KLASIFIKASI *AUTHOR MATCHING* PADA DATA BIBLIOGRAFI MENGGUNAKAN METODE *COST-SENSITIVE DEEP NEURAL NETWORK* (CSDNN)

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat

Memperoleh Gelar Sarjana Komputer

OLEH :

SUCI DWI LESTARI

09011181722021

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2021

HALAMAN PENGESAHAN

KLASIFIKASI *AUTHOR MATCHING* PADA DATA BIBLIOGRAFI MENGGUNAKAN METODE *COST-SENSITIVE DEEP NEURAL NETWORK* (CSDNN)

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh :

SUCI DWI LESTARI
09011181722021

Palembang, Juni 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 : Senin

Tanggal : 21 Juni 2021

Tim Penguji :

1. Ketua **: Rossi Passarella, M.Eng.**

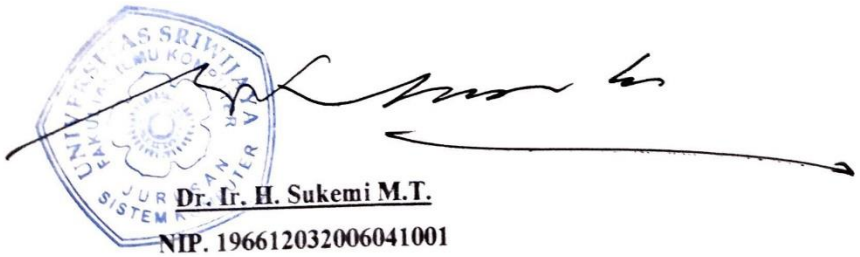
2. Sekretaris **: Rendyansyah, M.T.**

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

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

Mengetahui,

Ketua Jurusan Sistem Komputer



Dr. Ir. H. Sukemi M.T.

NIP. 196612032006041001

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama : Suci Dwi Lestari

NIM : 09011181722021

Judul : Klasifikasi *Author Matching* pada Data Bibliografi Menggunakan Metode *Cost-Sensitive Deep Neural Network* (CSDNN)

Hasil pengecekan *Software Turnitin* : 12%

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, 5 Juli 2021

Suci Dwi Lestari

KATA PENGANTAR

Bismillahirrahmanirrahim.

Assalamu'alaikum Warahmatullahi Wabarakatuh. Puji dan syukur penulis panjatkan kehadirat Allah SWT, karena berkat karunia dan rahmat-Nya sehingga penulis dapat menyelesaikan Tugas Akhir ini dengan judul "Klasifikasi *Author Matching* Pada Data Bibliografi Menggunakan Metode *Cost-Sensitive Deep Neural Network* (CSDNN)*".*

Pada penyusunan Tugas Akhir ini, tidak terlepas dari bantuan, bimbingan serta dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis mengucapkan rasa syukur dan terimakasih kepada yang terhormat :

1. Allah SWT, yang telah memberikan rahmat dan karunia-Nya sehingga penulisan Tugas Akhir ini dapat berjalan dengan lancar.
2. Orang tua saya yang saya cintai, serta adik laki-laki dan kakak perempuan saya, yang selalu memberikan semangat dan do'a, serta dukungan baik financial maupun dukungan lainnya.
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 Kompuer Fakutas Ilmu Komputer Universitas Sriwijaya
5. Bapak Firdaus, M.Kom. selaku Pembimbing Tugas Akhir Penulis yang telah meluangkan waktu nya serta memberikan bimbingan dan arahan kepada penulis sehingga penulis dapat menyelesaikan Tugas Akhir ini.
6. Bapak Rossi Passarella, S.T.,M.Eng, selaku Dosen Pembimbing Akademik di Jurusan Sistem Komputer.
7. Ibu Prof. Dr. Ir. Siti Nurmaini, M.T. selaku Head of Intelligent System Research Group (ISysRG) yang telah menerima penulis menjadi bagian dari team research group sehingga penulis dapat menyelesaikan Tugas Akhir.
8. Bapak Dr. Erwin, M.Si., selaku penguji Sidang Tugas Akhir penulis yang berkenan meluangkan waktunya guna menguji, memberi arahan, serta nasihat untuk penulis.

9. Kak Naufal Rachmatullah, S.Kom., M.T., Mbak Ade Irian Safitri, M.Kom. dan Mbak Annisa Darmawahyuni, M.Kom. yang telah banyak membantu penulis dan suportif memberikan semangat serta arahan kepada penulis sehingga dapat menyelesaikan Tugas Akhir ini dengan baik.
10. Team ISysRG Batch 2 terutama team teks processing (Qiliq, Annisa, Azis, Irvan, Wais, dan Jorgi) yang selalu saling menyemangati satu sama lain dan saling membantu agar dapat menyelesaikan Tugas Akhir ini.
11. AAPS (Alna, Annisa, Putri), Circle Bigbox (Ghina, Jannes, Farhan, Chay, Juno, Jorgi), dan Lia yang selalu ada dan turut membantu penulis dalam segala hal terutama penyelesaian Tugas Akhir.
12. Jurusan Sistem Komputer Reguler kelas A angkatan 2017 yang tidak dapat saya sebutkan satu persatu.

Penulis juga berterima kasih kepada semua pihak yang terlibat, baik secara langsung ataupun tidak langsung dalam penyelesaian Tugas Akhir ini. Tentunya dalam pembuatan Tugas Akhir ini, masih terdapat beberapa kekurangan dan kesalahan yang mungkin terjadi. Oleh karena itu sebagai bahan perbaikan kedepan penulis tentunya mengharapkan koreksi, saran, serta masukan terhadap isi dari Tugas Akhir ini. Akhir kata, semoga dengan Tugas Akhir ini, akan menjadi tambahan ilmu dan pengembangan wawasan kita terhadap klasifikasi author matching dan dapat menjadi bahan referensi terhadap mahasiswa yang memerlukan.

Palembang, Juli 2021

Suci Dwi Lestari

AUTHOR CLASSIFICATION IN BIBLIOGRAPHIC DATA USING COST-SENSITIVE DEEP NEURAL NETWORK (CSDNN)

Suci Dwi Lestari (09011181722021)

Computer Engineering Department, Computer Science Faculty, Universitas Sriwijaya

Email : sucidl27@gmail.com

Abstract

Author Name Ambiguity is an issue that occurs when publication records contain ambiguous or ambiguous author names, i.e. the same author may appear under different names, or different authors may have similar names. The method proposed in this research is Cost-Sensitive Deep Neural Network (CSDNN). The bibliographic dataset used is the DBLP Dataset by Jinseok Kim, et al. This research focuses on the use of classification methods, namely CSDNN and DNN. The main parameters of the research carried out are accuracy, precision, specificity, recall, and error rate, which are important parameters to determine the success rate of the method used in overcoming problems AND in particular finding the similarities of the authors. The CSDNN classification resulted achieves accuracy, precision, specificity, recall, and error rate which is 99.94%, 96.60%, 99.97%, 96.90%, and 0.000515. DNN classification resulted achieves accuracy, precision, specificity, recall, and error rate which is 99.94%, 99.97%, 96.90%, 96.78%, and 0.000501.

Keywords : *Author Matching, Digital Library, Bibliographic Data, Author Name Disambiguation, Cost-Sensitive Deep Neural Network, Deep Neural Network.*

KLASIFIKASI AUTHOR MATCHING PADA DATA BIBLIOGRAFI MENGGUNAKAN METODE COST-SENSITIVE DEEP NEURAL NETWORK (CSDNN)

Suci Dwi Lestari (09011181722021)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya

Email : sucidl27@gmail.com

Abstrak

Author Name Ambiguity merupakan suatu masalah yang terjadi apabila sekumpulan catatan publikasi berisi nama pengarang yang rancu atau ambigu, yaitu pengarang yang sama mungkin muncul dibawah nama yang berbeda, atau pengarang yang berbeda mungkin memiliki nama yang mirip. Metode yang diusulkan pada penelitian ini adalah *Cost-Sensitive Deep Neural Network* (CSDNN). Dataset bibliografi yang digunakan adalah *Dataset DBLP* oleh Jinseok Kim, dkk. Penelitian yang dilakukan berfokus dalam penggunaan metode klasifikasi yaitu CSDNN dan DNN. Parameter utama penelitian yang dilakukan adalah *accuracy*, *precision*, *spesifisity*, *recall*, dan *error rate* yang merupakan parameter penting untuk mengetahui tingkat keberhasilan metode yang dilakukan dalam mengatasi permasalahan AND khususnya identifikasi kesamaan *author*. Klasifikasi CSDNN menghasilkan nilai *accuracy* 99,94%, *precision* 96,60%, *spesifisity* 99,97%, *recall* 96,90%, dan *error rate* 0,000515. Klasifikasi DNN menghasilkan nilai *accuracy* 99,94%, *precision* 99,97%, *spesifisity* 96,90%, *recall* 96,78%, dan *error rate* 0,000501.

Kata Kunci : *Author Matching, Digital Library, Bibliographic Data, Author Name Disambiguation, Cost-Sensitive Deep Neural Network, Deep Neural Network*.

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

BAB I
PENDAHULUAN

1.1. Latar Belakang

Informasi penulis merupakan ukuran yang sangat diperlukan database bibliografi, baik pada pengambilan informasi maupun untuk bibliometrik. Ambiguitas nama penulis merupakan masalah umum dalam kedua kasus, satu nama dapat mewakili beberapa penulis dan seorang penulis dapat menerbitkan atau diindeks dengan beberapa nama [1]. Proses sederhana seperti *attribution* (mengesampingkan karya-karya yang diterbitkan secara anonim), adalah masalah utama yang belum terpecahkan dalam ilmu informasi. Sehingga, perlu untuk menganalisi metadata, terkadang teks, suatu karya tulis untuk membuat dugaan yang tepat tentang identitas penulis sebenarnya [2].

Author Name Ambiguity merupakan suatu masalah yang terjadi apabila sekumpulan catatan publikasi berisi nama pengarang yang rancu atau ambigu, yaitu pengarang yang sama mungkin muncul dibawah nama yang berbeda, atau pengarang yang berbeda mungkin memiliki nama yang mirip. Masalah ini menurunkan kualitas dan keandalan informasi yang diambil dari perpustakaan digital seperti *impact* penulis, *impact* organisasi, dan lain sebagainya. Oleh karena itu, *author name disambiguation* adalah tugas penting dalam perpustakaan digital. Terdapat dua pendekatan untuk *Author Name Disambiguation*, mengelompokkan catatan publikasi dari penulis yang sama dengan menemukan beberapa kesamaan diantara penulis (metode *author grouping*) atau secara langsung *assigning* ke penulis masing-masing (metode *author assignment*). Kedua metode tersebut akan mencoba membuat, memilih dan menggabungkan fitur berdasarkan kesamaan atribut (nama penulis, kata kunci, dan lainnya.) Dengan menggunakan beberapa *measure* seperti Jaccard, Jaro, dan lainnya, atau beberapa heuristik [3][4]. Pada

penelitian yang pernah dilakukan sebelumnya mengenai *author* matching, data dengan skala yang cukup besar, akan diproses menggunakan teknik berpasangan (*pairwise*) dengan melakukan kombinasi pada tiap atribut dataset dan data akan diberikan label 0 jika bukan merupakan author yang sama, label 1 jika merupakan author yang sama. Hal tersebut, menyebabkan ketidakseimbangan data (data *imbalance*), pada kondisinya, label 0 menjadi 98,9% dari jumlah data dibandingkan dengan, label 1 hanya sebesar 1,1%. Pada penelitian tersebut, digunakan metode deteksi anomali yang menggunakan algoritma *Isolation Forest*. Hasil yang didapatkan dari penelitan dengan akurasi yang cukup baik, yaitu mencapai akurasi 99,5% [5].

Data *imbalance* dapat menyebabkan kesalahan yang tidak terduga dan bahkan konsekuensi yang serius dalam analisis data, terutama dalam klasifikasi. Hal ini, dikarenakan distribusi kelas yang lebih cenderung menuntut algoritma klasifikasi menjadi bias ke kelas mayoritas. Oleh karena itu, konsep kelas minoritas tidak dipelajari secara memadai [6]. Pada kasus kelas *binary*, ketika jumlah *instances* di satu kelas secara signifikan melebihi jumlah *instances* di kelas lain. Oleh karena itu, konsep kelas minoritas tidak dipelajari secara memadai. Situasi ini menjadi kendala ketika mencoba untuk mengidentifikasi kelas minoritas, karena algoritma pembelajaran biasanya tidak disesuaikan dengan karakteristik tersebut [7]. Metode untuk menangani kelas *imbalance* dalam *machine learning* dapat dikelompokkan menjadi tiga kategori yaitu, teknik level data, metode *algorithm level*, dan *hybrid approaches*. Teknik level data akan mengurangi tingkat *imbalance* melalu berbagai metode pengambilan sampel data. Metode *algorithm level* akan menangani kelas *imbalance*, biasanya diterapkan dengan skema *weight* atau *cost*, termasuk memodifikasi *learner* yang mendasari atau *output* untuk mengurangi bias terhadap kelompok mayoritas. Terakhir, *hybrid approaches* secara strategis akan menggabungkan metode sampling dan metode algoritmik [8][9]. Maka metode pada penelitian tersebut yang berguna untuk menangani kelas *imbalance* di dataset digunakan lah metode *algorithm level*. Penggunaan algoritma *deep neural network*, seperti diketahui jika *deep neural network* telah banyak berhasil dalam berbagai bidang [10], seperti *computer vision* [11], *speech recognition* [12], serta *natural language processing* [13]. Serta menerapkan *cost-sensitive learning* seperti

diketahui berbagai metode *cost-sensitive learning* [14], telah dikembangan untuk menangani kesalahan dalam klasifikasi, *cost-sensitive learning* juga telah dianggap sebagai solusi yang baik untuk pembelajaran kelas *imbalance* [15]. Spesifiknya, metode yang akan digunakan ialah *cost sensitive deep neural network*. Terdapat penelitian sebelumya yang menggunakan metode *cost sensitive deep neural network*, yaitu mengenai prediksi pada catatan kesehatan elektronik pasien kemungkinan apakah pasien akan diterima kembali di rumah sakit tempat pasien di rawat. Penelitian tersebut mendapatkan hasil kinerja yang cukup baik untuk menjamin uji klinis aktual di rumah sakit [16].

Berdasarkan hal tersebut, akan dilakukan penelitian mengenai Klasifikasi *Author Matching* pada Data Bibliografi dengan Metode *Cost Sensitive Deep Neural Network* (CSDNN).

1.2. Tujuan dan Manfaat

1.2.1. Tujuan

Berikut tujuan nya antara lain :

1. Dapat melakukan penyelesaian pada kasus *Author Name Disambiguation* (AND) yaitu *case* penyamaan penulis (*author matching*) metode nya *Cost Sensitive Deep Neural Network* (CSDNN) dan *Deep Neural Network* (DNN).
2. Dapat melakukan penentuan teknik (metode) dan pendekatan terbaik penyelesaian permasalahan *Author Name Disambiguation* (AND) pada *case* penyamaan penulis (*author matching*).

1.2.2. Manfaat

Manfaat nya antara lain :

1. Membantu pemecahan permasalahan *Author Name Disambiguation* (AND) khususnya penyamaan penulis (*author matching*) dengan penggunaan metode *Cost Sensitive Deep Neural Network* (CSDNN).

2. Dapat dijadikan untuk bahan referensi pada penelitian berikutnya tentang *Author Name Disambiguation* (AND) pada studi kasus penyamaan penulis (*author matching*) dan dapat dikembangkan lebih lanjut.

1.3. Perumusan dan Batasan Masalah

1.3.1. Perumusan Masalah

Bagaimana memilih metode dan pendekatan terbaik untuk merampungkan permasalahan *Author Name Disambiguation* (AND) lebih tepatnya dalam studi kasus penyamaan penulis (*author matching*) di data bibliografi dengan penggunaan metode *Cost Sensitive Deep Neural Network* (CSDNN) untuk mendapatkan *output* yang akurat.

1.3.2. Batasan Masalah

Berikut batasan masalah Tugas Akhir ini, adalah :

1. Cakupan penelitian hanya pada permasalahan *Author Name Disambiguation* (AND) lebih tepatnya penyamaan penulis (*author matching*).

2. Basis bahasa pemrograman yang digunakan untuk penelitian adalah *Python*.

3. Bahan dataset yang digunakan untuk penelitian adalah *Dataset DBLP Labeled Data* yang merupakan hasil dari penelitian oleh Jinseok Kim et al. [17] dataset tersebut berasal dari website *dblp.org* dan telah melalui proses pembersihan (*Cleaning Process*).

4. Penelitian menggunakan metode CSDNN dan DNN, yang akan dilakukan perbandingan agar dapat ditentukan pembelajaran mesin (*Machine Learning*) terbaik pada permasalahan penyamaan penulis.

5. Hasil penelitian terbatas pada nilai Sensitivitas, spesifisitas, Presisi, *F1-Score*, *Error Rate,* dan Akurasi, serta persentase kebenaran guna untuk tolak ukur tingkat untuk kesamaan author.

1.4. Metodologi Penelitian

Metodologi penelitian untuk Tugas Akhir ini, adalah sebagai berikut :

1.4.1 Metode Studi Pustaka dan Literatur

Metode ini, akan melakukan proses mencari serta mengumpulkan berbagai rujukan berbentuk pustaka-pustaka baik yang ada di buku maupun internet tentang klasifikasi *Author Name Disambiguation* (AND) dengan menggunakan *Cost Sensitive Deep Neural Network* (CSDNN) dan *Deep Neural Network* (DNN).

1.4.2 Metode Konsultasi

Metode ini, dilakukan konsultasi oleh penulis baik secara langsung maupun tidak langsung kepada semua pihak narasumber yang memiliki pengetahuan dan wawasan yang baik yang nanti nya dapat membantu penulis dalam mengatasi permasalahan penulisan Tugas Akhir ini mengenai klasifikasi *Author Name Disambiguation* (AND) dengan menggunakan *Cost Sensitive Deep Neural Network* (CSDNN) dan *Deep Neural Network* (DNN).

1.4.3 Metode Pembuatan Model

Metode ini, akan dulakukan proses merancang pembuatan pemodelan dengan menggunakan program.

1.4.4 Metode Pengujian dan Validasi

Metode ini, akan menguji sistem yang telah dirancang sebelum untuk mengetahui batasan-batasan performa sistem apakah dapat mendapatkan hasil yang terbaik ataupun kebalikannya.

1.4.5 Metode Hasil dan Analisa

Metode ini, selanjutnya setelah pengujian maka akan dianalisa keseluruhan nya baik itu keunggulan maupun kelemahannya, dengan harapan menjadi bahan referensi pada penelitian selanjutnya.

1.4.6 Metode Penarikan Kesimpulan dan Saran

Pada metode ini, hasil dan analisa yang didapatkan dapat diambil kesimpulan dan saran untuk penelitian selanjutnya. Metode ini merupakan tahap akhir dari metodologi penelitian.

1.5. Sistematika Penulisan

Sistematika penulisan untuk Tugas Akhir ini adalah seperti berikut :

BAB I – PENDAHULUAN

Bab ini berisi mengenai Latar Belakang Masalah, Tujuan dan Manfaat, Perumusan Masalah dan Batasan Masalah, Metode Penelitian, dan Sistematika Penulisan dari penelitian yang dilakukan.

BAB II – TINJAUAN PUSTAKA

Bab ini berisi mengenai Dasar Teori, Konsep, dan Prinsip Dasar yang dibutuhkan untuk memecahkan masalah dalam penelitian yang dilakukan.

BAB III – METODOLOGI

Bab ini membahas secara spesifik mengenai teknik, metode, dan alur proses penelitian Tugas Akhir.

BAB IV – HASIL DAN PEMBAHASAN

Bab ini berisi mengenai hasil pengujian dan analisis yang diperoleh dari penelitian serta membahas hasil yang telah dicapai meliputi kelebihan dan kekurangan dari penelitian yang telah dilakukan.

BAB V – KESIMPULAN DAN SARAN

Bab ini berisi mengenai simpulan mengenai hasil penelitian yang dilakukan beserta saran untuk penelitian selanjutnya tepatnya mengenai *Author Matching*.

DAFTAR PUSTAKA

[1] A. Strotmann and D. Zhao, "Author name disambiguation: What difference does it make in author-based citation analysis?," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 9, pp. 1820–1833, 2012.

[2] N. R. Smalheiser and V. I. Torvik, "Author name disambiguation," *Annu. Rev. Inf. Sci. Technol.*, vol. 43, no. 1, p. 1, 2009.

[3] 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.

[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] Z. Yamani, S. Nurmaini, and D. P. Rini, "Author Matching Classification with Anomaly Detection Approach for Bibliomethric Repository Data," *Comput. Eng. Appl. J.*, vol. 9, no. 2, pp. 79–92, 2020.

[6] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in *2016 international joint conference on neural networks (IJCNN)*, 2016, pp. 4368–4374.

[7] V. López, A. Fernández, J. G. Moreno-Torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics," *Expert Syst. Appl.*, vol. 39, no. 7, pp. 6585–6608, 2012.

[8] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, 2019.

[9] B. Krawczyk, "Learning from imbalanced data: open challenges and future

directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.

[10] L. Zhang, T. Luo, F. Zhang, and Y. Wu, "A recommendation model based on deep neural network," *IEEE Access*, vol. 6, pp. 9454–9463, 2018.

[11] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE conference on computer vision and pattern recognition*, 2012, pp. 3642–3649.

[12] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1671–1675, 2015.

[13] E. Arisoy, T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Deep neural network language models," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, 2012, pp. 20–28.

[14] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, 2005.

[15] X.-Y. Liu and Z.-H. Zhou, "The influence of class imbalance on cost-sensitive learning: An empirical study," in *Sixth International Conference on Data Mining (ICDM'06)*, 2006, pp. 970–974.

[16] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. Ben Abdallah, and A. Kronzer, "Predicting hospital readmission via cost-sensitive deep learning," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 6, pp. 1968–1978, 2018.

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

[18] H. Jin, L. Huang, and P. Yuan, "Name disambiguation using semantic association clustering," in *2009 IEEE International Conference on e-*

Business Engineering, 2009, pp. 42–48.

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

[20] M. Imran, S. Z. H. Gillani, and M. Marchese, "A real-time heuristic-based unsupervised method for name disambiguation in digital libraries," *D-Lib Mag.*, vol. 19, no. 9, p. 1, 2013.

[21] C. Schulz, A. Mazloumian, A. M. Petersen, O. Penner, and D. Helbing, "Exploiting citation networks for large-scale author name disambiguation," *EPJ Data Sci.*, vol. 3, no. 1, p. 11, 2014.

[22] R. G. Cota, A. A. Ferreira, C. Nascimento, M. A. Gonçalves, and A. H. F. Laender, "An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 9, pp. 1853–1870, 2010.

[23] M.-C. Müller, "Semantic author name disambiguation with word embeddings," in *International Conference on Theory and Practice of Digital Libraries*, 2017, pp. 300–311.

[24] 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.

[25] W. W. Cohen, P. Ravikumar, S. E. Fienberg, and others, "A Comparison of String Distance Metrics for Name-Matching Tasks.," in *IIWeb*, 2003, vol. 2003, pp. 73–78.

[26] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 104–112, 2014.

[27] C. Zhang, T. Baldwin, H. Ho, B. Kimelfeld, and Y. Li, "Adaptive parser-centric text normalization," in *Proceedings of the 51st Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 1159–1168.

[28] J. L. Neto, A. A. Freitas, and C. A. A. Kaestner, "Automatic text summarization using a machine learning approach," in *Brazilian symposium on artificial intelligence*, 2002, pp. 205–215.

[29] S. Vijayarani, R. Janani, and others, "Text mining: open source tokenization tools-an analysis," *Adv. Comput. Intell. An Int. J.*, vol. 3, no. 1, pp. 37–47, 2016.

[30] V. Jha, N. Manjunath, P. D. Shenoy, and K. R. Venugopal, "Hsra: Hindi stopword removal algorithm," in *2016 international conference on microelectronics, computing and communications (MicroCom)*, 2016, pp. 1–5.

[31] J. Singh and V. Gupta, "Text stemming: Approaches, applications, and challenges," *ACM Comput. Surv.*, vol. 49, no. 3, pp. 1–46, 2016.

[32] 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.

[33] 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.

[34] J. Ye, "Cosine similarity measures for intuitionistic fuzzy sets and their applications," *Math. Comput. Model.*, vol. 53, no. 1–2, pp. 91–97, 2011.

[35] 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.

[36] 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.

[37] K. Mikawa, T. Ishida, and M. Goto, "A proposal of extended cosine measure for distance metric learning in text classification," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*, 2011, pp. 1741–1746.

[38] E. Zangerle, W. Gassler, and G. Specht, "On the impact of text similarity functions on hashtag recommendations in microblogging environments," *Soc. Netw. Anal. Min.*, vol. 3, no. 4, pp. 889–898, 2013.

[39] X. Lin, J. Zhu, Y. Tang, F. Yang, B. Peng, and W. Li, "A novel approach for author name disambiguation using ranking confidence," in *International Conference on Database Systems for Advanced Applications*, 2017, pp. 169–182.

[40] K. Kim, S. Rohatgi, and C. L. Giles, "Hybrid deep pairwise classification for author name disambiguation," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2369–2372.

[41] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Mach. Learn.*, vol. 42, no. 1, pp. 143–175, 2001.

[42] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in *Workshop on artificial intelligence for web search (AAAI 2000)*, 2000, vol. 58, p. 64.

[43] D. Khongorzul, S.-M. Lee, and M.-H. Kim, "OrdinalEncoder based DNN for Natural Gas Leak Prediction," *J. Korea Converg. Soc.*, vol. 10, no. 10, pp. 7–13, 2019.

[44] W. Fu and T. Menzies, "Easy over hard: A case study on deep learning," in *Proceedings of the 2017 11th joint meeting on foundations of software engineering*, 2017, pp. 49–60.

[45] P. Tapkan, L. Özbak\ir, S. Kulluk, and A. Baykaso\uglu, “A cost-sensitive classification algorithm: BEE-Miner,” *Knowledge-Based Syst.*, vol. 95, pp. 99–113, 2016.

[46] O. Gencoglu, T. Virtanen, and H. Huttunen, “Recognition of acoustic events using deep neural networks,” in *2014 22nd European signal processing conference (EUSIPCO)*, 2014, pp. 506–510.

[47] S. Nurmaini, A. Gani, and others, “Cardiac Arrhythmias Classification Using Deep Neural Networks and Principle Component Analysis Algorithm.,” *Int. J. Adv. Soft Comput. Its Appl.*, vol. 10, no. 2, 2018.

[48] M. Shah and R. Kapdi, “Object detection using deep neural networks,” in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2017, pp. 787–790.

[49] A. Bashar, “Survey on evolving deep learning neural network architectures,” *J. Artif. Intell.*, vol. 1, no. 02, pp. 73–82, 2019.

[50] N. Kriegeskorte and T. Golan, “Neural network models and deep learning,” *Curr. Biol.*, vol. 29, no. 7, pp. R231--R236, 2019.

[51] J. Sokolić, R. Giryes, G. Sapiro, and M. R. D. Rodrigues, “Robust large margin deep neural networks,” *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4265–4280, 2017.

[52] T. I. O. A. NUGRAHA and F. Firdaus, “KLASIFIKASI AUTHOR PADA DATA BIBLIOGRAFI MENGGUNAKAN DEEP NEURAL NETWORK DAN SUPPORT VECTOR MACHINE,” Sriwijaya University, 2019.

[53] 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.