

**PERANCANGAN MODEL DEEP NEURAL NETWORK
UNTUK KLASIFIKASI AUTHOR PADA DATA
PUBLIKASI INDONESIA**

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat

Memperoleh Gelar Sarjana Komputer



OLEH :

IRVAN FAHREZA

09011281722032

**JURUSAN SISTEM KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA
2021**

HALAMAN PENGESAHAN

**PERANCANGAN MODEL DEEP NEURAL NETWORK
UNTUK KLASIFIKASI AUTHOR
PADA DATA PUBLIKASI INDONESIA**

TUGAS AKHIR

**Program Studi Sistem Komputer
Jenjang S1**

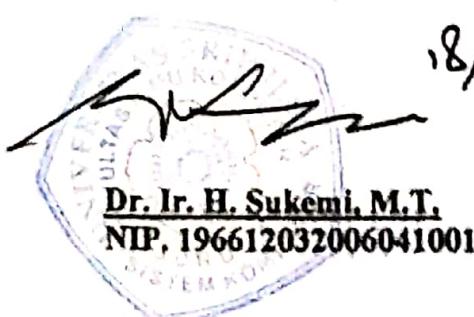
Oleh

**IRVAN FAHREZA
09011281722032**

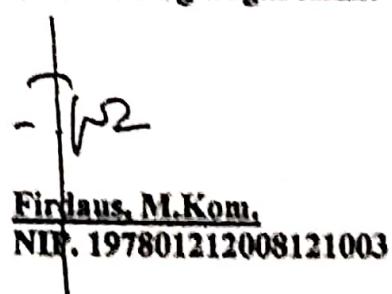
Indralaya, 29 Juli 2021

Mengetahui,

Ketua Jurusan Sistem Komputer



Pembimbing Tugas Akhir



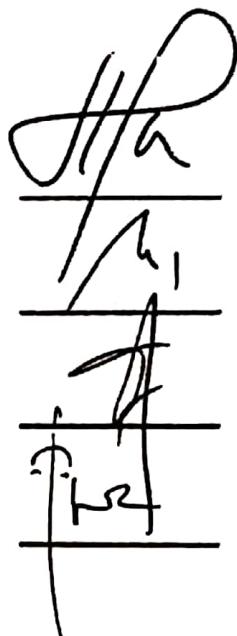
HALAMAN PERSETUJUAN

Telah diuji dan lulus pada:

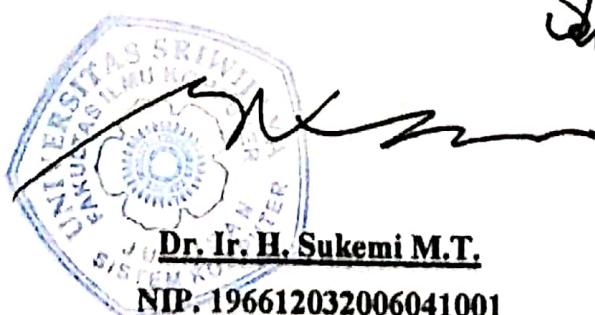
Hari : Rabu
Tanggal : 28 Juli 2021

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



HALAMAN PERNYATAAN

Yang bertanda tangan dibawah ini:

Nama : Irvan Fahreza
NIM : 090112811722032
Judul : Perancangan Model *Deep Neural Network*
Untuk Klasifikasi *Author* Pada Data Publikasi Indonesia

Hasil Pengecekan Software *iThenticate/Turnitin* : 10%

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, 29 Juli 2021



Irvan Fahreza

090112811722032

KATA PENGANTAR

Assalamu'alaikum Wr. Wb.

Alhamdulillahirabbil'alamin, segala puji dan syukur atas kehadirat Allah SWT. atas berkah, rahmat, dan hidayahNya yang senantiasa dilimpahkan kepada penulis, sehingga penulis dapat menyelesaikan Tugas Akhir ini yang berjudul **"Perancangan Model Deep Neural Network Untuk Klasifikasi Author Pada Data Publikasi Indonesia"**.

Dalam tugas akhir ini penulis menjelaskan mengenai perancangan model untuk melakukan klasifikasi dan identifikasi author pada suatu set publikasi digital dengan disertai data-data yang diperoleh penulis saat penelitian. Penulis berharap agar isi Tugas Akhir ini dapat bermanfaat bagi orang banyak serta menjadi topik bacaan yang tertarik untuk meneliti masalah *Author Name Disambiguation* (AND).

Dalam penyusunan Tugas Akhir ini banyak hambatan serta rintangan yang penulis hadapi, namun pada akhirnya dapat melaluinya berkat adanya bimbingan dan bantuan dari berbagai pihak, baik secara moral maupun spiritual. Untuk itu pada kesempatan ini penulis menyampaikan terimakasih kepada:

1. Kedua orang tua tercinta atas segala do'a, motivasi, serta dukungan baik moril, materil, dan spiritual.
2. Yth, Bapak Jaidan Jauhari, S.Pd., M.T., selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
3. Yth, Bapak Dr. Ir. H. Sukemi, M.T., selaku Ketua Jurusan Sistem Komputer Fakultas Ilmu Komputer Universitas Sriwijaya, dan selaku Pembimbing Akademik Jurusan Sistem Komputer.
4. Yth, Ibu Prof. Dr. Ir. Siti Nurmaini, M.T., selaku pembimbing Akademik Jurusan Sistem Komputer.

5. Bapak Firdaus, M.Kom., selaku Pembimbing Tugas Akhir yang telah berkenan meluangkan waktunya guna membimbing, memberikan saran dan motivasi serta bimbingan terbaik untuk penulis dalam menyelesaikan Tugas Akhir ini.
6. Kepada teman-teman tim penelitian ISysRG khususnya bidang Teks yaitu Qiliq, Azis, Suci, Annisa, Wais, dan Jorgi yang telah membantu dan mendukung penulis dalam menyelesaikan tugas akhir ini.
7. Kak Naufal, mba Ade, dan mba Annisa sebagai mentor dalam menyelesaikan tugas akhir.
8. Tama, Sultan, Jo, serta Iqbale sebagai teman yang telah banyak membantu sejak awal perkuliahan hingga saat ini.
9. Luthfi dan Ardi sebagai partner kuliah sejak awal perkuliahan hingga saat ini.
10. Teman-teman seperjuangan Sistem Komputer Angkatan 2017 serta semua pihak yang tidak dapat penulis sebutkan satu-persatu.

Penulis menyadari bahwa Tugas Akhir ini masih sangat jauh dari kata sempurna. Untuk itu kritik dan saran yang membangun sangatlah diharapkan penulis agar dapat segera diperbaiki sehingga Tugas Akhir ini dapat dijadikan sebagai masukkan ide dan pemikiran yang bermanfaat bagi semua pihak.

Wassalamu'alaikum Wr. Wb.

Palembang, Juli 2021

Irvan Fahreza

***DESIGNING OF DEEP NEURAL NETWORK MODEL
FOR AUTHOR CLASSIFICATION
IN INDONESIAN PUBLICATION DATA***

IRVAN FAHREZA (09011281722032)

*Computer Engineering Department, Computer Science Faculty, Sriwijaya
University*

Email : irvanfahreza45@gmail.com

Abstract

Author Name Disambiguation (AND) is a problem that occurs when a set of publications contains ambiguous names of authors, i.e. the same author may appear with different names (synonyms) in other published papers, or authors who may be different who may have the same name (homonym). In the final project, we will design a model with a Deep Neural Network (DNN). The dataset used in this final project uses primary data sourced from the Scopus website. This research focuses on integrating data from Indonesian authors. Parameters accuracy, sensitivity, and precision are standard benchmarks to determine the performance of the methods used to solve AND problems. The best DNN classification model achieves Accuracy 99.9936%, Sensitivity 93.1433%, Precision 94.3733%. Then for the highest performance measurement, the Non Synonym-Homonym case has 99.9967% Accuracy, 96.7388% Sensitivity, and 97.5102% Precision.

Keywords : *Author Name Disambiguation, Synonym, Homonym, Bibliographic Data, Deep Neural Network.*

**PERANCANGAN MODEL DEEP NEURAL NETWORK
UNTUK KLASIFIKASI AUTHOR
PADA DATA PUBLIKASI INDONESIA**

IRVAN FAHREZA (09011281722032)

Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya
Email : irvanfahreza45@gmail.com

Abstrak

Author Name Disambiguation (AND) adalah masalah yang terjadi ketika satu set publikasi berisi nama-nama *authors* (penulis) yang ambigu, yaitu *author* (penulis) yang sama mungkin saja muncul dengan nama yang berbeda (sinonim) dalam paper lain yang dipublikasinya, atau *author* (penulis) yang mungkin berbeda yang mungkin memiliki nama yang sama (homonim). Dalam tugas akhir akan merancang model dengan *classifier Deep Neural Network* (DNN). Dataset yang digunakan pada tugas akhir ini menggunakan data primer yang bersumber dari *website scopus*. Penelitian yang dilakukan berfokus pada pengintegrasian data *author* Indonesia. Parameter *accuracy*, *sensitivity* dan *precision* merupakan standar tolak ukur untuk mengetahui performa dari metode yang digunakan untuk mengatasi permasalahan AND. Model klasifikasi DNN terbaik mencapai Akurasi 99.9936%, Sensitivitas 93.1433%, Presisi 94.3733%. Kemudian untuk pengukuran kinerja tertinggi, kasus *Non Synonym-Homonym* memiliki *Accuracy* 99,9967%, *Sensitivity* 96,7388%, dan *Precision* 97,5102%.

Kata Kunci : *Author Name Disambiguation, Synonym, Homonym, Bibliographic Data, Deep Neural Network.*

DAFTAR ISI

	Halaman
HALAMAN PENGESAHAN	ii
HALAMAN PERSETUJUAN.....	iii
HALAMAN PERNYATAAN.....	iv
KATA PENGANTAR	v
ABSTRACT	vii
ABSTRAK	viii
DAFTAR ISI.....	ix
DAFTAR GAMBAR.....	xii
DAFTAR TABEL.....	xiv
DAFTAR LAMPIRAN	xv
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang.....	1
1.2. Tujuan dan Manfaat.....	3
1.2.1. Tujuan	3
1.2.2. Manfaat	3
1.3. Perumusan dan Batasan Masalah	3
1.3.1. Perumusan Masalah	3
1.3.2. Batasan Masalah.....	4
1.4. Metodologi Penelitian	4
1.4.1. Metode Literatur dan Studi Pustaka.....	4
1.4.2. Metode Konsultasi	4
1.4.3. Metode Penentuan Model	5

1.4.4. Metode Pengujian dan Validasi	5
1.4.5. Metode Hasil dan Analisa	5
1.4.6. Metode Penarikan Kesimpulan dan Saran	5
1.5. Sistematika Penulisan.....	5
BAB II TINJAUAN PUSTAKA.....	7
2.1. Author Name Disambiguation.....	7
2.1.1. Faktor Masalah AND	7
2.2. Normalisasi Text	8
2.2.1. <i>Stemming</i>	8
2.2.2. <i>Lemmatization</i>	9
2.2.3. <i>Tokenization</i>	10
2.2.4. <i>Case Folding</i>	11
2.3. Ekstraksi Fitur	12
2.3.1. <i>One Hot Encoder</i>	12
2.3.2. <i>Term Frequency - Inverse Document Frequency</i>	13
2.4. Normalisasi Data	14
2.4.1. <i>MinMax Scaler</i>	14
2.5. Reduksi Fitur	15
2.5.1. <i>Principal Component Analysis</i>	15
2.6. Klasifikasi.....	16
2.6.1. <i>Deep Neural Network</i>	16
2.7. <i>Performance Measurement</i>	19
BAB III METODOLOGI	21
3.1. Pendahuluan	21
3.2. Kerangka Kerja.....	21
3.3. Akuisisi Data	22

3.3.1. Komposisi Data.....	25
3.4. Pra-Pemrosesan Data.....	27
3.4.1. Pemrosesan Fitur.....	28
3.4.2. Penggabungan dan Reduksi fitur	33
3.5. Tuning.....	34
3.6. Klasifikasi.....	34
3.6.1. Klasifikasi DNN	34
3.7. Evaluasi Model.....	38
BAB IV HASIL DAN PEMBAHASAN.....	42
4.1. Hasil Akuisisi Data.....	42
4.2. Hasil Persiapan Data.....	42
4.3. Hasil Pra-pemrosesan Data.....	44
4.4. Hasil Klasifikasii	45
4.4.1. Hasil Klasifikasii <i>Deep Neural Network</i> (DNN)	45
BAB V KESIMPULAN DAN SARAN	54
5.1. Kesimpulan.....	54
5.2. Saran	54
DAFTAR PUSTAKA	55

DAFTAR GAMBAR

	Halaman
Gambar 2.1 Contoh <i>Stemming</i>	9
Gambar 2.2 Contoh Penerapan <i>Lemmatization</i>	10
Gambar 2.3 Contoh Penerapan <i>Tokenization</i>	11
Gambar 2.4 Contoh Penerapan <i>Case Folding</i>	11
Gambar 2.5 Penggunaan OHE.....	12
Gambar 2.6 Struktur <i>Deep Neural Network</i> (DNN).....	18
Gambar 3.1 Kerangka Kerja	21
Gambar 3.2 Akuisisi Data	23
Gambar 3.3 Tahap Pra-pemrosesan Data	27
Gambar 3.4 Flowchart Pra-pemrosesan Data	28
Gambar 3.5 Flowchart dengan Teknik OHE	29
Gambar 3.6 Flowchart fitur <i>Venue</i>	30
Gambar 3.7 Flowchart fitur <i>Year</i>	31
Gambar 3.8 Flowchart fitur <i>Co-Author(s)</i>	32
Gambar 3.9 Flowchart fitur <i>Title</i>	33
Gambar 3.10 Arsitektur DNN	35
Gambar 4.1 <i>Piie Chart</i> Komposisi Data.....	43
Gambar 4.2 Grafik terbaik Model <i>Accuracy</i> klasifikasi DNN	48
Gambar 4.3 Grafik terbaik Model <i>Loss</i> klasifikasi DNN	48
Gambar 4.4 Grafiik <i>Performance Measurement</i> model DNN	50
Gambar 4.5 <i>Confusion Matrix</i> Kasus <i>Synonym</i>	51
Gambar 4.6 <i>Confusion Matrix</i> Kasus <i>Homonym</i>	52

Gambar 4.7 *Confusion Matrix Kasus Synonym-Homonym* 52

Gambar 4.8 *Confusion Matrix Kasus Non Synonym-Homonym* 53

DAFTAR TABEL

	Halaman
Tabel 1. Contoh paper berisi nama ambigu.....	1
Tabel 2. Perhitungan TF-IDF	14
Tabel 3. Contoh Labeling.....	25
Tabel 4. Detail Tunning DNN	36
Tabel 5. Tabel Kebenaran <i>Confusion Matrix</i>	38s
Tabel 6. Deskripsi Data	42
Tabel 7. Komposisi Data	43
Tabel 8. Detail Fitur	44
Tabel 9. Hasil Klasifikasi DNN.....	45
Tabel 10. <i>Performance Measurement</i> DNN.....	49

DAFTAR LAMPIRAN

Lampiran 1. Form Perbaikan

Lampiran 2. Cek Plagiat

BAB I

PENDAHULUAN

1.1. Latar Belakang

Author Name Disambiguation adalah masalah yang terjadi ketika satu set publikasi berisi nama-nama *authors* (penulis) yang ambigu, yaitu *author* (penulis) yang sama mungkin saja muncul dengan nama yang berbeda (sinonim) dalam paper lain yang dipublikasikan, atau *author* (penulis) yang berbeda tetapi memiliki nama yang sama (homonim) [1].

Tabel 1.

Contoh paper berisi nama ambigu .

Paper No.1	<p>High corolla color variation of <i>Hoya coronaria</i> Blume in Belitung Island: Potential use and conservation. IOP Conference Series: Earth and Environmental Science.</p> <p>Authors: Rahayu S., Fakhurrozi Y. Research Center for Plant Conservation and Botanic Gardens-Bogor Botanic Gardens, Indonesian Institut of Sciences (LIPI), Jl. Ir. H. Juanda 13, Bogor, 16122, Indonesia</p>	<p>Mrs. Rahayu worked at Research Center for Plant Conservation and Botanic Gardens-Bogor Botanic Gardens, Indonesian Institut of Sciences (LIPI), Jl. Ir. H. Juanda 13, Bogor, 16122,</p>
Paper No.2	<p>The effect of socioscientific issues embedded in explanation-driven inquiry (EDI) learning model on high school students' conceptual understanding of reaction rate. AIP Conference Proceedings.</p> <p>Authors: Wahyuni E.S., Rahayu S., Yahmin. Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Negeri Malang, Jl Semarang 5, Malang, East Java, 65145, Indonesia</p>	<p>Mrs. Rahayu worked at Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Negeri Malang, Jl Semarang 5, Malang, East Java, 65145, Indonesia</p>
Paper No.3	<p>Characterization, identification, and analysis of bioactive compound of endophytic bacteria from <i>hoya multiflora</i> blume. Biodiversitas.</p> <p>Authors: Alvionita D.N., Rahayu S.R.I., Mubarik N.R. Research Center for Plant Conservation Botanic Gardens, Indonesian Institute of Sciences, Jl. Ir. Juanda 13, Bogor, 16122, Indonesia</p>	<p>Mrs. Rahayu worked at Research Center for Plant Conservation and Botanic Gardens-Bogor Botanic Gardens, Indonesian Institut of Sciences (LIPI), Jl. Ir. H. Juanda 13, Bogor, 16122,</p>

Tabel 1 menunjukkan contoh tentang ambiguitas *author* Indonesia dalam 3 publikasi yang berbeda. Nama *author* pada paper 1 dan paper 3 adalah contoh sinonim. Kedua paper tersebut merujuk pada ‘Rahayu S’ dari Indonesian Institut of

Sciences (LIPI). Sedangkan nama *author* pada paper 1 dan paper 2 merupakan contoh homonim dimana ‘Rahayu S’ pada paper 1 mengacu pada ‘Rahayu S’ dari Mrs. Rahayu worked at Indonesian Institut of Sciences (LIPI) dan ‘Rahayu S’ pada paper 2 mengacu pada ‘Rahayu S’ dari Universitas Negeri Malang.

Ambiguitas nama dalam konteks ini adalah masalah yang sangat kritis yang menarik banyak perhatian riset perpustakaan digital. Terutama, *author* dengan nama Indonesia adalah kasus yang ambigu menjadi salah satu tantangan untuk masalah ini. Oleh karena itu, fokus utama terletak pada disambiguasi *Author* Indonesia untuk mengintegrasikan data bibliografi.

Author Name Disambiguation (AND) berkait dengan pengelompokan *authors* (penulis) dengan nama yang sama ke dalam individu yang berbeda. Untuk mengatasi hal tersebut, sudah banyak penelitian yang telah menggunakan fitur *disambiguation* seperti nama *co-authorsnya*, judul paper atau publikasi, topik artikel, email/affiliasi, dll. Diantaranya, *co-authorship* adalah fitur yang paling berpengaruh, karena hubungan terdekat dengan *co-authorship* dapat membedakan identitas *authors* lebih jelas daripada fitur yang lainnya [2]. *Author Name Disambiguation* (AND) adalah tugas yang menantang bagi para sarjana yang menambah informasi bibliografi untuk pengetahuan ilmiah. Pendekatan konstruktif untuk menyelesaikan ambiguitas nama adalah dengan menggunakan algoritma komputer untuk mengidentifikasi nama penulis. Beberapa metode disambiguasi berbasis algoritma telah dikembangkan oleh data *scientist* [3].

Masalah ambiguitas nama penulis terkait erat dengan bidang penelitian lain seperti disambiguasi entitas [4]–[9], name disambiguation [1], [10], [11], varian nama [12], nama alias [13], dan arsitektur nama global [14]. Umumnya, ambiguitas nama penulis dapat diselesaikan menggunakan atribut yang berbeda seperti *Title*, *affiliation*, *co-authors*, *keywords*, *abstract*, *venue* dan tahun publikasi [10], [15]–[17].

1.2. Tujuan dan Manfaat

1.2.1. Tujuan

Tujuan dari penulisan Tugas Akhir ini, yaitu :

1. Dapat menyelesaikan permasalahan dalam *Author Name Disambiguation* (AND) menggunakan klasifikasi *Deep Neural Network* (DNN).
2. Mendapatkan model terbaik untuk penyelesaian permasalahan dalam *Author Name Disambiguations* (AND).

1.2.2. Manfaat

Manfaat yang didapatkan dari penulisan Tugas Akhir ini, yaitu :

1. Membantu menyelesaikan permasalahan dalam *Author Name Disambiguations* (AND) menggunakan *Classifier Deep Neural Network* (DNN).
2. Hasil dari tugas akhir ini dapat digunakan untuk referensi apabila ditemui masalah *Author Name Disambiguations* (AND).

1.3. Perumusan dan Batasan Masalah

1.3.1. Perumusan Masalah

Menentukan model dan metode terbaik untuk penyelesaian masalah *Author Name Disambiguations* (AND) pada data publikasi Indonesia menggunakan *Classifier Deep Neural Network* (DNN) untuk menghasilkan model terbaik.

1.3.2. Batasan Masalah

Berikut merupakan batasan masalah yang terdapat pada Tugas Akhir ini :

1. Penelitian ini termasuk kedalam masalah *Author Name Disambiguation* (AND).
2. Dataset yang digunakan pada penelitian ini merupakan dataset primer bersumber dari *scopus*.
3. *Python* merupakan bahasa pemrograman yang digunakan dalam penelitian ini.
4. Hasil pengukuran pada penelitian ini berdasarkan nilai *Specificity*, *Sensitivity*, *Accuracy*, *Precision*, *Error Rate*, dan *F1-Score* sebagai tolak ukur performa terhadap kecocokan dari *authors* dengan label.

1.4. Metodologi Penelitian

Metodologi yang digunakan pada tugas akhir adalah :

1.4.1. Metode Literatur dan Studi Pustaka

Metode pengumpulan referensi yaitu dengan membuat literatur review yang terdapat pada internet dan beberapa sumber lainnya yang dapat dipertanggung jawabkan sumbernya yang berhubungan dengan tema *Author Name Disambiguation* (AND) yang dikerjakan penulis.

1.4.2. Metode Konsultasi

Penulis melakukan konsultasi ke orang-orang yang mampu atau memiliki pengetahuan yang baik dalam penulis dalam Tugas Akhir tentang masalah *Author Name Disambiguation* (AND) dengan menggunakan *Classifier Deep Neural Network* (DNN).

1.4.3. Metode Penentuan Model

Metode merupakan metode dimana modelakan ditentukan dan dirancang menggunakan program *python* berdasarkan tema Tugas Akhir yang dikerjakan.

1.4.4. Metode Pengujian dan Validasi

Melakukan pengujian dan validasi guna melihat batasan kinerja sistem agar mendapatkan hasil metode yang terbaik.

1.4.5. Metode Hasil dan Analisa

Hasil dan Analisa Tugas Akhir ini akan dianalisa kekurangan dan kelebihannya, sehingga kedepannya jika terdapat masalah yang sama seperti yang dikerjakan penulis, Tugas Akhir ini dapat menjadi referensi untuk penelitian selanjutnya.

1.4.6. Metode Penarikan Kesimpulan dan Saran

Berdasarkan Hasil dan Analisa, setelah itu akan ditarik kesimpulan dan saran untuk penelitian tentang masalah ini, agar untuk penelitian selanjutnya dapat menghasilkan *output* yang lebih baik lagi.

1.5. Sistematika Penulisan

Dalam mempermudah penyusunan Tugas Akhir ini dan juga membuat isi dari setiap bab yang ada pada Tugas Akhir ini lebih jelas, maka dibuat sistematika penulisan sebagai berikut :

BAB I – PENDAHULUAN

BAB I ini berisikan Latar Belakang, Tujuan dan Manfaat, serta Metodologi penelitian yang dilakukan.

BAB II – TINJAUAN PUSTAKA

Tinjauan Pustaka berisikan Dasar Teori dan Konsep Dasar yang dipakai serta dibutuhkan dalam memecahkan masalah yang ada pada penelitian ini.

BAB III – METODOLOGI

Bab ini membahas secara rinci tentang metode, teknik, serta alur proses yang dilakukan.

BAB IV – HASIL DAN PEMBAHASAN

Pada BAB akan menjelaskan berupa hasil pengujian metode serta analisis dari penelitian dan melakukan pembahasan dari hasil yang telah didapatkan dengan cara menganalisis kekurangan dan kelebihan dari penelitian yang sudah dilakukan.

BAB V – KESIMPULAN DAN SARAN

Pada bab terakhir ini merupakan kesimpulan berdasarkan hasil penelitian serta saran untuk penelitian selanjutnya agar dapat mendapatkan metode yang lebih baik khususnya tentang Penelitian yang sudah dikerjakan.

DAFTAR PUSTAKA

- [1] A. A. Ferreira, M. A. Gonçalves, and A. H. F. Laender, “A brief survey of automatic methods for author name disambiguation,” *SIGMOD Rec.*, vol. 41, no. 2, pp. 15–26, 2012.
- [2] I. S. Kang *et al.*, “On co-authorship for author disambiguation,” *Inf. Process. Manag.*, 2009.
- [3] Firdaus *et al.*, “Author identification in bibliographic data using deep neural networks,” *Telkomnika (Telecommunication Comput. Electron. Control.)*, vol. 19, no. 3, pp. 911–919, 2021.
- [4] I. Bhattacharya and L. Getoor, “Collective entity resolution in relational data,” *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, pp. 1–35, 2007.
- [5] E. L. Murnane, B. Haslhofer, and C. Lagoze, “RESLVE: Leveraging user interest to improve entity disambiguation on short text,” *WWW 2013 Companion - Proc. 22nd Int. Conf. World Wide Web*, pp. 1275–1283, 2013.
- [6] A. Chisholm and B. Hachey, “Entity Disambiguation with Web Links,” *Trans. Assoc. Comput. Linguist.*, vol. 3, pp. 145–156, 2015.
- [7] S. Oramas, L. Espinosa-Anke, M. Sordo, H. Saggion, and X. Serra, “ELMD: An automatically generated Entity Linking gold standard dataset in the Music Domain,” *Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016*, pp. 3312–3317, 2016.
- [8] A. Krzywicki, W. Wobcke, M. Bain, J. Calvo Martinez, and P. Compton, “Data mining for building knowledge bases: Techniques, architectures and applications,” *Knowl. Eng. Rev.*, vol. 31, no. 2, pp. 97–123, 2016.
- [9] L. Zhu, M. Ghasemi-Gol, P. Szekely, A. Galstyan, and C. A. Knoblock, “Unsupervised entity resolution on multi-type graphs,” *Lect. Notes*

Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9981 LNCS, pp. 649–667, 2016.

- [10] B. W. On, J. Kang, D. Lee, and P. Mitra, “Comparative study of name disambiguation problem using a scalable blocking-based framework,” *Proc. ACM/IEEE Jt. Conf. Digit. Libr.*, pp. 344–353, 2005.
- [11] D. Shin, T. Kim, J. Choi, and J. Kim, “Author name disambiguation using a graph model with node splitting and merging based on bibliographic information,” *Scientometrics*, vol. 100, no. 1, pp. 15–50, 2014.
- [12] G. Louppe, H. T. Al-Natsheh, M. Susik, and E. J. Maguire, “Ethnicity sensitive author disambiguation using semi-supervised learning,” *Commun. Comput. Inf. Sci.*, vol. 649, pp. 272–287, 2016.
- [13] I. Johannes, C. Scholtes, F. Peter, and E. Maes, “System and method for authorship disambiguation and alias resolution in electronic data,” vol. 2, no. 12, 2016.
- [14] R. L. Pyle, “Towards a global names architecture: The future of indexing scientific names,” *Zookeys*, vol. 2016, no. 550, pp. 261–281, 2016.
- [15] A. A. Ferreira, A. Veloso, M. A. Gonçalves, and A. H. F. Laender, “Effective self-training author name disambiguation in scholarly digital libraries,” *Proc. ACM Int. Conf. Digit. Libr.*, pp. 39–48, 2010.
- [16] S. Elliott, “Survey of author name disambiguation: 2004 to 2010,” *Libr. Philos. Pract.*, vol. 2010, no. NOVEMBER, 2010.
- [17] L. V. B. Esperidião *et al.*, “Reducing Fragmentation in Incremental Author Name Disambiguation,” *Jidm*, vol. 5, no. 3, pp. 293–307, 2014.
- [18] H. N. Tran, T. Huynh, and T. Do, “Author name disambiguation by using deep neural network,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8397 LNAI, no. PART 1, pp. 123–132, 2014.

- [19] V. I. Torvik and N. R. Smalheiser, “Author name disambiguation in MEDLINE,” *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 3, 2009.
- [20] M. da S. Conrado., V. A. L. Gutierrez., S. O. Rezende., H. S. Baird, and D. P. Lopresti, *Evaluation of normalization techniques in text classification for portuguese*, vol. 3517. 2005.
- [21] H. Hassan and A. Menezes, “Social text normalization using contextual graph Random Walks,” *ACL 2013 - 51st Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, vol. 1, pp. 1577–1586, 2013.
- [22] M. A. G. Jivani, “A Comparative Study of Stemming Algorithms,” *Oxford Handb. Cult. Psychol.*, vol. 2, no. 6, pp. 1930–1938, 2012.
- [23] R. Alkula, “From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software,” *Inf. Retr. Boston.*, vol. 4, no. 3–4, pp. 195–208, 2001.
- [24] R. Krovetz, “Viewing morphology as an inference process,” *Artif. Intell.*, vol. 118, no. 1–2, pp. 277–294, 2000.
- [25] A. Pirkola, “Morphological typology of languages for IR,” *J. Doc.*, vol. 57, no. 3, pp. 330–348, 2001.
- [26] D. Harman, “How effective is suffixing?,” *J. Am. Soc. Inf. Sci.*, vol. 42, no. 1, pp. 7–15, 1991.
- [27] D. A. Hull, “Stemming algorithms: A case study for detailed evaluation,” *J. Am. Soc. Inf. Sci.*, vol. 47, no. 1, pp. 70–84, 1996.
- [28] K. Tuomo *et al.*, “Authors : Stemming and lemmatization in the clustering of finnish text conference on Information and knowledge management Editors of work : Pages : Stemming and Lemmatization in the Clustering of Finnish Text Documents,” 2004.

- [29] I. Bruzzi and A. Benigni, “Development of a Stemming Algorithm,” *Clin. Exp. Pharmacol. Physiol.*, vol. 23, no. 4, pp. 349–353, 1996.
- [30] M. Rashmi, C. D. Manning, P. Raghavan, and H. Schütze, “Introduction to Information Retrieval Systems,” *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 3, no. 4, pp. 2051–2054, 2015.
- [31] K. Ingason, S. Helgadóttir, H. Loftsson, and E. Rögnvaldsson, *A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI)*, vol. 2718, no. January. 2008.
- [32] V. Balakrishnan and E. Lloyd-yemoh, “Stemming and lemmatization: A comparison of retrieval performances,” pp. 174–179.
- [33] O. Ozturkmenoglu and A. Alpkocak, “Comparison of different lemmatization approaches for information retrieval on Turkish text collection,” *INISTA 2012 - Int. Symp. Innov. Intell. Syst. Appl.*, no. July, 2012.
- [34] T. Verma, R. Renu, and D. Gaur, “Tokenization and Filtering Process in RapidMiner,” *Int. J. Appl. Inf. Syst.*, vol. 7, no. 2, pp. 16–18, 2014.
- [35] B. Habert *et al.*, “Towards Tokenization Evaluation,” *Zeitschrift fuer Met. Res. Adv. Tech.*, vol. 74, no. 4, pp. 233–237, 1983.
- [36] C. Seger, “An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing,” *Degree Proj. Technol.*, p. 41, 2018.
- [37] D. Jiang, W. Lin, and N. Raghavan, “A novel framework for semiconductor manufacturing final test yield classification using machine learning techniques,” *IEEE Access*, vol. 8, pp. 197885–197895, 2020.
- [38] J. Attenberg, K. Weinberger, A. Dasgupta, A. Smola, and M. Zinkevich, “Collaborative email-spam filtering with the hashing-trick,” *6th Conf. Email Anti-Spam, CEAS 2009*, no. May 2014, 2009.

- [39] Y. T. Zhang, L. Gong, and Y. C. Wang, “Improved TF-IDF approach for text classification,” *J. Zhejiang Univ. Sci.*, vol. 6 A, no. 1, pp. 49–55, 2005.
- [40] C. Buckley and G. Salton, “Term weighting approaches in automatic text retrieval.” 1988.
- [41] G. Salton, “Developments in automatic text retrieval,” *Science (80-.)*, vol. 253, no. 5023, pp. 974–980, 1991.
- [42] T. Joachims, “Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.” .
- [43] C. Sitaula, “MANTIC TEXT CLUSTERING USING ENHANCED VECTOR SPACE MODEL USING NEPALI LANGUAGE,” vol. 4, no. 4, pp. 41–46, 2012.
- [44] M. Alhawarat, M. Hegazi, and A. Hilal, “Processing the Text of the Holy Quran: a Text Mining Study,” *Int. J. Adv. Comput. Sci. Appl.*, 2015.
- [45] T. Jayalakshmi and A. Santhakumaran, “Statistical Normalization and Back Propagationfor Classification,” *Int. J. Comput. Theory Eng.*, vol. 3, no. 1, pp. 89–93, 2011.
- [46] A. Pandey and A. Jain, “Comparative Analysis of KNN Algorithm using Various Normalization Techniques,” *Int. J. Comput. Netw. Inf. Secur.*, vol. 9, no. 11, pp. 36–42, 2017.
- [47] K. Pearson, “ LIII. On lines and planes of closest fit to systems of points in space ,” *London, Edinburgh, Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, 1901.
- [48] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, 1933.
- [49] I. Jolliffe, “Principal Component Analysis for Special Types of Data,” *Anal. Princ. Compon. Types, Spec.*, pp. 338–372, 2002.

- [50] F. E. Steffens, “What is principal components analysis?,” *Semin. Princ. components Anal. Atmos. Earth Sci. Pretoria, 1983, (Council Sci. Ind. Res. Pretoria, Natl. Program. Weather. Clim. Atmos. Res. CSIR-S-334)*, vol. 26, no. 3, pp. 3–16, 1983.
- [51] P. J. Phillips *et al.*, “Overview of the face recognition grand challenge,” *Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005*, vol. I, pp. 947–954, 2005.
- [52] S. Asadi, C. D. V. S. Rao, and V. Saikrishna, “A Comparative study of Face Recognition with Principal Component Analysis and Cross-Correlation Technique,” *Int. J. Comput. Appl.*, vol. 10, no. 8, pp. 17–21, 2010.
- [53] S. J. Kwon, “Artificial neural networks,” *Artif. Neural Networks*, pp. 1–426, 2011.
- [54] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Comput.*, vol. 18, no. April, pp. 1527–1554, 2006.
- [55] Y. Lecun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [56] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, “Audio-visual speech recognition using deep learning,” *Appl. Intell.*, vol. 42, no. 4, pp. 722–737, 2015.
- [57] G. Wu, W. Lu, G. Gao, C. Zhao, and J. Liu, “Regional deep learning model for visual tracking,” *Neurocomputing*, vol. 175, no. PartA, pp. 310–323, 2015.
- [58] J. Hu, D. Chen, and J. Du, “State estimation for a class of discrete nonlinear systems with randomly occurring uncertainties and distributed sensor delays,” *Int. J. Gen. Syst.*, vol. 43, no. 3–4, pp. 387–401, 2014.

- [59] J. Hu, Z. Wang, D. Chen, and F. E. Alsaadi, “Estimation, filtering and fusion for networked systems with network-induced phenomena: New progress and prospects,” *Inf. Fusion*, vol. 31, pp. 65–75, 2016.
- [60] J. Hu, Z. Wang, S. Liu, and H. Gao, “A variance-constrained approach to recursive state estimation for time-varying complex networks with missing measurements,” *Automatica*, vol. 64, pp. 155–162, 2016.
- [61] Q. Liu, Z. Wang, X. He, and D. H. Zhou, “Event-Based H_∞ Consensus Control of Multi-Agent Systems with Relative Output Feedback: The Finite-Horizon Case,” *IEEE Trans. Automat. Contr.*, vol. 60, no. 9, pp. 2553–2558, 2015.
- [62] J. Song and Y. Niu, “Resilient finite-time stabilization of fuzzy stochastic systems with randomly occurring uncertainties and randomly occurring gain fluctuations,” *Neurocomputing*, vol. 171, pp. 444–451, 2016.
- [63] H. Yang, Z. Wang, H. Shu, F. E. Alsaadi, and T. Hayat, “Almost sure H_∞ sliding mode control for nonlinear stochastic systems with Markovian switching and time-delays,” *Neurocomputing*, vol. 175, no. PartA, pp. 392–400, 2015.
- [64] Å. Å., “Finite Frequency Property-Based Robust Control Analysis and Synthesis,” no. 2, pp. 3962–3967, 2004.
- [65] N. Zeng, Z. Wang, and H. Zhang, “Inferring nonlinear lateral flow immunoassay state-space models via an unscented Kalman filter,” *Sci. China Inf. Sci.*, vol. 59, no. 11, pp. 1–10, 2016.
- [66] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science (80-.).*, vol. 313, no. 5786, pp. 504–507, 2006.
- [67] N. Hou, H. Dong, Z. Wang, W. Ren, and F. E. Alsaadi, “Non-fragile state estimation for discrete Markovian jumping neural networks,” *Neurocomputing*, vol. 179, pp. 238–245, 2016.

- [68] F. Yang, H. Dong, Z. Wang, W. Ren, and F. E. Alsaadi, “A new approach to non-fragile state estimation for continuous neural networks with time-delays,” *Neurocomputing*, vol. 197, pp. 205–211, 2016.
- [69] Y. Yu, H. Dong, Z. Wang, W. Ren, and F. E. Alsaadi, “Design of non-fragile state estimators for discrete time-delayed neural networks with parameter uncertainties,” *Neurocomputing*, vol. 182, pp. 18–24, 2016.
- [70] Y. Yuan and F. Sun, “Delay-dependent stability criteria for time-varying delay neural networks in the delta domain,” *Neurocomputing*, vol. 125, pp. 17–21, 2014.
- [71] J. Zhang, L. Ma, and Y. Liu, “Passivity analysis for discrete-time neural networks with mixed time-delays and randomly occurring quantization effects,” *Neurocomputing*, vol. 216, pp. 657–665, 2016.
- [72] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” 2012.
- [73] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, “Deep neural networks segment neuronal membranes in electron microscopy images,” *Adv. Neural Inf. Process. Syst.*, vol. 4, pp. 2843–2851, 2012.
- [74] C. Szegedy *et al.*, “Going deeper with convolutions,” *Res. Methods Appl. Settings*, pp. 319–338, 2021.
- [75] D. Cires and U. Meier, “Multi-column Deep Neural Networks for Image Classification,” pp. 3642–3649, 2012.
- [76] R. Collobert, L. Bottou, J. Weston, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural Language Processing (Almost) from Scratch Ronan,” *Proc. - 2017 IEEE 3rd Int. Conf. Collab. Internet Comput. CIC 2017*, vol. 2017-Janua, pp. 328–338, 2017.
- [77] R. Socher *et al.*, “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank,” *Empir. Methods Nat. Lang. Process.*, no.

October, pp. 1631–1642, 2004.

- [78] S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional neural networks for human action recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2013.
- [79] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3361–3368, 2011.
- [80] G. Montavon *et al.*, “Machine learning of molecular electronic properties in chemical compound space,” *New J. Phys.*, vol. 15, pp. 0–16, 2013.
- [81] J. F. Burnham, “Scopus database: A review,” *Biomed. Digit. Libr.*, vol. 3, pp. 1–9, 2006.