PENERAPAN *CORPUS-BASED TEXT SIMILARITY* SEBAGAI PENGUKUR KESAMAAN FITUR PADA DATA BIBLIOGRAFI UNTUK MENINGKATKAN AKURASI KLASIFIKASI KESAMAAN PENULIS

TUGAS AKHIR

Diajukan Untuk Melengkapi Salah Satu Syarat

Memperoleh Gelar Sarjana Komputer

OLEH :

AZIS MULKI RAFANI

09011281722034

JURUSAN SISTEM KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJAYA

2021

HALAMAN PENGESAHAN

PENERAPAN *CORPUS-BASED TEXT SIMILARITY* SEBAGAI PENGUKUR KESAMAAN FITUR PADA DATA BIBLIOGRAFI UNTUK MENINGKATKAN AKURASI KLASIFIKASI KESAMAAN PENULIS

TUGAS AKHIR

Program Studi Sistem Komputer
Jenjang S1

Oleh

AZIS MULKI RAFANI
09011281722034

Indralaya, 29 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 : Senin

Tanggal : 05 Juli 2021

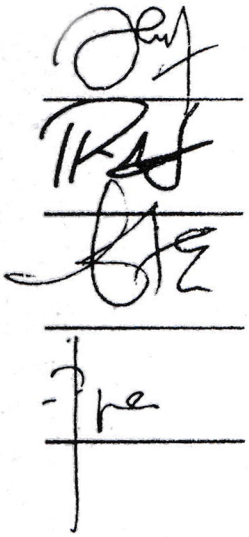
Tim Penguji :

1. Ketua : Ahmad Fali Oklilas, M.T.

2. Sekretaris : Rahmat Fadli Isnanto, S.Si., M.Sc.

3. Penguji : Dr. Ir. Bambang Tutuko, 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 : **Azis Mulki Rafani**
NIM : **090112811722034**
Judul : **Penerapan *Corpus-Based Text Similarity* Sebagai Pengukur Kesamaan Fitur Pada Data Bibliografi Untuk Meningkatkan Akurasi Klasifikasi Kesamaan Penulis**

Hasil Pengecekan Software *iThenticate/Turnitin* : 7%

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

Azis Mulki Rafani
09011281722034

HALAMAN PERSEMBAHAN

وَوَجَدَكَ ضَآلًّا فَهَدٰىۖ

"Dan Dia mendapatimu sebagai seorang yang bingung, lalu Dia memberikan petunjuk” (Q.S. 93:7)

"意志あるところに道はある。"

"Don't stop when you are tired. Stop when you are done!”

"Selalu ada harapan bagi mereka yang sering berdoa, selalu ada jalan bagi mereka yang sering berusaha"

"Tugas Akhir ini kupersembahkan untuk kedua orang tua ku tercinta, keempat saudaraku dan keluarga besar yang senantiasa memberikan semangat serta do'a yang tiada tara sehingga tugas akhir ini dapat selesai"

"Last but not least, I wanna thank me” (Azis Mulki Rafani, S.Kom)"

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh, puji dan syukur penulis panjatkan kepada Allah SWT yang telah memberikan nikmat iman, nikmat sehat, taufik, karunia dan rahmat-Nya sehingga penulis dapat menyelesaikan Tugas Akhir ini yang berjudul "*Penerapan Corpus-Based Text Similarity* Sebagai Pengukur Kesamaan Fitur Pada Data Bibliografi Untuk Meningkatkan Akurasi Klasifikasi Kesamaan Penulis".

Penulis berharap agar tulisan ini dapat bermanfaat bagi banyak orang dan menjadi bahan bacaan yang menarik bagi peneliti tentang permasalahan Author Name Disambiguation (AND) terutama pada kasus author matching.

Pada penyusunan Tugas Akhir ini, tidak terlepas dari bantuan, bimbingan serta dukungan dari berbagai pihak. Oleh sebab itu, pada kesempatan ini penulis ingin mengucapkan rasa terima kasih kepada :

1. Orang tua saya tercinta, serta Uda dan Uni-uni saya, yang telah membimbing saya dengan penuh kasih sayang dengan selalu memberikan do'a, semangat, motivasi, serta dukungan baik secara finansial maupun spiritual.
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 Kompuer Fakutas Ilmu Komputer Universitas Sriwijaya
4. Bapak Dr. Erwin, S.Si, M.Si, selaku Dosen Pembimbing Akademik di Jurusan Sistem Komputer
5. Bapak Firdaus, S.T., M.Kom., selaku Pembimbing Tugas Akhir yang telah berkenan meluangkan waktunya guna membimbing, memberikan arahan dan saran serta motivasi yang tiada henti dalam memberikan bimbingan terbaik untuk penulis dalam menyelesaikan Tugas Akhir ini.

6. 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.
7. 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 memberikan semangat serta motivasi kepada penulis sehingga dapat menyelesaikan Tugas Akhir ini dengan baik.
8. Team ISysRG terutama team teks processing (Qiliq, Irvan, Suci, Annisa, Wais, dan Jorgi) yang selalu saling memberikan support dan menyemangati satu sama lain serta saling membantu agar dapat menyelesaikan Tugas Akhir ini.
9. Sinta Bella, yang selalu setia menemani, memberikan semangat, motivasi serta turut membantu penulis dalam segala hal terutama penyelesaian Tugas Akhir ini.
10. Teman-teman seperjuangan Sistem Komputer Angkatan 2017, serta semua pihak yang tidak dapat penulis sebutkan satu-persatu.
11. Last but not least, i want to thank me, i want to thank me for believing in me, i want to thank me for doing all this hard work, i want to thank me for having no days off, i want to thank me for never quitting, i want to thank me for being a giver, i want to thank me for just being me at all times.

Tentunya dalam pembuatan Tugas Akhir ini, masih jauh dari kata sempurna. Oleh karena itu sebagai bahan perbaikan kedepan penulis tentunya mengharapkan kritik dan saran yang membangun terhadap isi dari Tugas Akhir ini. Sehingga dapat menjadi bahan bacaan yang bermanfaat guna menambah wawasan terutama untuk yang tertarik pada penelitian tentang klasifikasi author matching.

Wassalamu'alaikum Warahmatullahi Wabarakatuh.

Indralaya, Juni 2021
Penulis,

Azis Mulki Rafani
Nim. 09011281722034

IMPLEMENTATION OF CORPUS-BASED TEXT SIMILARITY AS FEATURE MEASUREMENT ON BIBLIOGRAPHIC DATA TO IMPROVE THE ACCURACY OF AUTHOR MATCHING CLASSIFICATION

AZIS MULKI RAFANI (09011281722034)

Computer Engineering Department, Computer Science Faculty, Sriwijaya University

Email : azismulki1@gmail.com

ABSTRACT

Author Name Disambiguation (AND) was a case of ambiguity of the author's name that occurred in a publication in the Digital Library (DL) database which was caused by Synonymity and Homonymity conditions in the author's name. In this final project, it proposed the implementation of Corpus-based Text Similarity using Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction with Deep Neural Network (DNN) classifier, Support Vector Machine and Random Forest. The dataset used was the DBLP Labeled Data Dataset by Jinseok Kim, et al. This research focused on feature extraction in data processing in order to create effective features for use in classification. Parameters accuracy, precision, and recall were benchmarks to determine the level of success of the method used to overcome AND problems in the case of author matching. Of the 2 approaches and 3 classifiers, the best results were obtained in the BOW approach using the Random Forest classifier, which had accuracy, precision and recall results of 99.80%, 99.84% and 99.95%.

Keywords **:** *Author Name Disambiguation, Synonym, Homonym, Bibliographic Data, Digital Library, BOW, TF-IDF, Deep Neural Network, Support Vector Machine, Random Forest*

PENERAPAN *CORPUS-BASED TEXT SIMILARITY* SEBAGAI PENGUKUR KESAMAAN FITUR PADA DATA BIBLIOGRAFI UNTUK MENINGKATKAN AKURASI KLASIFIKASI KESAMAAN PENULIS

AZIS MULKI RAFANI (09011281722034)
Jurusan Sistem Komputer, Fakultas Ilmu Komputer, Universitas Sriwijaya
Email : azismulki1@gmail.com

ABSTRAK

Author Name Disambiguation (AND) adalah kasus ambiguitas nama penulis yang terjadi pada suatu publikasi dalam database *Digital Library* (DL) yang disebabkan karena kondisi *Synonymity* Dan *Homonymity* pada nama penulis (*author*). Dalam tugas akhir ini mengusulkan penerapan *Corpus-based Text Similarity* dengan menggunakan *feature extraction Bag of Words* (BOW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF) dengan *classifier Deep Neural Network* (DNN), *Support Vector Machine* dan *Random Forest*. Dataset yang digunakan adalah *Dataset DBLP Labeled Data* oleh Jinseok Kim, dkk. Penelitian yang dilakukan berfokus pada ekstraksi fitur dalam pengolahan data guna menciptakan fitur yang efektif untuk digunakan dalam klasifikasi. Parameter *accuracy*, *precision*, dan *recall* merupakan tolak ukur untuk mengetahui tingkat keberhasilan dari metode yang digunakan untuk mengatasi permasalahan AND pada kasus *author matching*. Dari 2 pendekatan dan 3 *classifier*, hasil terbaik didapatkan pada pendekatan BOW menggunakan *classifier Random Forest*, yang memiliki hasil *accuracy*, *precision* dan *recall* sebesar 99,80%, 99,84% dan 99,95%.

Kata Kunci : *Author Name Disambiguation, Synonym, Homonym, Bibliographic Data, Digital Library,* BOW, TF-IDF, *Deep Neural Network, Support Vector Machine, Random Forest*

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

DAFTAR LAMPIRAN

BAB 1
PENDAHULUAN

1.1. Latar Belakang

Database bibliografi termasuk sejumlah besar data dari perpustakaan digital[1]. Penulis atau peneliti dapat memiliki nama yang mirip, dapat memiliki berbagai cara untuk menulis nama lengkap mereka, atau penulis yang berbeda dapat menggunakan banyak nama, situasi ini menyebabkan ambiguitas terhadap penulis [2], [3], [4]. Ketidakjelasan nama penulis dalam database bibliografi telah lama dianggap sebagai masalah yang penting [5]. Secara khusus, ambiguitas dari nama seorang penulis adalah masalah yang terjadi jika penulis dengan nama yang sama (homonim) atau variasi nama yang berbeda untuk orang yang sama (sinonim)[2], [3], [6], [7], [8], [9] termasuk kurangnya standar dalam publikasi database bibliografi pada *digital libraries* [2].

Author Name Disambiguation (AND) adalah masalah *big* data yang tak kunjung hilang [8], [10], [11]. Seringkali terjadinya perbedaan penulisan nama mengakibatkan timbulnya persepsi bahwa orang yang sama seringkali dianggap sebagai orang yang berbeda [8]. Masalah ambiguitas nama ini memperburuk kinerja pencarian informasi di perpustakaan digital dan pencarian web [3], [12], [13]. Sebab hal ini bukanlah hal yang sepele untuk membedakan referensi nama tersebut, terutama ketika informasi tentang mereka sangat terbatas. Kebanyakan studi yang ada menggunakan fitur seperti alamat email, kata-kata yang sering digunakan oleh penulis, dll. Namun, informasi tersebut tidak selalu tersedia karena privasi atau terlalu mahal untuk didapat [13]. Semakin banyak dokumen dan kutipan yang diterbitkan setiap tahun, sistem apa pun yang dibangun di atas data ini harus terus dilatih ulang dan diklasifikasikan ulang agar tetap relevan dan bermanfaat [10]. Penelitian terbaru, *Deep Neural Network* (DNN) mampu menunjukkan kemampuan yang kuat untuk melakukan banyak tugas, terutama proses klasifikasi. Di sisi lain, terdapat teknik pembelajaran yang lain, yaitu SVM dan Random Forest yang dikenal karena kemampuannya dalam melakukan klasifikasi [14]. Sebelum diklasifikasi dengan metode *classifier*, atribut dataset diolah terlebih dahulu (nama

penulis, judul, dan lainnya) dengan menggunakan *measure* seperti *Bag Of Words* (BOW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF) [7], [8].

Dalam klasifikasi teks, metode *classifier* tidak dapat memproses teks secara langsung. Sebagai gantinya, diperlukan teknik ekstraksi fitur untuk mengubah teks menjadi bentuk numerik. Teknik ekstraksi memiliki efek yang besar pada akurasi klasifikasi. Oleh karena itu, sebelum dilakukannya proses klasifikasi akan diterapkan *corpus-based text similarity* sebagai ekstraksi fitur, yaitu *Bag of words* (BOW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF) yang dikenal handal dalam hal ini. Dimana pada penelitian sebelumnya, hasil analisis BOW dan TF-IDF menunjukkan akurasi klasifikasi yang lebih baik dan memiliki dampak positif pada kinerja klasifikasi [15].

Umumnya, ambiguitas nama penulis diselesaikan dengan menggunakan atribut publikasi yang berbeda seperti rekan penulis, kata judul, kata kunci, referensi, kata abstrak, tempat dan tahun publikasi. Namun, setiap *Digital Library* (DL) tidak menyediakan semua atribut ini, mereka hanya memberikan sedikit informasi tentang atribut ini dan tidak mungkin dilakukan dalam skala besar. Lebih dari itu, beberapa tahun terakhir jumlah data publikasi semakin membludak yang diterima dan diimpor di DL membuat masalah ambiguitas nama menjadi lebih parah. Metode yang menyelesaikan masalah ambiguitas nama di DL disebut disambiguasi nama penulis (AND) [2], [16].

Pada penelitian sebelumnya dengan kumpulan data yang berskala besar, teknik *pairwise* yang digunakan adalah dengan melakukan kombinasi dari masing-masing atribut dalam kumpulan data AND dan dengan menentukan kelas biner untuk setiap penulis yang sama, dimana penulis yang tidak sama akan diberi dengan nilai 0 dan penulis yang sama akan diberi nilai 1. Teknik tersebut menghasilkan data *imbalanced* yang sangat tinggi dimana kelas 0 sebesar 98.9% dan kelas 1 sebesar 1.1% dari jumlah data tersebut. Dari hasil tersebut dapat dianalisis, dimana kelas 1 dapat dianggap dan diolah sebagai anomali data dari keseluruhan data. Oleh karena itu, metode deteksi anomali yang dipilih dalam penelitian ini menggunakan algoritma *Isolation Forest* sebagai pengklasifikasiannya. Dengan hasil yang didapat sangat memuaskan dari segi akurasi yaitu mencapai 99.5% [17].

Basis penelitian adalah pengembangan serta kelanjutan dari penelitian yang telah ada sebelumnya, data dan metode diperlukan guna mendukung dan menemukan dan diperlukan data dan metode yang saling mendukung untuk menemukan jalan keluar terbaik atas masalah AND, terkhusus pada identifikasi penulis. *Deep learning neural network* (DNN) adalah metode klasifikasi dan pengambilan keputusan yang relatif stabil saat memproses varian data dalam pembelajaran mesin (ML). Penelitian terbaru [18], [19], [20] telah menunjukkan kemampuan yang kuat dalam pembelajaran fitur di banyak tugas. Fitur internal yang dipelajari oleh DNN relatif stabil untuk varian dalam data jika data pelatihan cukup representative [18]. Hal ini membantu menangani kesalahan sebuah kutipan, yang merupakan tantangan terbuka yang ditunjukkan oleh Ferreira et al. [2]. Selain itu, penggunaan *neural network* memiliki keunggulan yaitu dapat membangun model umum. Model ini dapat membedakan nama penulis secara bertahap ketika catatan publikasi baru dimasukkan ke dalam kumpulan data [21]. Tanpa bantuan langsung dari para ahli, metode *deep learning* mampu mempelajari fitur dengan baik [21].

Support Vector Machine (SVM) adalah metode klasifikasi yang populer dan efisien. Baru-baru ini, banyak aplikasi yang dikembangkan oleh para peneliti karena meningkatnya minat pada metode SVM. SVM juga telah banyak digunakan dalam pengolahan citra, klasifikasi gambar, klasifikasi dokumen, pengenalan karakter dari tulisan tangan. Selain itu juga SVM memiliki kemampuan klasifikasi yang cepat dan akurat [22].

Random Forest adalah pembelajaran ensemble guna memecahkan masalah klasifikasi dan regresi. RF dikenal dapat menangani tugas-tugas pelatihan pada sebuah dataset dengan hasil yang baik. Selain itu juga RF relatif mudah diterapkan dan baru-baru ini para peneliti telah menunjukkan peningkatan minat terhadap RF. Di mana RF memiliki kelebihan dalam melakukan prediksi di luar sampel dengan cepat, tidak sensitif terhadap fitur yang berdimensi tinggi dan parameter yang diatur hanya sedikit. Sebab itulah RF dikenal sebagai salah satu metode klasifikasi yang handal dan juga mudah diterapkan [23].

Dari ketiga metode *classifier* tersebut sangat sering dijumpai dalam penelitian AND. Pada klasifikasi DNN, SVM ataupun RF nilai akurasi menjadi

tolak ukur dari performa klasifikasi tersebut apakah menghasilkan nilai yang bagus atau tidak terhadap dataset yang digunakan. Terkhusus pada permasalahan AND, nilai *Precision* dan *Recall* merupakan parameter penting untuk menentukan kinerja penyelesaian klasifikasi dengan benar untuk masalah AND yang dihadapi, yaitu target yang diidentifikasi oleh penulis. Berdasarkan hal tersebut, penelitian ini akan menggunakan *Deep Neural Network* (DNN), *Support Vector Machine* (SVM) dan *Random Forest* untuk mengklasifikasikan penulis berdasarkan data bibliografi [24], [25].

1.2 Tujuan dan Manfaat

1.2.1. Tujuan

Pada Tugas Akhir ini bertujuan untuk:

1. Penerapan *Corpus-Based Text Similarity* pada kasus *author matching*.
2. Dapat menentukan *similarity measure* dan metode klasifikasi terbaik guna memecahkan masalah *Author Name Disambiguation* (AND) pada kasus *author matching*.

1.2.2. Manfaat

Manfaat dari penulisan Tugas Akhir ini, yaitu :

1. Dapat menyelesaikan permasalahan *Author Name Disambiguation* (AND) pada kasus *author matching*.
2. Hasil dari *similarity measure* dan metode klasifikasi dapat dijadikan referensi pada studi kasus *author matching*.

1.3 Perumusan Masalah

Bagaimana menerapkan pendekatan *Corpus-based Text Similarity* dalam studi kasus *author matching* pada data bibliografi untuk mendapatkan hasil yang akurat.

1.4 Batasan Masalah

Adapun batasan masalah yang terdapat pada penelitian tugas akhir ini, yaitu:

1. Penelitian yang digarap mencakup kasus *Author Name Disambiguation* (AND) terukhusus pada kasus *author matching*.
2. Bahasa pemrograman *Python* merupakan basis bahasa pada penelitian ini.
3. Dataset DBLP Labeled Data yang digarap oleh Jinseok Kim et al [12] merupakan dataset bibliografi sebagai bahan yang digunakan dalam penelitian ini. Dataset tersebut bersumber dari dblp.org dan data yang digunakan ini telah melalui tahap pembersihan (*Cleaning Process*).
4. Penelitian ini menggunakan pendekatan BOW dan TF-IDF dengan *classifier* DNN, SVM dan *Random Forest*. Hasil dari ketiga metode yang digunakan ini nantinya akan dibandingkan guna menentukan pembelajaran mesin (*Machine Learning*) terbaik dalam mengatasi permasalahan terutama pada kasus AND.
5. Hasil penelitian ini nantinya menghasilkan *performance measurement* berupa nilai Akurasi, *Spesifisity*, Presisi, *Recall*, *Error-Rate* dan F1-*Score* sebagai patokan guna melihat performa dari metode yang digunakan serta hasil persentase kebenaran dalam permasalahan *synonym*, *homonym* dan *non synonym homonym* untuk melihat tingkat kesamaan penulis.

1.5 Metodologi Penelitian

Penulisan pada Tugas Akhir ini menggunakan metodologi penelitian sebagai berikut:

1.5.1 Metode Studi Pustaka dan Literatur

Pada metode ini pencarian dilakukan untuk mengumpulkan referensi dalam bentuk literatur yang ada pada internet dan buku terhadap klasifikasi *Author Name Disambiguation* (AND) dengan menerapkan *Corpus-based Text Similarity* menggunakan *classifier Deep Neural Network* (DNN), *Support Vector Machine* (SVM) dan *Random Forest*.

1.5.2 Metode Konsultasi

Pada metode ini, penulis berkonsultasi kepada pihak narasumber yang memiliki pengetahuan dan wawasan baik secara secara langsung ataupun tidak langsung guna memecahkan kasus penulisan Tugas Akhir penulis mengenai klasifikasi *Author Name Disambiguation* (AND) dengan menerapkan *Corpus-based Text Similarity* menggunakan *classier Deep Neural Network* (DNN), *Support Vector Machine* (SVM) dan *Random Forest*.

1.5.3 Metode Pembuatan Model

Pada metode ini akan merancang pembuatan model dengan menerapkan *Corpus-based Text Similarity* yang dilakukan menggunakan program berbasis *python*.

1.5.4. Metode Pengujian dan Validasi

Pada metode ini akan dilakukan uji coba terhadap kinerja sistem yang sudah dibuat untuk melihat batasan kinerja, apakah membuahkan hasil yang baik atau malah sebaliknya.

1.5.5 Metode Hasil dan Analisa

Setelah dilakukan pengujian dengan metode ini, semua kelebihan dan kekurangan akan dianalisa, dengan harapan dapat menjadi acuan yang baik dalam penelitian selanjutnya.

1.5.6 Metode Penarikan Kesimpulan dan Saran

Metode ini merupakan tahap terakhir dari metodologi penelitian. Dalam metode ini nantinya akan diambil sebuah kesimpulan dan saran terhadap penelitian selanjutnya berdasarkan dari hasil dan analisis yang diperoleh.

1.6 Sistematika Penelitian

Pada tugas akhir ini sistematika penulisan yang digunakan adalah sebagai berikut:

BAB I PENDAHULUAN

Sebagai landasan terhadap penelitian yang dilakukan, bab ini akan membahas secara sistematis dari latar belakang penelitian, tujuan penelitian, rumusan masalah serta sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Pada bab ini akan membahas penjelasan tentang teori, konsep serta prinsip dasar dalam memecahkan masalah terhadap penelitian yang dilakukan.

BAB III METODOLOGI PENELITIAN

Pada bab ini menjelaskan proses penelitian, dimulai dari teknik, metode, dan jalannya proses penelitian.

BAB IV HASIL DAN ANALISIS

Pada bab ini membahas tentang hasil tes dan analisis yang didapatkan termasuk kelebihan dan kekurangan dari hasil yang dicapai terhadap penelitian yang telah dilakukan.

BAB V KESIMPULAN

Pada bab ini membahas kesimpulan yang diambil dari hasil dan analisis penelitian yang dilakukan, dan saran penelitian lebih lanjut tentang *Author Matching*.

DAFTAR PUSTAKA

[1] M. Shoaib, A. Daud, and T. Amjad, “Author name disambiguation in bibliographic databases: A survey,” *arXiv*, pp. 1–24, 2020.

[2] 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, doi: 10.1145/2350036.2350040.

[3] 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, doi: 10.1007/s13369-018-3099-0.

[4] 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.

[5] V. I. Torvik and N. R. Smalheiser, “Author name disambiguation in MEDLINE,” *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 3, pp. 1–29, 2009, doi: 10.1145/1552303.1552304.

[6] P. Gnoyke, K. Kumar, and M. Kumaresh, “Author Name Disambiguation by Clustering based on Deep Learned Pairwise Similarities,” no. May, pp. 0–12, 2020.

[7] F. Firdaus, M. Anshori, S. P. Raflesia, A. Zarkasi, M. Afrina, and S. Nurmaini, “Deep Neural Network Structure to Improve Individual Performance based Author Classification,” *Comput. Eng. Appl. J.*, vol. 8, no. 1, pp. 77–83, 2019, doi: 10.18495/comengapp.v8i1.264.

[8] Z. YAMANI, S. NURMAINI, FIRDAUS, M. Naufal R, and W. K. SARI, “Author Matching Using String Similarities and Deep Neural Networks,” vol. 172, no. Siconian 2019, pp. 474–479, 2020, doi: 10.2991/aisr.k.200424.073.

[9] T. Backes, “The impact of name-matching and blocking on author disambiguation,” *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 803–812, 2018, doi: 10.1145/3269206.3271699.

[10] Z. Zhao, J. Rollins, L. Bai, and G. Rosen, “Incremental Author Name Disambiguation for Scientific Citation Data,” *Proc. - 2017 Int. Conf. Data*

Sci. Adv. Anal. DSAA 2017, vol. 2018-Janua, pp. 175–183, 2017, doi: 10.1109/DSAA.2017.17.

[11] S. Zhang, X. E., and T. Pan, "A Multi-Level Author Name Disambiguation Algorithm," *IEEE Access*, vol. 7, pp. 104250–104257, 2019, doi: 10.1109/ACCESS.2019.2931592.

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

[13] W. Zhang, Z. Yan, and Y. Zheng, "Author name disambiguation using graph node embedding method," *Proc. 2019 IEEE 23rd Int. Conf. Comput. Support. Coop. Work Des. CSCWD 2019*, pp. 410–415, 2019, doi: 10.1109/CSCWD.2019.8791898.

[14] 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.

[15] F. M. Al-kharboush and M. A. Al-hagery, "Features Extraction Effect on the Accuracy of Sentiment Classification Using Ensemble Models," vol. 10, no. 3, pp. 2019–2022, 2021, doi: 10.21275/SR21303123511.

[16] 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.

[17] Z. Yamani, S. Nurmaini, D. Palupi, F. Firdaus, and A. Darmawahyuni, "Author Matching Classification with Anomaly Detection Approach for Bibliomethric Repository Data," *Comput. Eng. Appl.*, vol. 9, no. 2, pp. 79–92, 2020, [Online]. Available: https://comengapp.unsri.ac.id/index.php/comengapp/article/view/335.

[18] D. Yu, M. L. Seltzer, J. Li, J. T. Huang, and F. Seide, "Feature learning in deep neural networks – Studies on speech recognition tasks," *1st Int. Conf. Learn. Represent. ICLR 2013 - Conf. Track Proc.*, pp. 1–9, 2013.

[19] Y. Zhang, J. Gao, and H. Zhou, "Breeds Classification with Deep Convolutional Neural Network," *ACM Int. Conf. Proceeding Ser.*, pp. 145–

151, 2020, doi: 10.1145/3383972.3383975.

[20] E. Ueda, Y. Hirohata, T. Hino, and T. Yamashina, “Lower limit of pressure measurement using a spinning rotor gauge,” *Vacuum*, vol. 44, no. 5–7, pp. 587–589, 1993, doi: 10.1016/0042-207X(93)90102-G.

[21] 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, doi: 10.1007/978-3-319-05476-6_13.

[22] S. Ougiaroglou, K. I. Diamantaras, and G. Evangelidis, “Exploring the effect of data reduction on Neural Network and Support Vector Machine classification,” *Neurocomputing*, vol. 280, pp. 101–110, 2018, doi: 10.1016/j.neucom.2017.08.076.

[23] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki, “Random forest ensembles and extended multiextinction profiles for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 202–216, 2018, doi: 10.1109/TGRS.2017.2744662.

[24] 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, doi: 10.1145/1816123.1816130.

[25] A. F. Santana, M. A. Gonçalves, A. H. F. Laender, and A. A. Ferreira, “On the combination of domain-specific heuristics for author name disambiguation: the nearest cluster method,” *Int. J. Digit. Libr.*, vol. 16, no. 3–4, pp. 229–246, 2015, doi: 10.1007/s00799-015-0158-y.

[26] “International Meetings,” *Nature*, vol. 221, no. 5177, pp. 295–295, 1969, doi: 10.1038/221295a0.

[27] 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.

[28] J. Bollen, L. Alamos, M. A. Rodriguez, and H. Van De Sompel, “The Largest Scholarly Semantic Network ... Ever . Categories and Subject

Descriptors," 2007.

[29] A. Jinha, "Article 50 million: An estimate of the number of scholarly articles in existence," *Learn. Publ.*, vol. 23, no. 3, pp. 258–263, 2010, doi: 10.1087/20100308.

[30] 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, doi: 10.1007/s11192-014-1289-4.

[31] H. Han, W. Xu, H. Zha, and C. L. Giles, "A hierarchical naive Bayes mixture model for name disambiguation in author citations," p. 1065, 2005, doi: 10.1145/1066677.1066920.

[32] D. Han, S. Liu, Y. Hu, B. Wang, and Y. Sun, "ELM-based name disambiguation in bibliography," *World Wide Web*, vol. 18, no. 2, pp. 253–263, 2015, doi: 10.1007/s11280-013-0226-4.

[33] X. Liu, "Full-Text Citation Analysis : A New Method to Enhance," *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. July, pp. 1852–1863, 2013, doi: 10.1002/asi.

[34] L. V. B. Esperidião *et al.*, "Reducing Fragmentation in Incremental Author Name Disambiguation," *Jidm*, vol. 5, no. 3, pp. 293–307, 2014, [Online]. Available: https://seer.lcc.ufmg.br/index.php/jidm/article/view/721%0Ahttp://dblp.uni-trier.de.

[35] J. Zhu, X. Wu, X. Lin, C. Huang, G. P. C. Fung, and Y. Tang, "A novel multiple layers name disambiguation framework for digital libraries using dynamic clustering," *Scientometrics*, vol. 114, no. 3, pp. 781–794, 2018, doi: 10.1007/s11192-017-2611-8.

[36] T. Amjad, A. Daud, D. Che, and A. Akram, "MuICE: Mutual Influence and Citation Exclusivity Author Rank," *Inf. Process. Manag.*, vol. 52, no. 3, pp. 374–386, 2016, doi: 10.1016/j.ipm.2015.12.001.

[37] T. Amjad, A. Daud, A. Akram, and F. Muhammed, "Impact of mutual influence while ranking authors in a co-authorship network," *Kuwait J. Sci.*, vol. 43, no. 3, pp. 101–109, 2016.

[38] L. Shu, B. Long, and W. Meng, “A latent topic model for complete entity,” *Proc. - Int. Conf. Data Eng.*, pp. 880–891, 2009, doi: 10.1109/ICDE.2009.29.

[39] I. Type, C. Paper, and A. Dimitris, “The Universal Author Identifier System (UAI_Sys) - The University of Arizona Campus Repository,” 2012, [Online]. Available: http://arizona.openrepository.com/arizona/handle/10150/105755.

[40] A. Daud, J. Li, L. Zhou, and F. Muhammad, “Knowledge discovery through directed probabilistic topic models: A survey,” *Front. Comput. Sci. China*, vol. 4, no. 2, pp. 280–301, 2010, doi: 10.1007/s11704-009-0062-y.

[41] J. Tang, A. C. M. Fong, B. Wang, and J. Zhang, “A unified probabilistic framework for name disambiguation in digital library,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 975–987, 2012, doi: 10.1109/TKDE.2011.13.

[42] K. Kim, S. Rohatgi, and C. Lee Giles, “Hybrid deep pairwise classification for author name disambiguation,” *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 2369–2372, 2019, doi: 10.1145/3357384.3358153.

[43] T. Backes, “Effective Unsupervised Author Disambiguation with Relative Frequencies,” *Proc. ACM/IEEE Jt. Conf. Digit. Libr.*, pp. 203–212, 2018, doi: 10.1145/3197026.3197036.

[44] I. Hussain and S. Asghar, “LUCID: Author name disambiguation using graph Structural Clustering,” *2017 Intell. Syst. Conf. IntelliSys 2017*, vol. 2018-Janua, no. September, pp. 406–413, 2018, doi: 10.1109/IntelliSys.2017.8324326.

[45] A. N. Ngomo and P. K. Eds, *7th International Conference on Knowledge Engineering and Semantic Web, KESW 2016*, vol. 649. 2016.

[46] Z. Jianqiang and G. Xiaolin, “Comparison research on text pre-processing methods on twitter sentiment analysis,” *IEEE Access*, vol. 5, no. c, pp. 2870–2879, 2017, doi: 10.1109/ACCESS.2017.2672677.

[47] A. I. Kadhim, Y. N. Cheah, and N. H. Ahamed, “Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering,” *Proc. - 2014 4th Int. Conf. Artif. Intell. with Appl. Eng.*

Technol. ICAIET 2014, pp. 69–73, 2015, doi: 10.1109/ICAIET.2014.21.

[48] D. Sebastian and K. A. Nugraha, "Text normalization for Indonesian abbreviated word using crowdsourcing method," *2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019*, pp. 529–532, 2019, doi: 10.1109/ICOIACT46704.2019.8938463.

[49] N. Hanafiah, A. Kevin, C. Sutanto, Fiona, Y. Arifin, and J. Hartanto, "Text Normalization Algorithm on Twitter in Complaint Category," *Procedia Comput. Sci.*, vol. 116, pp. 20–26, 2017, doi: 10.1016/j.procs.2017.10.004.

[50] M. Javed and S. Kamal, "Normalization of unstructured and informal text in sentiment analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 10, pp. 78–85, 2018, doi: 10.14569/IJACSA.2018.091011.

[51] M. Allahyari *et al.*, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," *arXiv*, 2017.

[52] 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.

[53] A. Singh and S. Kumar, "A novel dice similarity measure for IFSs and its applications in pattern and face recognition," *Expert Syst. Appl.*, vol. 149, p. 113245, 2020, doi: 10.1016/j.eswa.2020.113245.

[54] J. Ye, "Cosine similarity measures for intuitionistic fuzzy sets and their applications," *Math. Comput. Model.*, vol. 53, no. 1–2, pp. 91–97, 2011, doi: 10.1016/j.mcm.2010.07.022.

[55] R. Subhashini and V. J. S. Kumar, "Evaluating the performance of similarity measures used in document clustering and information retrieval," *Proc. - 1st Int. Conf. Integr. Intell. Comput. ICIIC 2010*, pp. 27–31, 2010, doi: 10.1109/ICIIC.2010.42.

[56] A. Seal, A. Karlekar, O. Krejcar, and C. Gonzalo-Martin, "Fuzzy c-means clustering using Jeffreys-divergence based similarity measure," *Appl. Soft Comput. J.*, vol. 88, p. 106016, 2020, doi: 10.1016/j.asoc.2019.106016.

[57] K. Mikawa, T. Ishida, and M. Goto, "A proposal of extended cosine measure for distance metric learning in text classification," *Conf. Proc. -*

IEEE Int. Conf. Syst. Man Cybern., pp. 1741–1746, 2011, doi: 10.1109/ICSMC.2011.6083923.

[58] M. Gabryel, "The bag-of-words method with different types of image features and dictionary analysis," *J. Univers. Comput. Sci.*, vol. 24, no. 4, pp. 357–371, 2018.

[59] J. Yang, Y. Li, C. Gao, and Y. Zhang, "Measuring the short text similarity based on semantic and syntactic information," *Futur. Gener. Comput. Syst.*, vol. 114, pp. 169–180, 2021, doi: 10.1016/j.future.2020.07.043.

[60] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, "A survey on the techniques, applications, and performance of short text semantic similarity," *Concurr. Comput.* , vol. 33, no. 5, pp. 1–17, 2021, doi: 10.1002/cpe.5971.

[61] J. Wang and Y. Dong, "Measurement of text similarity: A survey," *Inf.*, vol. 11, no. 9, pp. 1–17, 2020, doi: 10.3390/info11090421.

[62] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Syst. Appl.*, vol. 66, pp. 1339–1351, 2016, doi: 10.1016/j.eswa.2016.09.009.

[63] C. Z. Liu, Y. X. Sheng, Z. Q. Wei, and Y. Q. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," *2018 IEEE Int. Conf. Intell. Robot. Control Eng. IRCE 2018*, no. 2, pp. 69–73, 2018, doi: 10.1109/IRCE.2018.8492945.

[64] Z. Zhu, J. Liang, D. Li, H. Yu, and G. Liu, "Hot Topic Detection Based on a Refined TF-IDF Algorithm," *IEEE Access*, vol. 7, no. c, pp. 26996–27007, 2019, doi: 10.1109/ACCESS.2019.2893980.

[65] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[66] 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, 2019, doi: 10.3390/en12010161.

[67] T. H. Nguyen and J. D. Zucker, "Enhancing metagenome-based disease prediction by unsupervised binning approaches," *Proc. 2019 11th Int. Conf.*

Knowl. Syst. Eng. KSE 2019, pp. 1–5, 2019, doi: 10.1109/KSE.2019.8919295.

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

[69] 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.

[70] 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."

[71] 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, doi: 10.1109/access.2020.3034680.

[72] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 8599–8603, 2013, doi: 10.1109/ICASSP.2013.6639344.

[73] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv*, pp. 1–12, 2017.

[74] 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.

[75] 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.

[76] G. Lai, Y. Yang, W. C. Chang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks," *arXiv*, pp. 95–104, 2017.

[77] 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.

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

[79] N. O. Attoh-Okine, "Analysis of learning rate and momentum term in backpropagation neural network algorithm trained to predict pavement performance," *Adv. Eng. Softw.*, vol. 30, no. 4, pp. 291–302, 1999, doi: 10.1016/S0965-9978(98)00071-4.

[80] D. A. Pisner and D. M. Schnyer, *Support vector machine*. Elsevier Inc., 2019.

[81] C. Gold and P. Sollich, "Model selection for support vector machine classification," *Neurocomputing*, vol. 55, no. 1–2, pp. 221–249, 2003, doi: 10.1016/S0925-2312(03)00375-8.

[82] A. Sarkar, S. Chatterjee, W. Das, and D. Datta, "Text Classification using Support Vector Machine Anurag," *Int. J. Eng. Sci. Invent.*, vol. 8, no. 2, pp. 33–37, 2015.

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

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

[85] 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.

[86] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[87] 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.

[88] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random

forest variable selection methods for classification prediction modeling," *Expert Syst. Appl.*, vol. 134, pp. 93–101, 2019, doi: 10.1016/j.eswa.2019.05.028.

[89] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, no. 1, pp. 93–104, 2012, doi: 10.1016/j.isprsjprs.2011.11.002.

[90] 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.

[91] Tin Kam Ho, "Random Decision Forests Tin Kam Ho Perceptron training," *Proc. 3rd Int. Conf. Doc. Anal. Recognit.*, pp. 278–282, 1995, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/598994/.

[92] Y. Amit and D. Geman, "Shape Quantization and Recognition with Randomized Trees," *Neural Comput.*, vol. 9, no. 7, pp. 1545–1588, 1997, doi: 10.1162/neco.1997.9.7.1545.

[93] Y. L. Pavlov, "Random forests," *Random For.*, pp. 1–122, 2019, doi: 10.1201/9780429469275-8.

[94] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved Random Forest for Classification," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4012–4024, 2018, doi: 10.1109/TIP.2018.2834830.

[95] 2018) (Al Amrani, Lazaar, El Kadiri., "Random Forest and Support Vector Machine based Hybrid Approach to SA --RF.pdf." .

[96] 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.

[97] 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.

[98] 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.

[99] C. P. Wolf, “Some lessons learned,” *Accid. Three Mile Isl. Hum. Dimens.*, pp. 215–232, 2019, doi: 10.4324/9780429048647-19.

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