PENERAPAN PENGUKUR KESAMAAN ATRIBUT JUDUL BERBASIS *STRING* PADA DATA BIBLIOGRAFI UNTUK MENINGKATKAN KINERJA KLASIFIKASI KESAMAAN PENULIS

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat

Memperoleh Gelar Sarjana Komputer

OLEH :

MOHAMMAD EL QILIQSANDY

09011281722064

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2022

HALAMAN PENGESAHAN

PENERAPAN PENGUKUR KESAMAAN ATRIBUT JUDUL BERBASIS STRING PADA DATA BIBLIOGRAFI UNTUK MENINGKATKAN KINERJA KLASIFIKASI KESAMAAN PENULIS

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh

MOHAMMAD EL QILIQSANDY
09011281722064

Indralaya, Juli 2022

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

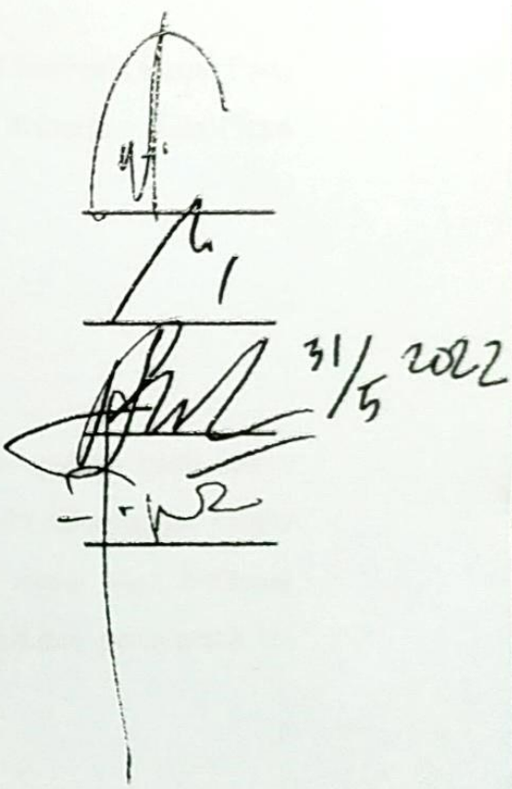
Tim Penguji :

1. Ketua : Ahmad Zarkasi, S.T., M.T.

2. Sekretaris : Adi Hermansyah, M.T.

3. Penguji : Dr. Erwin, M.Si.

4. Pendamping : Firdaus, S.T., M.Kom.



Mengetahui,

Ketua Jurusan Sistem Komputer

Dr. Ir. H. Sukemi M.T.

NIP. 196612032006041001

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Mohammad El Qiliqsandy

NIM : 09011281722064

Judul : Penerapan Pengukur Kesamaan Atribut Judul Berbasis String Pada Data Bibliografi Untuk Meningkatkan Kinerja Klasifikasi Kesamaan Penulis

Hasil pengecekan *Software Turnitin* : %

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.

Palembang, Juli 2022

Mohammad El Qiliqsandy

KATA PENGANTAR

Assalamu'alaikum Wr.Wb.

Alhamdulilahirabbil'alamin, puji dan syukur penulis panjatkan kepada Allah SWT yang telah melimpahkan nikmat, taufik, dan hidayah-Nya yang sangat besar dan tidak pernah berhenti kepada penulis sehingga penulis dapat menyelesaikan Tugas Akhir ini yang berjudul **"Penerapan Pengukur Kesamaan Atribut Judul Berbasis String Pada Data Bibliografi Untuk Meningkatkan Kinerja Klasifikasi Kesamaan Penulis"**.

Dalam tugas akhir ini penulis menjelaskan mengenai pengolahan nilai kemiripan untuk identifikasi dan klasifikasi penulis terhadap suatu publikasi dengan disertai data-data yang diperoleh penulis saat melakukan peneilitian dan pengujian data. Penulis berharap agar tulisan ini dapat bermanfaat bagi orang banyak dan menjadi bahan bacaan bagi yang tertarik untuk meneliti permodelan pada bidang data science untuk sub-topik Author Name Disambiguation (AND) khususnya di bidang.

Pada kesempatan ini, dengan segala kerendahan hati penulis mengucapkan terima kasih kepada semua pihak atas bantuan, bimbingan, dan saran yang telah diberikan dalam menyelesaikan Tugas Akhir ini, antara lain:

1. Orang tua saya tercinta yang telah membesarkan saya dengan penuh kasih sayang dan selalu mengajarkan saya dalam berbuat hal yang baik. Terima kasih untuk segala doa, motivasi dan dukungannya baik moril, materiil maupun spiritual selama ini.

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

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

4. Bapak Firdaus, S.T., M.Kom, selaku Dosen Pembimbing Akademik sekaligus Pembimbing Tugas Akhir yang telah berkenan meluangkan waktunya guna membimbing, memberikan saran dan motivasi serta bimbingan terbaik untuk penulis dalam menyelesaikan Tugas Akhir ini.
5. Teman-teman laboratorium ISYSRG yang menemani perjalanan ini. Irvan, Suci, Annisa, Wais, Azis, Jorgi dari kelompok teks yang dari awal sama-sama tidak tahu sampai sekarang.

Penulis menyadari bahwa Tugas Akhir ini masih sangat jauh dari kata sempurna. Untuk itu kritik dan saran dst.

Wassalamu'alaikum Wr. Wb.

Indralaya, Juli 2022
Penulis,

Mohamad El Qiliqsandy
NIM. 09011281722064

Application of Title Attribute Equality Measurement Based on String in Bibliographic Data to Improve The Performance of Authors Similarity Classification

Mohammad El Qiliqsandy (09011281722064)

Computer Engineering Department, Computer Science Faculty, Sriwijaya University

Email : elqiliqsandy@gmail.com

ABSTRACT

Author Name Disambiguation (AND) is a problem of name ambiguity to the publication in a Digital Library (DL) database caused by the Homonymy and Synonymy of the author's name. The proposed method is a Deep Neural Network (DNN) and Support Vector Machine (SVM) to classify data. The DBLP Labeled Data dataset by Jinseok Kim, et. al. is used for the classification task. This study concerned with processing data with the techniques of normalization and transformation data to create an effective feature for classification. The performance evaluation of the research conducted is accuracy, precision, and recall. The parameters are important to evaluate the AND classification process, especially the identification of the author. For the result, DNN achieves accuracy, precision, and recall, which is 99.98%, 97.71%, and 97.83%, respectively. In addition, SVM produces accuracy, precision, and recall 99.98%, 95.33%, 95.09%, respectively. From the comparison of the two classification methods, DNN outperformed SVM for data classification and author identification.

Keywords : *Digital Library, Bibliographic Data, Author Name Disambiguation, Homonym, Synonym, Deep Neural Network, Support Vector Machine*

PENERAPAN PENGUKUR KESAMAAN ATRIBUT JUDUL BERBASIS STRING PADA DATA BIBLIOGRAFI UNTUK MENINGKATKAN KINERJA KLASIFIKASI KESAMAAN PENULIS

Mohammad El Qiliqsandy (09011281722064)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya

Email : elqiliqsandy@gmail.com

ABSTRAK

Author Name Disambiguation (AND) merupakan permasalahan ambiguitas nama terhadap suatu publikasi pada database *Digital Library* (DL) yang disebabkan karena kondisi *Homonymity* dan *Synonymity* nama penulis (*author*). Metode yang diusulkan dalam adalah *Deep Neural Network* (DNN) dan *Support Vector Machine* (SVM). Dataset bibliografi yang digunakan adalah *Dataset DBLP Labeled Data* oleh Jinseok Kim, dkk. Penelitian yang dilakukan berfokus dalam pengolahan data dengan berbagai macam teknik *Normalization* dan *Transformation Data* untuk menciptakan suatu fitur yang efektif untuk digunakan dalam klasifikasi. Parameter utama penelitian yang dilakukan adalah *accuracy*, *precision*, dan *recall* yang merupakan parameter penting untuk mengetahui tingkat keberhasilan metode yang dilakukan dalam mengatasi permasalahan AND khususnya identifikasi *author*. Klasifikasi DNN menghasilkan nilai *accuracy* 99,98%, *precision* 97,71%, dan *recall* 97,83%. Klasifikasi SVM menghasilkan nilai *accuracy* 99,98%, *precision* 95,33%, dan *recall* 95,09%. DNN menjadi metode terbaik untuk melakukan klasifikasi data dan identifikasi *author*.

Kata Kunci : Digital Library, Bibliographic Data, Author Name Disambiguation, Homonym, Synonym, Deep Neural Network, Support Vector Machine

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

BAB 1
PENDAHULUAN

1.1. Latar Belakang

Author Name Disambiguation (AND) adalah permasalahan yang terjadi ketika terdapat satu *author* yang sama tetapi muncul dengan nama yang berbeda pada paper lain yang dipublikasikannya (sinonim) atau dua *author* yang berbeda tetapi memiliki nama yang sama (homonim)[1]. Menurut penelitian [2], AND sedikit sulit dipahami karena berbeda dari kasus *disambiguation* biasa pada bidang *natural language processing* lainnya dimana untuk membedakan nama penulis diperlukannya data *meta*, yaitu data di luar konteks tulisan. Hal ini terjadi karena dalam sebuah karya tulis, nama atau identitas penulis biasanya tidak tercantum di dalam tulisannya. Metadata yang dimaksud meliputi nama penulis, nama-nama rekan penulis, judul tulisan, tahun terbit, dan berbagai macam data lain tergantung dari organisasi yang menyimpan data bibliografi tersebut seperti DBLP , PubMed, dan MEDLINE[3]–[6].

AND memiliki dua kasus utama sebagai standar penyelesaiannya, yaitu penelitian untuk membedakan dua orang berbeda yang namanya mirip atau sama, kemudian seorang individu yang penulisan namanya berbeda. *First name* atau *last name* seorang penulis bisa sama dengan orang lain, seperti yang terjadi pada data yang dimiliki *database* MEDLINE dengan 2/3 penulis memiliki nama yang sama dengan penulis lain[6]. Hal ini bisa merusak beberapa bagian dari sistem data bibliografi seperti sistem pencarian dan sistem skor atau index seorang individu yang terbagi ke lebih dari satu akun[7].

Pada penelitian-penelitian yang pernah dilakukan sebelumnya, telah digunakan berbagai metode menggunakan *machine learning* maupun mengaplikasikan algoritma *non machine learning*. Penelitian terbaru [8], [9] mengarah ke pembuatan dataset karena jumlah data yang semakin membesar membuat penelitian-penelitian yang menggunakan dataset berbeda sulit dibandingkan.

Untuk menyelesaikan masalah AND, penelitian terdahulu telah menggunakan metode seperti *Graph-based*[10], [11] dan *Supervised Machine Learning*[12], [13]. Prapemrosesan dataset juga diteliti untuk memperbagus model sebelum diklasifikasi, seperti dataset *Initial* dari [14]. Penelitian ini berfokus pada bagian prapemrosesan dataset dimana sebelum masuk ke model *machine learning*, data teks harus diolah menjadi angka. Untuk mengubah data teks menjadi angka, beberapa penelitian menggunakan berbagai metode untuk mencari nilai kemiripan atau *similarity*. Nilai kemiripan bisa dicari dengan berbagai macam algoritma [15]. Penelitian ini akan menggunakan algoritma *string-based similarity* atau metode kemiripan teks berbasis *string* dimana metode yang digunakan yaitu *Jaro-Winkler*, *Jaccard* dan *Cosine*.

1.2 Tujuan dan Manfaat

1.2.1 Tujuan

Tujuan dari Tugas Akhir ini, yaitu:

1. Meningkatkan akurasi *Author Matching* pada kasus *Author Name Disambiguation* (AND) menggunakan metode kemiripan teks berbasis *string* dan *Deep Neural Network* (DNN).

2. Mencari metode kemiripan teks terbaik untuk menyelesaikan masalah *Author Name Disambiguation* (AND) pada kasus *Author Matching*.

1.2.2 Manfaat

Manfaat dari penulisan Tugas Akhir ini, yaitu:

1. Dapat memberikan informasi jelas mengenai kesamaan penulis (*Author Matching*) berdasarkan metode-metode kemiripan teks berbasis *string*.

2. Dapat memberikan referensi jelas untuk penelitian mengenai *Author Matching* di bidang *Author Name Disambiuation* kedepannya.

1.3 Perumusan Masalah

Bagaimana memilih metode kemiripan teks terbaik untuk digunakan pada permasalahan *Author Name Disambiguation* (AND) terutama pada kasus kesamaan penulis (*author matching*) menggunakan metode *classifier Deep Neural Network* (DNN), *Random Forest* dan *Support Vector Machine* (SVM) untuk mendapatkan kinerja terbaik dari hasil penelitian yang dilakukan.

1.4 Batasan Masalah

Batasan masalah pada penelitian tugas akhir ini adalah sebagai berikut:

1. Penelitian dilakukan terhadap kasus *author matching* di bidang *Author Name Disambiguation* (AND).
2. Penelitian dilakukan menggunakan bahasa pemrograman *Python*.
3. Dataset yang digunakan dibuat oleh peneliti sebelumnya, Jinseok Kim et al [16]. Dataset ini merupakan dataset bibliografi bersumber dari organisasi DBLP dan data yang diambil sudah bersih.
4. Penelitian ini menggunakan metode kemiripan teks berbasis *string* yaitu *jaro-winkler*, *jaccard*, dan *cosine* dengan *classifier* DNN, SVM dan *Random Forest*. Performa dari semua metode yang digunakan dibandingkan untuk menentukan cara terbaik dalam mengatasi permasalahan kasus *author matching*.
5. Penelitian ini menghasilkan nilai-nilai yang dapat mengukur performa metode yang digunakan berupa nilai Akurasi, Presisi, *Spesifisity*, *Recall*, F1-*Score* dan *Error-Rate*, serta persentase kebenaran dalam kasus sinonim, homonim dan non sinonim homonim.

1.5 Metodologi Penelitian

Pada Tugas Akhir ini, metodologi yang digunakan dalam melakukan penelitian adalah sebagai berikut:

1.5.1 Metode Studi Pustaka dan Literatur

Sebelum memulai penelitian, penulis terlebih dahulu melakukan studi pustaka mengenai metode kemiripan teks, klasifikasi *Author Name Disambiguation* (AND) dan kemajuan yang telah dikembangkan oleh peneliti-peneliti sebelumnya, sehingga penulis bisa memahami secara penuh apa yang akan diteliti.

1.5.2 Metode Konsultasi

Pada bagian ini, penulis terlebih dahulu berkonsultasi dengan para narasumber yang mengetahui materi dan berwawasan di bidang permasalahan AND yang dikerjakan dalam Tugas Akhir ini.

1.5.3 Metode Pembuatan Model

Penulis membuat program yang membantu perancangan model pembelajaran mesin untuk menyelesaikan permasalahan AND. Program tersebut merupakan sebuah instruksi yang apabila diberi dataset, akan mengubah masukan menjadi keluaran.

1.5.4. Metode Pengujian dan Validasi

Model yang telah dibuat selanjutnya diuji untuk mengetahui kecacatan atau kekurangan yang bisa dibenarkan untuk meningkatkan akurasi dari model klasifikasi.

1.5.5 Metode Hasil dan Analisa

Hasil akhir dari pengerjaan tugas akhir dianalisis kelebihan dan kekurangannya agar dijadikan informasi yang mampu membantu penelitian-penelitian kedepannya.

1.5.6 Metode Penarikan Kesimpulan dan Saran

Metode terakhir yaitu mengambil kesimpulan dengan merangkum secara ringkas, padat dan jelas dengan maksud memberikan informasi secara terstruktur dan baik, serta untuk memberi saran dan ide untuk penelitian kedepannya.

1.6 Sistematika Penelitian

Untuk memperjelas isi-isi dari tugas akhir, disusunlah sistematika penulisan seperti di bawah ini:

BAB I PENDAHULUAN

Bab ini digunakan untuk memberitahu secara rinci dan jelas perihal latar belakang, tujuan, manfaat, perumusan dan batasan masalah, serta metodologi dan sistematika penelitian.

BAB II TINJAUAN PUSTAKA

Bab ini menjelaskan dasar teori dan materi-materi mengenai masalah dan penyelesaian dari penelitian yang digunakan dalam pembuatan tugas akhir ini.

BAB III METODOLOGI PENELITIAN

Metodologi yang digunakan dibahas secara rinci perihal teknik, metode, dan alur proses yang dilakukan dalam penelitian.

BAB IV HASIL DAN ANALISIS

Bagian berikutnya menjelaskan hasil dari penelitan yang telah dikerjakan termasuk kehebatan dan kekurangan dari metode yang digunakan.

BAB V KESIMPULAN

Bab terakhir ini memberitahu hasil penelitian secara ringkas, padat dan jelas ,disertakan saran untuk penelitian selanjutnya khususnya tentang Tugas Akhir yang dikerjakan.

DAFTAR PUSTAKA

[1] J. Wu and X. H. Ding, "Author name disambiguation in scientific collaboration and mobility cases," *Scientometrics*, vol. 96, no. 3, pp. 683–697, 2013, doi: 10.1007/s11192-013-0978-8.

[2] J. Huang, S. Ertekin, and C. L. Giles, "Efficient name disambiguation for large-scale databases," Springer Verlag, 2006. doi: 10.1007/11871637_53.

[3] M. Ley, "The DBLP computer science bibliography: Evolution, research issues, perspectives," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2002, vol. 2476, pp. 1–10, doi: 10.1007/3-540-45735-6_1.

[4] W. Liu *et al.*, "Author name disambiguation for PubMed," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 4, pp. 765–781, Apr. 2014, doi: 10.1002/asi.23063.

[5] M. Song, E. H. J. Kim, and H. J. Kim, "Exploring author name disambiguation on PubMed-scale," *J. Informetr.*, vol. 9, no. 4, pp. 924–941, 2015, doi: 10.1016/j.joi.2015.08.004.

[6] V. I. Torvik and N. R. Smalheiser, "Author name disambiguation in MEDLINE," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 3, Jul. 2009, doi: 10.1145/1552303.1552304.

[7] J. Gomide, H. Kling, and D. Figueiredo, "Name usage pattern in the sinonim ambiguity problem in bibliographic data," *Scientometrics*, vol. 112, no. 2, pp. 747–766, Aug. 2017, doi: 10.1007/s11192-017-2410-2.

[8] S. Subramanian, D. King, D. Downey, and S. Feldman, "S2AND: A Benchmark and Evaluation System for Author Name Disambiguation," 2021, [Online]. Available: http://arxiv.org/abs/2103.07534.

[9] L. Zhang, W. Lu, and J. Yang, "LAGOS-AND: A Large, Gold Standard Dataset for Scholarly Author Name Disambiguation,"معرفت ادیان , vol. 4, no. 3, pp. 57–71, Apr. 2021, [Online]. Available: http://marefateadyan.nashriyat.ir/node/150.

[10] Y. Ma, Y. Wu, and C. Lu, "A graph-based author name disambiguation method and analysis via information theory," *Entropy*, vol. 22, no. 4, pp. 1–17, 2020, doi: 10.3390/E22040416.

[11] K. M. Pooja, S. Mondal, and J. Chandra, “A Graph Combination With Edge Pruning-Based Approach for Author Name Disambiguation,” *J. Assoc. Inf. Sci. Technol.*, vol. 71, no. 1, pp. 69–83, 2020, doi: 10.1002/asi.24212.

[12] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, “Two supervised learning approaches for name disambiguation in author citations,” 2004. doi: 10.1145/996350.996419.

[13] J. Kim, J. Kim, and J. Owen-Smith, “Ethnicity-based name partitioning for author name disambiguation using supervised machine learning,” *J. Assoc. Inf. Sci. Technol.*, no. January, pp. 1–16, 2021, doi: 10.1002/asi.24459.

[14] S. Milojević, “Accuracy of simple, initials-based methods for author name disambiguation,” *J. Informetr.*, vol. 7, no. 4, pp. 767–773, Oct. 2013, doi: 10.1016/j.joi.2013.06.006.

[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, Apr. 2013, doi: 10.5120/11638-7118.

[16] J. Kim, “Evaluating author name disambiguation for digital libraries: a case of DBLP,” *Scientometrics*, vol. 116, no. 3, pp. 1867–1886, Sep. 2018, doi: 10.1007/s11192-018-2824-5.

[17] I. Hussain and S. Asghar, “Incremental author name disambiguation using author profile models and self-citations,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 27, no. 5, pp. 3665–3681, 2019, doi: 10.3906/elk-1806-132.

[18] A. P. de Carvalho, A. A. Ferreira, A. H. F. Laender, and M. A. Gonçalves, “Incremental Unsupervised Name Disambiguation in Cleaned Digital Libraries,” *J. Inf. Data Manag.*, vol. 2, no. 573871, p. 289, 2011.

[19] L. V. B. Esperidião *et al.*, “Reducing Fragmentation in Incremental Author Name Disambiguation,” *Jidm*, vol. 5, no. 3, pp. 293–307, 2014.

[20] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender, “A brief survey of automatic methods for author name disambiguation,” *SIGMOD Record*, vol. 41, no. 2. pp. 15–26, Aug. 2012, doi: 10.1145/2350036.2350040.

[21] H. Han, H. Zha, and C. L. Giles, “Name disambiguation in author citations using a K-way spectral clustering method,” 2005, doi: 10.1145/1065385.1065462.

[22] I. Hussain and S. Asghar, “A survey of author name disambiguation techniques: 2010-2016,” *Knowl. Eng. Rev.*, vol. 32, pp. 1–24, Dec. 2017, doi: 10.1017/S0269888917000182.

[23] D. Zhang, J. Tang, J. Li, and K. Wang, “A constraint-based probabilistic framework for name disambiguation,” in *International Conference on Information and Knowledge Management, Proceedings*, 2007, pp. 1019–1022, doi: 10.1145/1321440.1321600.

[24] H. Han, H. Zha, W. Xu, and C. L. Giles, “A hierarchical naive bayes mixture model for name disambiguation in author citations,” in *Proceedings of the ACM Symposium on Applied Computing*, 2005, vol. 2, no. 1, pp. 1065–1069, doi: 10.1145/1066677.1066920.

[25] C. A. D’Angelo, C. Giuffrida, and G. Abramo, “A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 2, pp. 257–269, Feb. 2011, doi: 10.1002/asi.21460.

[26] M. J. Denny and A. Spirling, “Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It,” *Polit. Anal.*, vol. 26, no. 2, pp. 168–189, 2018, doi: 10.1017/pan.2017.44.

[27] T. Wolf *et al.*, “Transformers: State-of-the-art natural language processing,” 2019. doi: 10.18653/v1/2020.emnlp-demos.6.

[28] E. Haddi, X. Liu, and Y. Shi, “The Role of Text Pre-processing in Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[29] A. I. Kadhim, “An Evaluation of Preprocessing Techniques for Text Classification,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 6, pp. 22–32, 2018, [Online]. Available: https://sites.google.com/site/ijcsis/.

[30] Z. Yao and C. Ze-Wen, “Research on the construction and filter method of stop-word list in text preprocessing,” *Proc. - 4th Int. Conf. Intell. Comput. Technol. Autom. ICICTA 2011*, vol. 1, pp. 217–221, 2011, doi: 10.1109/ICICTA.2011.64.

[31] J. Singh and V. Gupta, *A systematic review of text stemming techniques*, vol. 48, no. 2. Springer Netherlands, 2017.

[32] 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, doi: 10.26555/ijain.v4i1.152.

[33] William W Cohen, Pradeep Ravikumar, and Stephen, "A Comparison of String Distance Metrics for Name-Matching Tasks William," *Proc. IJCAI-2003 Work.*, pp. 73--78, 2003.

[34] I. Atoum and A. Otoom, "Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 9, 2016, doi: 10.14569/ijacsa.2016.070917.

[35] W. W. Yaoshu Wang, Jianbin Qin, "Efficient Approximate Entity Matching Using Jaro-Winkler Distance," in *Web Information Systems Engineering – WISE 2017*, 2017, pp. 231–239.

[36] J. M. Keil, "Efficient bounded Jaro-winkler similarity based search," *Lect. Notes Informatics (LNI), Proc. - Ser. Gesellschaft fur Inform.*, vol. P-289, pp. 205–214, 2019, doi: 10.18420/btw2019-13.

[37] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," *Lect. Notes Eng. Comput. Sci.*, vol. 2202, no. March, pp. 380–384, 2013.

[38] V. Thada and V. Jaglan, "Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm," *Int. J. Innov. Eng. Technol.*, vol. 2, no. 4, pp. 202–205, 2013.

[39] B. B. Tirkey and B. S. Saini, *Proposing model for recognizing user position*, vol. 1045. 2020.

[40] D. Harlianto, S. Mardiyati, D. Lestari, A. H. Zili, and S. Devila, "Indonesia tuberculosis morbidity rate forecasting using recurrent neural network," *AIP Conf. Proc.*, vol. 2242, no. June, 2020, doi: 10.1063/5.0010445.

[41] S. Prof and J. Basilio, "Predicting revenue generation in an online retail website using machine learning algorithm in Data Analytics Annadurai Srinivasan National College of Ireland."

[42] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv*, pp. 1–12,

2017.

[43] 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, doi: 10.1016/j.neucom.2016.12.038.

[44] D. A. Bashar, "Survey on Evolving Deep Learning Neural Network Architectures," *J. Artif. Intell. Capsul. Networks*, vol. 2019, no. 2, pp. 73–82, 2019, doi: 10.36548/jaicn.2019.2.003.

[45] A. Bhandare, M. Bhide, P. Gokhale, and R. Chandavarkar, "Applications of Convolutional Neural Networks," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 5, pp. 2206–2215, 2016, [Online]. Available: http://ijcsit.com/docs/Volume 7/vol7issue5/ijcsit20160705014.pdf.

[46] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018, doi: 10.1016/j.patcog.2017.10.013.

[47] S. Nurmaini, P. R. Umi, R. M. Naufal, and A. Gani, "Cardiac arrhythmias classification using Deep Neural Networks and principle component analysis algorithm," *Int. J. Adv. Soft Comput. its Appl.*, vol. 10, no. 2, pp. 14–32, 2018.

[48] T. I. O. A. NUGRAHA and F. Firdaus, "Klasifikasi Author Pada Data Bibliografi Menggunakan Deep Neural Network Dan Support Vector Machine," 2019.

[49] D. Mao and J. R. Edwards, *Simulation of chemically-reacting gas-solid flowfields using a preconditioning strategy*. 2003.

[50] A. Ukil and I. Systems, *Power Systems Abhisek Ukil Intelligent Systems and Signal Processing in Power Engineering*. .

[51] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," *IEEE Access*, vol. 6, no. c, pp. 33789–33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

[52] A. Patle and D. S. Chouhan, "SVM kernel functions for classification," *2013 Int. Conf. Adv. Technol. Eng. ICATE 2013*, 2013, doi: 10.1109/ICAdTE.2013.6524743.

[53] V. F. Rodriguez-Galiano, M. Chica-Olmo, F. Abarca-Hernandez, P. M.

Atkinson, and C. Jeganathan, “Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture,” *Remote Sens. Environ.*, vol. 121, pp. 93–107, 2012, doi: 10.1016/j.rse.2011.12.003.

[54] A. T. Azar, H. I. Elshazly, A. E. Hassanien, and A. M. Elkorany, “A random forest classifier for lymph diseases,” *Comput. Methods Programs Biomed.*, vol. 113, no. 2, pp. 465–473, 2014, doi: 10.1016/j.cmpb.2013.11.004.

[55] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[56] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, “Random forest classification of multisource remote sensing and geographic data,” *Int. Geosci. Remote Sens. Symp.*, vol. 2, no. C, pp. 1049–1052, 2004, doi: 10.1109/igarss.2004.1368591.

[57] 2018) (Al Amrani, Lazaar, El Kadiri., “Random Forest and Support Vector Machine based Hybrid Approach to SA --RF.pdf.” .

[58] 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, doi: 10.1007/s11192-018-2865-9.