan mempublishnya ke dalam bibliografi database memiliki data ejaan nama yang beragam. Beberapa contoh bibiliografi *online* akan berkurang kualitas informasinya jika terjadi suatu kasus berupa keambiguitas pada nama. Maksud dari ambiguitas disini adalah penulis bisa saja konsisten dengan menuliskan namanya selalu serupa pada setiap penelitian yang telah dibuat. Bisa juga mereka memperkenalkan diri dengan urutan karakter nama dengan banyak cara. Kasus tersebut akan menjadi masalah serius karena akan mengurangi efisiensi dalam pengambilan informasi pada tempat publikasi online. Sehingga telah banyak dilakukan penelitian mengenai masalah ini[1]–[3] dengan berbagai macam pendekatan dan salah satunya dengan menggunakan *supervised machine learning*[4].

Nama penulis merupakan data teks sehingga dalam pengolahannya dapat masuk kedalam kategori *Word Sense Disambiguation* (WSD) terlebih jika terjadi kebingungan dalam menganalisis nama[5]. Cabang dari WSD itu sendiri adalah permasalahan *Author Name Disambiguation* (AND). Salah satu kasus yang dapat diatasi adalah ketaksaan nama. Misal ada dua nama yaitu “Ridwan Hakim” dan “Ridwan Hanan”. Kedua nama itu dapat disingkat menjadi Ridwan H dan menjadi tampak serupa sehingga ketika mencari nama tersebut bisa saja bukan orang yang dimaksud[6]. Adanya kelengkapan dalam sebuah kepenulisan hasil penelitian dapat membantu mengatasi masalah tersebut dengan menggunakan *author grouping*[4], [7].

Jika ingin meneliti tentang permasalahan kemiripan nama antara dua orang tentu dibutuhkan sebuah pendekatan yang dapat mengukur tingkat kemiripan. Perbedaan antara dokumen teks dapat terbagi menjadi dua yaitu perbedaan yang didasari oleh urutan karakter teks dan perbedaan berdasarkan arti atau relasi antar

kata dua teks tersebut. Perbedaan berdasarkan arti kemudian dapat dihitung dengan metode *semantic similarity*[7]. Kedua penulis kemudian dapat disandingkan untuk mencari kemiripannya. Dalam dataset yang berbentuk dataframe, penulis dapat dikombinasi baris per barisnya. Setelah itu fitur berupa isi dataset akan terbentuk berdasarkan nilai dari tingkat kemiripannya[8].

DNN, adalah salah satu metode *supervised learning* dengan bentuk jaringan neural yang memiliki banyak lapisan. DNN relative stabil digunakan untuk data *training* yang cukup representatif.[9] Beberapa penelitian mengenai *image classification*[10] dan pengenalan suara[11] telah menggunakan DNN sebagai *classifier* dan memperoleh hasil yang bagus. Untuk di bidang disambiguasi nama juga telah dilakukan dengan menggunakan metode DNN[9]. Untuk itu, penelitian ini akan mengaplikasikan struktur DNN untuk menghasilkan akurasi serta performa dalam proses klasifikasi dengan menggunakan pendekatan *author grouping*. Perlu digaris bawahi di sini ialah fitur "judul penelitian" akan dicari kesamaannya dengan menggunakan pengukuran kesamaan berbasis pengetahuan. Hal ini dikarenakan "judul penelitian" termasuk dalam kategori *short text*[12] dimana memiliki kata atau kalimat yang memiliki makna secara semantik.

1.2. Tujuan dan Manfaat

1.2.1. Tujuan

Dalam penelitian ini terdapat beberapa tujuan penulisan Tugas Akhir, yaitu:

1. Menggunakan pengukuran kesamaan berbasis pengetahuan terhadap atribut judul penelitian untuk meningkatkan akurasi.
2. Menggunakan *Classifier* DNN untuk mengatasi permasalahan mengenai kesesuaian penulis.

1.2.2. Manfaat

Dalam Penyusunan tugas akhir, diharapkan mampu memberi manfaat berupa:

1. Dapat mengaplikasikan pendekatan fungsi kesamaan berbasis pengetahuan dan metode *supervised learning* sebagai cara untuk mengatasi perkara kesesuaian penulis.

2. Dapat menjadi bahan refrensi untuk terkait permasalahan AND terkhusus untuk kasus kesesuaian penulis.

1.3. Perumusan dan Batasan Masalah

1.3.1. Perumusan Masalah

Bagaimana cara meningkatkan akurasi terhadap dataset dengan menggunakan pengukuran kesamaan *Knowledge-Based Text Similarity* untuk diterapkan pada kasus kesesuaian penulis?

1.3.2. Batasan Masalah

1. Tema yang diambil dalam penelitian adalah *Author Name Disambiguation*.

2. Penelitian diproses dengan menggunakan bahasa pemrograman *python*.

3. Mengambil dataset yang telah bersih dan digunakan oleh peneliti lain[13].

4. Menggunakan *performance measurement* sebagai takaran keberhasilan penelitian.

1.4. Metodologi Penelitian

Untuk menyelesaikan Tugas Akhir diterapkan beberapa metode seperti:

1.4.1. Metode Studi Pustaka dan Literatur

Metode ini dilakukan dengan melakukan pencarian berbagai macam refrensi atau sumber baik dari internet atau buku guna menunjang keberhasilan penelitian. Sumber yang dicari berpusat pada tema penelitian yang diusung yaitu AND dengan menggunakan metode DNN dan *Knowledge-based Text Similarity*.

1.4.2. Metode Konsultasi

Metode ini dilakukan dengan melakukan konsultasi, wawancara, atau bertanya kepada orang yang ahli perihal permasalahan yang akan diteliti.

1.4.3. Metode Pembuatan Model

Penggunaan bahasa pemrograman diterapkan dalam penelitian ini dan bertujuan untuk membangun sebuah model yang nantinya dapat diaplikasikan.

1.4.4. Metode Pengujian dan Validasi

Metode ini ditujukan terhadap hasil penelitian yang telah divalidasi sehingga kedepannya dapat menjadi bahan refrensi dan membuat penelitian lebih baik lagi.

1.4.5. Metode Hasil dan Analisa

Metode ini ditujukan terhadap hasil penelitian yang telah dinilai dan dianalis sehingga dapat dijadikan ukuran keberhasilan untuk penelitian selanjutnya.

1.4.6. Metode Penarikan Kesimpulan dan Saran

Penarikan kesimpulan dan saran akan dilakukan pada metode ini sehingga dapat menjadi acuan atau sumber refrensi untuk penelitan lebih lanjut.

1.5. Sistematika Penulisan

Dalam penelitian ini, dibutuhkan sebuah sistematika penulisan dari setiap Bab agar dapat mempermudah menyusun Tugas Akhir yang kemudian dapat dilihat sistematika tersebut sebagai berikut ini:

BAB I – PENDAHULUAN

Pada bab pertama akan menunjukkan pemaparan yang sistematis mengenai latar belakang, tujuan penelitian, rumusan masalah, dan sistematika penulisan.

BAB II – TINJAUAN PUSTAKA

Bab kedua berisi penjelasan mengenai Konsep, Dasar Teori, dan Prinsip Dasar yang diperlukan untuk menyelesaikan permasalahan dalam penelitian.

BAB III – METODOLOGI

Pada bab ketiga akan dijelaskan mengenai proses dan tahapan metodologi yang akan dilakukan dalam penelitian. Metodologi yang digunakan akan dibahas secara rinci tentang teknik, metode, dan alur proses yang dilakukan dalam penelitian.

BAB IV – HASIL DAN PEMBAHASAN

Pada bab ke empat akan terdapat penjelasan mengenai pembahasan hasil, hasil pengujian, analisis, kelebihan dan kekurang dari penelitian yang telah dilaksanakan.

BAB V – KESIMPULAN DAN SARAN

Pada bab kelima akan terdapat kesimpulan yang ditarik berdasarkan hasil penelitian serta saran yang dapat membangun mengenai tugas akhir bertemakan AND.

DAFTAR PUSTAKA

[1] I. Hussain and S. Asghar, "Author Name Disambiguation by Exploiting Graph Structural Clustering and Hybrid Similarity," *Arab. J. Sci. Eng.*, vol. 43, no. 12, pp. 7421–7437, 2018.

[2] A. A. Ferreira, A. Veloso, M. A. Gonçalves, and A. H. F. Laender, "Effective self-training author name disambiguation in scholarly digital libraries," in *Proceedings of the 10th annual joint conference on Digital libraries*, 2010, pp. 39–48.

[3] Q. Lin, B. Wang, Y. Du, X. Wang, Y. Li, and S. Chen, "Disambiguating authors by pairwise classification," *Tsinghua Sci. Technol.*, vol. 15, no. 6, pp. 668–677, 2010.

[4] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender, "A brief survey of automatic methods for author name disambiguation," *Acm Sigmod Rec.*, vol. 41, no. 2, pp. 15–26, 2012.

[5] G. Mann and D. Yarowsky, "Unsupervised personal name disambiguation," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, 2003, pp. 33–40.

[6] D. K. Sanyal, P. K. Bhowmick, and P. P. Das, "A review of author name disambiguation techniques for the PubMed bibliographic database," *J. Inf. Sci.*, p. 0165551519888605, 2019.

[7] W. H. Gomaa, A. A. Fahmy, and others, "A survey of text similarity approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13–18, 2013.

[8] Z. YAMANI, S. NURMAINI, W. K. SARI, and others, "Author Matching Using String Similarities and Deep Neural Networks," in *Sriwijaya International Conference on Information Technology and Its Applications*

(SICONIAN 2019), 2020, pp. 474–479.

[9] H. N. Tran, T. Huynh, and T. Do, "Author name disambiguation by using deep neural network," in *Asian Conference on Intelligent Information and Database Systems*, 2014, pp. 123–132.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[11] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[12] M. Shoaib, A. Daud, M. Sikandar, and H. Khiyal, "Improving Similarity Measures for Publications with Special Focus on Author Name Disambiguation," 2015.

[13] M. Shoaib, A. Daud, and T. Amjad, "Author Name Disambiguation in Bibliographic Databases: A Survey," *arXiv Prepr. arXiv2004.06391*, 2020.

[14] D. D. Prasetya, A. P. Wibawa, and T. Hirashima, "The performance of text similarity algorithms," *Int. J. Adv. Intell. Informatics*, vol. 4, no. 1, pp. 63–69, 2018.

[15] I. Hussain and S. Asghar, "A survey of author name disambiguation techniques: 2010-2016.," *Knowl. Eng. Rev.*, vol. 32, p. e22, 2017.

[16] R. Mihalcea, C. Corley, C. Strapparava, and others, "Corpus-based and knowledge-based measures of text semantic similarity," in *Aaai*, 2006, vol. 6, no. 2006, pp. 775–780.

[17] S. Jain, S. KR, and R. Jindal, "Identification of New Parameters for Ontology Based Semantic Similarity Measures," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 6, no. 20, 2019.

[18] S. Patwardhan, S. Banerjee, and T. Pedersen, “Using measures of semantic relatedness for word sense disambiguation,” in *International conference on intelligent text processing and computational linguistics*, 2003, pp. 241–257.

[19] B. A. H. Murshed, H. D. E. Al-Ariki, and S. Mallappa, “Semantic Analysis Techniques using Twitter Datasets on Big Data: Comparative Analysis Study.,” *Comput. Syst. Sci. Eng.*, vol. 35, no. 6, pp. 495–512, 2020.

[20] A. Gupta, A. Kumar, and J. Gautam, “A survey on semantic similarity measures,” *Int. J. Innov. Res. Sci. Technol.*, vol. 3, no. 12, pp. 243–247, 2017.

[21] Y. Shao, G. Xu, M. Xu, and L. Dong, “An Automatic Question Answering Method for Small-Scale Corpus,” in *Journal of Physics: Conference Series*, 2020, vol. 1621, no. 1, p. 12113.

[22] G. Zhu and C. A. Iglesias, “Computing semantic similarity of concepts in knowledge graphs,” *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 72–85, 2016.

[23] B. Jia, X. Huang, and S. Jiao, “Application of semantic similarity calculation based on knowledge graph for personalized study recommendation service,” *Educ. Sci. Theory Pract.*, vol. 18, no. 6, 2018.

[24] K. Jayakodi, M. Bandara, and I. Perera, “An automatic classifier for exam questions in Engineering: A process for Bloom’s taxonomy,” in *2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 2015, pp. 195–202.

[25] P. Sravanthi and B. Srinivasu, “Semantic similarity between sentences,” *Int. Res. J. Eng. Technol.*, vol. 4, no. 1, pp. 156–161, 2017.

[26] X. Zhu, F. Li, H. Chen, and Q. Peng, “An efficient path computing model for measuring semantic similarity using edge and density,” *Knowl. Inf. Syst.*, vol. 55, no. 1, pp. 79–111, 2018.

[27] A. A. Ilham, I. Nurtanio, and others, “Image search optimization with web scraping, text processing and cosine similarity algorithms,” in *2020 IEEE International Conference on Communication, Networks and Satellite (Comnetsat)*, 2020, pp. 346–350.

[28] R. Subhashini and V. J. S. Kumar, “Evaluating the performance of similarity measures used in document clustering and information retrieval,” in *2010 first international conference on integrated intelligent computing*, 2010, pp. 27–31.

[29] A. Pawar and V. Mago, “Calculating the similarity between words and sentences using a lexical database and corpus statistics,” *arXiv Prepr. arXiv1802.05667*, 2018.

[30] A. Gupta and K. K. Goyal, “Classification of Semantic Similarity Technique between Word Pairs using Word Net.”

[31] N. Varghese and M. Punithavalli, “Semantic Similarity Analysis on Knowledge Based and Prediction Based Models.”

[32] Y.-Y. Lee, H. Ke, T.-Y. Yen, H.-H. Huang, and H.-H. Chen, “Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement,” *J. Assoc. Inf. Sci. Technol.*, vol. 71, no. 6, pp. 657–670, 2020.

[33] Y. Cai, Q. Zhang, W. Lu, and X. Che, “A hybrid approach for measuring semantic similarity based on IC-weighted path distance in WordNet,” *J. Intell. Inf. Syst.*, vol. 51, no. 1, pp. 23–47, 2018.

[34] J. J. Lastra-D\’\iaz, A. Garc\’\ia-Serrano, M. Batet, M. Fernández, and F. Chirigati, “HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset,” *Inf. Syst.*, vol. 66, pp. 97–118, 2017.

[35] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, “Improving text preprocessing for student complaint document

classification using sastrawi," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 874, no. 1, p. 12017.

[36] A. A. Farisi, Y. Sibaroni, and S. Al Faraby, "Sentiment analysis on hotel reviews using Multinomial Na{\"\i}ve Bayes classifier," in *Journal of Physics: Conference Series*, 2019, vol. 1192, no. 1, p. 12024.

[37] P. P. Ramadhani and S. Hadi, "Text classification on the Instagram caption using support vector machine," in *Journal of Physics: Conference Series*, 2021, vol. 1722, no. 1, p. 12023.

[38] V. C. Mawardi, N. Susanto, and D. S. Naga, "Spelling correction for text documents in Bahasa Indonesia using finite state automata and Levinshtein distance method," in *MATEC Web of Conferences*, 2018, vol. 164, p. 1047.

[39] P. M. Prihatini, I. Putra, I. Giriantari, and M. Sudarma, "Indonesian text feature extraction using gibbs sampling and mean variational inference latent dirichlet allocation," in *2017 15th International Conference on Quality in Research (QiR): International Symposium on Electrical and Computer Engineering*, 2017, pp. 40–44.

[40] I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, "Study of hoax news detection using na{\"\i}ve bayes classifier in Indonesian language," in *2017 11th International Conference on Information & Communication Technology and System (ICTS)*, 2017, pp. 73–78.

[41] H. N. Rohman and I. Asror, "Automatic detection of argument components in text using multinomial Nave Bayes clasiffier," in *Journal of Physics: Conference Series*, 2019, vol. 1192, no. 1, p. 12034.

[42] Z. Yao and C. Ze-wen, "Research on the construction and filter method of stop-word list in text preprocessing," in *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, 2011, vol. 1, pp. 217–221.

[43] R. Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the

Accuracy of Text Classification using Stemming Method, A Case of Informal Indonesian Conversation," 2020.

[44] S. N. Khera and Divya, "Predictive modelling of employee turnover in Indian IT industry using machine learning techniques," *Vision*, vol. 23, no. 1, pp. 12–21, 2018.

[45] T. Zhang, S. Song, S. Li, L. Ma, S. Pan, and L. Han, "Research on gas concentration prediction models based on LSTM multidimensional time series," *Energies*, vol. 12, no. 1, p. 161, 2019.

[46] T. H. Nguyen and J.-D. Zucker, "Enhancing metagenome-based disease prediction by unsupervised binning approaches," in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, 2019, pp. 1–5.

[47] B. B. Tirkey and B. S. Saini, "Proposing model for recognizing user position," in *First international conference on sustainable technologies for computational intelligence*, 2020, pp. 155–162.

[48] D. Harlianto, S. Mardiyati, D. Lestari, A. H. Zili, and S. Devila, "Indonesia tuberculosis morbidity rate forecasting using recurrent neural network," in *AIP Conference Proceedings*, 2020, vol. 2242, no. 1, p. 30006.

[49] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Mach. Learn.*, vol. 20, no. 3, pp. 197–243, 1995.

[50] V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," *Int. J. Artif. Intell. Appl.*, vol. 3, no. 2, p. 85, 2012.

[51] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv Prepr. arXiv1702.04595*, 2017.

[52] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of

deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[53] S. Sangsavate, S. Tanthanongsakkun, and S. Sinthupinyo, "Stock Market Sentiment Classification from FinTech News," in *2019 17th International Conference on ICT and Knowledge Engineering (ICT&KE)*, 2019, pp. 1–4.

[54] J. Kim, "Evaluating author name disambiguation for digital libraries: a case of DBLP," *Scientometrics*, vol. 116, no. 3, pp. 1867–1886, 2018.

[55] J. Kim and J. Kim, "The impact of imbalanced training data on machine learning for author name disambiguation," *Scientometrics*, vol. 117, no. 1, pp. 511–526, 2018.