BAB I
PENDAHULUAN

1.1. Latar Belakang

Di era digital saat ini, perpustakaan digital (*digital library*) telah menjadi pilar penting dalam ekosistem akademik. Keunggulannya, seperti kapasitas penyimpanan yang masif dengan biaya rendah serta kemudahan aksesibilitas data kapan pun dan di mana pun, telah mendorong pertumbuhannya secara eksponensial (Weiss, 2016). Repositori – repositori besar seperti DBLP, Google Scholar, PubMed, dan CiteSeer tidak hanya berfungsi sebagai gudang literatur, tetapi juga menyediakan layanan fungsional yang memfasilitasi penemuan ilmiah (Ferreira dkk., 2020). Namun, di balik manfaat tersebut, perpustakaan digital juga menghadapi tantangan signifikan terkait integritas dan kualitas data. Berbagai jenis kesalahan sering terjadi, mulai dari kesalahan tipografis, hasil pemindaian (*scanning*) yang tidak sempurna, hingga inkonsistensi metadata akibat format kutipan yang berbeda (Ferreira dkk., 2012). Di antara berbagai tantangan ini, salah satu masalah paling krusial dan persisten adalah ambiguitas nama penulis.

Disambiguasi Nama Penulis atau *Author Name Disambiguation* (AND) adalah proses untuk menentukan secara akurat apakah sekumpulan nama penulis dalam berbagai publikasi merujuk pada individu yang sama atau berbeda (Lewis dkk., 2006; Zhu dkk., 2014). Masalah ini muncul karena berbagai faktor, seperti adanya penulis dengan nama yang identik atau mirip, variasi penulisan nama (misalnya, penggunaan inisial), perubahan nama pribadi, kesalahan pengetikan, hingga transliterasi nama dari aksara non-Romawi (McKay dkk., 2010; Shoaib dkk., 2020). Secara spesifik, ambiguitas ini terwujud dalam dua bentuk utama (Hussain & Asghar, 2017; Shen dkk., 2017) yaitu: Sinonim (*Synonyms*) satu penulis yang sama muncul dengan beberapa variasi nama yang berbeda di berbagai publikasi dan Homonim (*Homonyms*) beberapa penulis yang berbeda memiliki nama yang sama persis. kegagalan dalam mengatasi masalah AND dapat menyebabkan atribusi karya yang salah dan analisis sitasi yang tidak akurat.

Berbagai solusi telah diusulkan untuk mengatasi masalah AND, yang secara umum dapat dikategorikan sebagai pendekatan *supervised learning*, *unsupervised learning*, dan berbasis graph. Di antara berbagai metode tersebut, teknik pencocokan berpasangan (*pairwise matching*) telah menjadi salah satu pendekatan yang paling dominan dan efektif. Pendekatan ini bekerja dengan membandingkan dua entri publikasi secara langsung untuk menentukan apakah keduanya merujuk pada individu penulis yang sama. (K. Kim dkk., 2019) menggunakan pendekatan *pairwise matching* dengan memasangkan publikasi yang terkait untuk menentukan penulis yang sama atau tidak menggunakan fitur *structure-based Term Frequency-Inverse Document Frequency* (TF-IDF) yang selanjutnya dilakukan pencarian kesamaan antar atribut menggunakan *cosine similarity*. Penelitian ini menggunakan algoritma *Support Vector Machine* (SVM)*, Random Forest* (RF)*, Gradient Boost Tree* (GBT)*,* dan *Deep Neural Network* (DNN) dalam melakukan klasifikasi AND. Hasil dari penelitian ini mendapatkan nilai MAP sebesar 0.8929 pada algoritma GBT, 0.8907 pada algoritma RF, 0.8878 pada algoritma SVM, dan 0.8771 pada algoritma DNN. (Boukhers & Asundi, 2023) menggunakan teknik *Char2Vec* untuk membandingkan dua referensi dengan memanfaatkan representasi vektor dari fitur seperti *co-authors,* judul, dan sumber publikasi. Penggunaan *embedding character Char2Vec* digunakan untuk mencari representasi vektor dari fitur. Penelitian ini juga menggunakan algoritma *neural network* dalam melakukan klasifikasi. (Ebrahimi dkk., 2023) menggunakan teknik *pairwise* dalam menemukan kecocokan antar penulis dengan menentukan prioritas dan bobot dari kriteria data yang kemudian dilakukan klasifikasi menggunakan algoritma *Analytical Hierarchy Process* dan *Fuzzy Delphi Method*.

Meskipun pendekatan *pairwise matching* terbukti efektif, metode ini secara inheren menciptakan sebuah tantangan baru yang signifikan, yaitu ketidakseimbangan data (*imbalanced data*). Sebagai representasi empiris dari masalah tersebut, penelitian ini menggunakan *dataset* berlabel turunan DBLP (*DBLP-derived labeled data*) yang dikembangkan oleh (J. J. Kim & Kim, 2018). *Dataset* ini dipilih karena secara eksplisit mendemonstrasikan permasalahan ketimpangan data pada skenario pelatihan sistem *author name disambiguation*, di

mana implementasi pencocokan berpasangan menghasilkan rasio yang sangat asimetris antara jumlah pasangan penulis yang berbeda (kelas mayoritas) berbanding pasangan penulis yang sama (kelas minoritas). Ketidakseimbangan ekstrem ini dapat menyebabkan model klasifikasi menjadi bias, cenderung memprediksi kelas mayoritas, dan pada akhirnya gagal mengenali kasus positif yang langka namun krusial (J. J. Kim & Kim, 2018; Manzoor dkk., 2022). Karena banyak penelitian sebelumnya tidak secara eksplisit menangani masalah kelas tidak seimbang ini secara mendalam pada tahap pra-pemrosesan, kondisi ini menciptakan celah penelitian (*research gap*) yang jelas untuk dieksplorasi.

Untuk mengisi celah tersebut, penelitian ini mengusulkan sebuah kerangka kerja yang secara spesifik dirancang untuk mengatasi masalah data tidak seimbang dalam konteks klasifikasi AND pada himpunan data DBLP. Solusi yang diajukan menerapkan algoritma *Variational Autoencoder* (VAE), sebuah model generatif *deep learning*, sebagai teknik *oversampling* tingkat lanjut. VAE dipilih karena kemampuannya yang unggul dalam mempelajari distribusi fitur tersembunyi (*latent space*) dari kelas minoritas. Dengan melatih model VAE secara eksklusif pada himpunan data kelas minoritas, algoritma ini mampu membangkitkan sampel-sampel data sintetis baru yang bervariasi dan berkualitas tinggi, sehingga menyeimbangkan proporsi kelas sebelum tahap klasifikasi (Ai dkk., 2023; Utyamishev & Partin-Vaisband, 2019).

Dengan demikian, penelitian ini mengusulkan sebuah alur kerja dua tahap: pertama, menyeimbangkan dataset berpasangan menggunakan VAE, dan kedua, melatih sebuah *Deep Neural Network* (DNN) pada dataset yang telah seimbang tersebut untuk melakukan klasifikasi akhir. Pendekatan hibrida ini diharapkan dapat menghasilkan model klasifikasi AND yang lebih kuat, akurat, dan tidak bias, sehingga memberikan solusi yang lebih optimal dibandingkan metode-metode sebelumnya yang tidak secara khusus menangani tantangan data tidak seimbang.

1.2. Identifikasi Masalah

Berdasarkan latar belakang yang telah dijelaskan, identifikasi masalah dari penelitian ini adalah :

1) Ambiguitas Nama Penulis

   Kesalahan dalam membedakan penulis dikarenakan mempunyai banyak penamaan cara yang sama dapat mengakibatkan kesalahan dalam atribusi publikasi. Kesalahan ambiguitas nama penulis mempunyai dua bentuk yaitu: sinonim dan homonim. Sinomin menjadi masalah karena penulis yang sama dapat tercatat dengan nama yang berbeda akibat transliterasi, perubahan nama, atau variasi gaya tulisan. Sedangkan, homonim menjadi masalah karena penulis yang berbeda tetapi memiliki nama yang sama sehingga dapat menyebabkan kebingungan dalam hal pengelompokkan publikasi.

2) Ketidakseimbangan Data

   Ketidakseimbangan data (*imbalanced data*) menjadi salah satu tantangan utama dalam melakukan klasifikasi *author name disambiguation*. Tahapan *Pre-Processing* data sebelum melakukan klasifikasi pada *author name disambiguation* sering menghasilkan data yang tidak seimbang antara kasus positif dan negatif sehingga dapat membuat sistem klasifikasi tidak akurat dan efisien, risiko *overfitting* pada kelas mayoritas, dan kesulitan dalam mendeteksi kasus ambiguitas yang relevan.

1.3. Perumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1) Bagaimana masalah ketidakseimbangan data memengaruhi efektivitas klasifikasi *author name disambiguation*?
2) Bagaimana cara mengatasi masalah data yang tidak seimbang (*imbalanced data*) guna meningkatkan akurasi identifikasi penulis yang ambigu?
3) Bagaimana mengembangkan arsitektur model klasifikasi menggunakan algoritma *deep learning* untuk menyelesaikan masalah *author name disambiguation*?
4) Sejauh mana integrasi algoritma *variational autoencoder* (VAE) dengan *deep neural network* (DNN) dapat meningkatkan performa klasifikasi dalam menyelesaikan masalah AND ?

1.4. Tujuan Penelitian

Tujuan umum penelitian ini adalah menghasilkan sebuah sistem untuk menangani data yang tidak seimbang pada tahap pemrosesan menggunakan algoritma VAE, serta untuk melakukan klasifikasi *author name disambiguation* dengan algoritma DNN. Guna mencapai tujuan umum tersebut, penelitian ini memiliki rincian tujuan sebagai berikut:

1) Mengembangkan solusi untuk mengatasi masalah *imbalanced data* dalam *author name disambiguation* (AND) guna meningkatkan akurasi identifikasi penulis yang ambigu.
2) Menerapkan algoritma *variational autoencoder* (VAE) untuk menghasilkan data sintetis dari kelas minoritas, sehingga dapat menyeimbangkan dataset pada masalah AND.
3) Mengintegrasikan algoritma *variational autoencoder* (VAE) dengan *deep neural network* (DNN) untuk meningkatkan performa klasifikasi dalam menyelesaikan masalah AND.
4) Mengukur kinerja dari model klasifikasi yang telah dirancang menggunakan *metric performance* dengan parameter *accuracy, precision, recall, specificity* dan *f-measure.*

1.5. Kebaruan (*Novelty*)

Kebaruan dalam penelitian ini terletak pada penggunaan algoritma *Variational Autoencoder* (VAE) dalam menyelesaikan masalah ketidakseimbangan data dalam proses klasifikasi *author name disambiguation* (AND. Pendekatan ini menawarkan solusi inovatif untuk mengatasi tantangan ketidakseimbangan data (*imbalanced data*) dengan memanfaatkan kemampuan VAE untuk menghasilkan data sintetis bagi kelas minoritas sekaligus menangkap struktur laten dari dataset. Selain itu, penelitian ini mengintegrasikan VAE dengan Deep Neural Network (DNN) untuk memanfaatkan kekuatan VAE dalam menyeimbangkan data dan kemampuan DNN dalam mengenali pola kompleks, sehingga meningkatkan performa klasifikasi secara keseluruhan. Kombinasi teknik generatif dan

diskriminatif yang diusulkan ini memberikan kontribusi baru yang efektif dalam menyelesaikan tantangan AND, baik dalam aspek teknis maupun implementasinya.

1.6. Manfaat Penelitian

Manfaat utama penelitian ini adalah menyediakan landasan teoretis dan praktis untuk penerapan algoritma VAE dalam menangani data tidak seimbang serta algoritma DNN dalam klasifikasi disambiguasi nama penulis. Secara lebih rinci, manfaat penelitian ini dijabarkan sebagai berikut:

1) Memberikan kontribusi penelitian di dalam bidang ambiguasi nama penulis.
2) Dapat menjadi salah satu solusi dalam mengatasi masalah data yang tidak seimbang dengan menggunakan dengan menggunakan algoritma VAE serta algoritma DNN dalam melakukan klasifikasi *author name disambiguation.*
3) Hasil dari penelitian ini dapat menjadi referensi untuk penelitian selanjutnya yang terkait.

1.7. Pembatasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1) Penelitian ini berfokus pada identifikasi penulis (*author*) dalam konteks masalah disambiguasi nama penulis (*Author Name Disambiguation*).
2) Proses pra-pemrosesan data menggunakan teknik pencocokan berpasangan *(pairwise matching),* kombinasi*,* dan pengukuran kesamaan data *(similarity).*
3) Sumber data yang digunakan adalah dataset bibliografi DBLP Labeled Data (J. J. Kim & Kim, 2018) yang telah melalui tahap pembersihan, dengan atribut yang dimanfaatkan meliputi: *author name, author id, paper id, author, year, venue, title, NaN*.
4) Validasi kinerja model akan dievaluasi menggunakan metrik performa yang terdiri atas parameter *accuracy, precision, recall, specificity,* dan *F-measure*.

1.8. Sistematika Penulisan

Sistematika penulisan ini bertujuan untuk memberikan gambaran yang jelas mengenai struktur dan isi setiap bab dalam penelitian. Adapun rincian sistematika tersebut adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini menguraikan dasar pemikiran penelitian yang mencakup latar belakang, identifikasi dan perumusan masalah, tujuan, serta manfaat yang diharapkan. Selain itu, bab ini juga menjelaskan kebaruan (novelty) yang ditawarkan, batasan masalah untuk menjaga fokus penelitian, dan diakhiri dengan sistematika penulisan yang menjadi panduan struktur laporan.

BAB II TINJAUAN PUSTAKA DAN LANDASAN TEORI

Bab ini menyajikan dua bagian utama. Bagian pertama adalah tinjauan pustaka yang merangkum penelitian-penelitian terdahulu yang relevan dengan topik disambiguasi nama penulis. Bagian kedua memaparkan landasan teori yang kokoh mengenai konsep dan teknologi yang digunakan, seperti *machine learning*, *deep learning*, *Variational Autoencoder* (VAE), *Deep Neural Network* (DNN), hingga metode evaluasi model.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan secara rinci mengenai rancangan dan langkah-langkah pelaksanaan penelitian. Pembahasan mencakup sumber data yang digunakan, prosedur penelitian yang sistematis mulai dari persiapan data, pra-pemrosesan, ekstraksi fitur, penanganan data tidak seimbang menggunakan VAE, hingga proses klasifikasi dengan DNN.

BAB IV HASIL DAN ANALISA

Bab ini memaparkan seluruh hasil dari eksperimen yang telah dilakukan. Hasil tersebut mencakup kinerja model pada data latih dan data uji yang kemudian dianalisis secara mendalam. Pada bab ini, validitas dan performa sistem yang dibangun akan diukur dan dibahas menggunakan berbagai metrik evaluasi untuk menguji hipotesis penelitian.

BAB V KESIMPULAN

Bab ini berisi rangkuman dari keseluruhan hasil penelitian dan analisis. Peneliti akan menarik kesimpulan yang menjawab rumusan masalah dan tujuan penelitian. Selain itu, bab ini juga akan mengemukakan keterbatasan yang dihadapi selama penelitian serta memberikan saran untuk pengembangan penelitian selanjutnya di masa mendatang.