AUTHOR NAMES DISAMBIGUATION DALAM KLASIFIKASI AUTHOR MATCHING DENGAN MENGGUNAKAN METODE MACHINE LEARNING PADA PENDEKATAN ANOMALY DETECTION

OLEH :

ZAQQI YAMANI A

09042611822005

PROGRAM MAGISTER ILMU KOMPUTER

FAKULTAS ILMU KOMPUTER

UNIVERSITAS SRIWIJYA

TAHUN 2020

BAB I
PENDAHULUAN

Bab ini menjelaskan tentang latar belakang diangkatnya tema permasalahan tesis ini, yang kemudian dari masalah dilakukan perumusan masalah. Di dalam bab ini pula dibahas mengenai batasan masalah, tujuan penelitian serta metodologi penulisan dalam penelitian ini.

1.1. Latar Belakang

Digital Library (DL) saat ini merupakan salah satu pusat literasi yang paling berkembang pesat, salah satunya pada bidang akademik. Hal ini dipengaruhi berbagai faktor seperti pemotongan anggaran untuk perpustakaan tradisional, ruang penyimpanan yang hampir tak terbatas dengan biaya yang jauh lebih rendah, kemudahan penggunaan dan tidak ada batas fisik dari penelitian (Palfrey, 2016; Weiss, 2016). Beberapa DL, antara lain DBLP1, MEDLINE2, CiteSeer3 arXiv4, MAS5, Google Scholar6, dan BDBComp7 secara luas digunakan oleh para peneliti untuk menemukan literatur ilmiah untuk penelitian dan penemuan mereka (Nicholson and Bennett, 2016). Selain memberikan beberapa informasi dan analisis yang berguna bagi pengambilan keputusan dan menyediakan konten berkualitas tinggi pada penelitian (Mitra *et al.*, 2007), DL juga memiliki beberapa sumber kesalahan antara lain adalah tipografi, pemindaian dan konversi data, menemukan dan mengganti, menyalin dan menempel, meta data, perangkat lunak pengumpulan kutipan yang tidak sempurna, format kutipan yang berbeda, nama-nama penulis yang ambigu, pembuatan konten terdesentralisasi dan singkatan dari judul tempat publikasi yang ambigu, dll (Ferreira, Gonçalves and Laender, 2012).

Ambiguitas Nama Penulis atau yang lebih dikenal dengan *Author Name Disambiguation* (AND) adalah salah satu masalah yang menurunkan kualitas dan keandalan informasi yang diperoleh dari DL (Tran, Huynh and Do, 2014). Konten DL dan kualitas layanan sangat dipengaruhi oleh masalah ambiguitas nama penulis dalam kutipan dan dianggap sebagai salah satu masalah tersulit yang dihadapi oleh para peneliti perpustakaan digital dari penelitian (Hussain and Asghar, 2017). AND menjadi sebuah masalah ketika satu set catatan publikasi berisi nama penulis yang

menimbulkan lebih dari satu interpretasi, yaitu penulis yang sama dapat muncul dengan nama yang berbeda (Ferreira, Gonçalves and Laender, 2012). Hal tersebut menjadi poin yang mengurangi kualitas dari informasi serta mengurangi pula keandalan informasi tersebut karena berdampak pada informasi terhadap penulis, organisasi dan hal lain yang ditampilkan sebagai bagian dari catatan publikasi tersebut (Müller, Reitz and Roy, 2017).

Beberapa penelitian telah dilakukan dalam rangka mengklasifikasikan AND, terutama pada proses klasifikasi *author matching*. Proses klasifikasi tersebut mengedepankan proses *pairwise* dan *distance* dari nama-nama author (Wang *et al.*, 2013). Klasifikasi terhadap data *author* diharapkan memberikan interpretasi yang tepat dan prediksi serta akurasi yang tinggi ketika dijalankan pada setiap dataset AND dan *bibliography*.

Dalam perkembangannya, AND dalam klasifikasi *author matching* menciptakan tantangan yang menakutkan dalam teknik disambiguasi karena sering menarik kesimpulan yang salah pada data publikasi yang tidak lengkap (Song, Kim and Kim, 2015). Ada sejumlah solusi yang dijalankan terhadap hal tersebut, diantaranya dengan *un-supervised* yaitu dilakukan berdasarkan kesamaan catatan bibliografi atau pola penulisan yang bersifat umum (Milojević, 2013), model NDMC atau multi step clustering yang dilakukan dengan menyamaratakan nama penulis atau menggabungkan karakteristik singkat dari informasi data publikasi (Gu *et al.*, 2016). Selain itu, ada teknik lain yang pernah digunakan yaitu LUCID, di mana dilakukan dengan menggunakan algoritma pendeteksian komunitas dan operasi grafik yang di akhir fase teknik ini tetap menggunakan fungsi kemiripan dari data publikasi tersebut (Hussain and Asghar, 2018b). Lalu ada teknik sistem analisis visual yang disebut NameClarifier yang secara interaktif mengelompokkan nama penulis dalam publikasi di dalam lingkaran tertentu, lalu menghitung dan memvisualisasikan kesamaan antara nama ambigu dan yang telah dikonfirmasi di Digital Library (Shen *et al.*, 2017). Namun, keempat metode tersebut tidak mementingkan akurasi dari proses klasifikasi data publikasi. Semua metode tersebut memberikan gambaran yang sama dalam teknik menuju klasifikasi yaitu dengan melakukan *pairwise* lalu menemukan kemiripan dari data yang dipasangkan tersebut.

Di sisi lain, *Machine Learning* juga telah banyak digunakan untuk melakukan proses klasifikasi AND dan menghasilkan kinerja yang memuaskan (Hussain and Asghar, 2017). Diantaranya dengan *supervised AND techniques* dengan menggunakan *boosted tree classification* yang fokus pada proses pemfilteran dan pencocokan nama dan afiliasi dalam sebuah publikasi (Wang *et al.*, 2013). *un supervised AND techniques* dengan pendekatan teori DST (*Dempster-Shafer Theory)* yaitu menghitung kesamaan fitur tingkat tinggi seperti afiliasi, tempat, content, rekan penulis, kutipan, korelasi Web(Wu *et al.*, 2014). Selanjutnya semi *supervised* AND *techniques* dengan menggunakan Algoritma mendeteksi kesamaan di antara objek. Mereka membangun matriks dua dimensi untuk penulisan bersama dan hubungan topik dan menghitung jarak antara dua simpul dengan bantuan jarak *Euclidean* (Zhu and Li, 2013). Selain itu, ada *Graph-Based AND techniques* dengan menggunakan algoritma *multi-level Graph Partinging* (MGP), dan algoritma *Multi-Level Graph Partinging* dan *Merging* (MGPM) (On, Lee and Lee, 2012). dan *Graph Based AND techniques* yang menggunakan kesamaan antara catatan *bibliografi* dan kelompok catatan baru untuk penulis dengan catatan kutipan yang sama di DL, atau untuk penulis baru ketika bukti kesamaan tidak cukup kuat. Beberapa *heuristik* khusus digunakan untuk memeriksa apakah referensi dari catatan kutipan baru milik penulis yang sudah ada di DL atau milik penulis yang baru (yaitu, penulis tanpa catatan kutipan di DL), menghindari menjalankan proses disambiguasi di seluruh DL (de Carvalho *et al.*, 2011). Namun, metode diatas menghasilkan kinerja akurasi, presisi, spesifisitas dan sensitivitas kurang memuaskan. Dan sama seperti sebelumnya, semua metode tersebut dilakukan dengan melakukan *pairwise,* lalu dilanjutkan dengan menghitung jarak untuk menghasilkan *similarity* dari data author yang dipasangkan.

Selain itu, metode lain yang pernah dilakukan untuk melakukan klasifikasi terdapat *author matching* yaitu dengan menggunakan *deep structure*. salah satu metode yang menggunakan struktur tersebut adalah metode *Deep Neural Network.* Arsitektur ini memiliki dua komponen utama. Dalam komponen pertama, data diambil sebagai input dan representasi data dihitung dengan mencari kemiripan dari data. Komponen kedua mengambil set fitur dasar sebagai inputnya dan mempelajari fitur-fitur di dalamnya adalah lapisan tersembunyi untuk menyamarkan nama

pembuatnya (Tran, Huynh and Do, 2014). Metode ini telah menghasilkan akurasi tinggi pada nilai 99,31%. Sama seperti penelitian sebelumnya, proses yang dilakukan adalah dengan *pairwise*. Namun, penelitian ini dilakukan dengan data yang sedikit dimana *pairwise* data menghasilkan hanya sekitar 30.537 data.

Dari berbagai metode yang telah dilakukan, dapat ditarik kesimpulan bahwa pra pemrosesan sebelum masuk pada proses klasifikasi adalah dengan melakukan *pairwise* dan dilanjutkan dengan menghitung jarak (*similarity)* dari data yang telah dipasangkan. Dalam prosesnya, teknik *pairwise* pada data author akan menimbulkan *imbalanced* data yang tinggi, dimana semakin banyak data yang dipasangkan akan menimbulkan tingkat *imbalanced* yang semakin tinggi pula atau dengan kalimat lain dimana data negatif menjadi sangat dominan dibanding data yang positif (Kim and Kim, 2018). Hal itu menjadikan hasil proses klasifikasi menjadi meragukan karena proses pencarian data yang bernilai positif menjadi sulit dan bisa dikatakan seperti mencari data yang langka.

Proses diatas sama halnya dengan dengan proses klasifikasi pada deteksi intrusi pada sistem komputer, dimana deteksi pada gangguan sistem komputer seperti mencari barang langka di dalam proses yang secara dominan berjalan normal (Ferreira, Gonçalves and Laender, 2012). Selain itu, kasus pada deteksi gangguan pada *network* pun memiliki masalah yang sama, dimana hal-hal yang dianggap gangguan atau *instrusion* merupakan hal yang sulit dikenali atau dicari karena persentase keberadaannya sangat kecil dibandingkan proses yang ada dan dianggap bukan merupakan sebuah *intrusion* (Dawoud, Shahristani and Raun, 2019). Adapun teknik yang dilakukan untuk mengklasifikan dua kasus diatas adalah dengan menggunakan metode *machine learning* pada pendekatan *Anomaly Detection* yang dapat menghasilkan nilai akurasi pada data negatif maupun data positif.

Berdasarkan hal tersebut, tingkat *imbalanced* yang tinggi pada proses klasifikasi *Author matching* dapat dikategorikan sama dengan yang terjadi pada proses *intrusi* atau gangguan pada sistem komputer dan *network.* Oleh karena itu, pendekatan *anomaly detection* akan menghasilkan tingkat akurasi yang tinggi pada proses klasifikasi. Sama seperti pada penggunaan *anomaly detection* di intrusi sistem komputer, model yang digunakan untuk proses klasifikasi pada penelitian

ini juga dengan menggunakan model algoritma *isolationForest* dan *local Outlier Factor (LOF).*

1.2. Perumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka perumusan masalah yang diambil pada penelitian ini adalah tentang "Bagaimana membuat sistem pengklasifikasi ambiguitas nama penulis (*author name disambiguation)* dalam *author matching* dengan menggunakan *Machine Learning* pada pendekatan *Anomaly Detection?*".

Dalam penelitian ini, perumusan masalah dijabarkan dalam bentuk pertanyaan sebagai berikut :

a. Bagaimana proses pra pengolahan data *Author Matching* dari suatu *digital library?*
b. Bagaimana menemukan model *Anomaly Detection* yang terbaik dengan menggunakan algoritma *IsolationForest* dan *Local Outlier Factor* atau *auto encoder* untuk melakukan klasifikasi pada data yang berdimensi tinggi?
c. Bagaimana mengukur hasil kinerja dari model *Anomaly Detection* berdasarkan parameter akurasi, sensitivitas, spesifisitas, presisi dan F1 score?

1.3. Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah :

1. System hanya berupa simulasi untuk melakukan proses klasifikasi data penulis (*author matching)* pada data di DL.
2. Dataset yang digunakan adalah data nama-nama penulis lengkap dengan judul jurnal serta tahun terbit yang diambil dari dataset SCAD-zBMATH dengan judul featured-dataset-merged yang terdiri dari 11.924 baris data publikasi (Müller, Reitz and Roy, 2017).
3. Pra pengolahan akan menggunakan teknik *pairwise* dan *similarity*.

1.4. Tujuan

Adapun tujuan dari penelitian ini adalah :

1. Melakukan proses pra pengolahan data publikasi yang memiliki disambiguitas untuk menginterpretasi dan mengelompokkan penulis berdasarkan kemiripan nama, judul, tahun dan nama kecil dari penulis.
2. Menganalisis model klasifikasi dari Anomaly Detection dalam mengklasifikasikan penulis yang sama atau bukan pada judul tulisan dan tahun terbit yang berbeda.
3. Mengukur kinerja klasifikasi Anomaly Detection berdasarkan parameter akurasi, sensitivitas, spesifisitas, presisi dan F1 score.

1.5. Manfaat

Manfaat dari penelitian ini adalah menjadi landasan penggunaan metode anomaly detection dalam klasifikasi Author Names Disambiguation terutama dalam konsep Author Matching. Selain itu, manfaat dari penelitian ini adalah sebagai berikut :

1. Anomaly Detection dapat menjadi metode tambahan baru dalam proses disambiguasi terhadap nama penulis.
2. Model klasifikasi yang dipakai dalam penelitian ini dapat digunakan lebih lanjut untuk meningkatka hasil accuracy dan presisi dalam proses AND terutama pada konsep Author Matching.

1.6. Metodologi Penulisan

Metodologi penulisan pada tesis ini terdiri dari lima bab sebagai berikut:

BAB I : PENDAHULUAN

Bab I berisi pendahuluan berupa latar belakang, perumusan masalah, tujuan dan manfaat dari topik yang dipilih.

BAB II: TINJAUAN PUSTAKA

Bab II berisi kerangka teori dan pustaka yang berhubungan dengan klasifikasi *Author Matching* pada data DL dengan menggunakan metode *Anomaly Detection* yang mengacu pada beberapa penelitian jurnal publikasi.

BAB III : METODOLOGI PENELITIAN

Bab III berisi metodologi yang menjelaskan secara bertahap dan terperinci tentang langkah-langkah yang digunakan untuk mencari, mengumpulkan dan menganalisa yang berkaitan dengan *author Matching*. Metodologi ini menjelaskan pendekatan atau algoritma *author matching*, serta model yang digunakan sehingga tujuan dari penulisan dapat tercapai.

BAB IV: HASIL DAN ANALISA SEMENTARA

Bab IV berisi hasil pengujian yang telah dilakukan, data-data yang diambil dari pengujian tersebut akan dianalisa menggunakan berbagai macam teknik, selain itu di bab ini juga membahas kevalidasian dari sistem yang telah dibuat.

BAB V: KESIMPULAN

BAB V berisi tentang kesimpulan apa yang diperoleh oleh penulis serta merupakan jawaban dari setiap tujuan yang ingin dicapai.

DAFTAR PUSTAKA

Amin, A. *et al.* (2016) 'Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study', *IEEE Access*, 4(Ml), pp. 7940–7957. doi: 10.1109/ACCESS.2016.2619719.

Berzins, K. *et al.* (2012) 'A boosted-trees method for name disambiguation', *Scientometrics*, 93(2), pp. 391–411. doi: 10.1007/s11192-012-0681-1.

de Carvalho, A. P. *et al.* (2011) 'Incremental Unsupervised Name Disambiguation in Cleaned Digital Libraries', *Journal of Information and Data Management*, 2(573871), p. 289.

Cheng, Z., Zou, C. and Dong, J. (2019) 'Outlier detection using isolation forest and local outlier', *Proceedings of the 2019 Research in Adaptive and Convergent Systems, RACS 2019*, pp. 161–168. doi: 10.1145/3338840.3355641.

Cota, R. G. *et al.* (2010) 'An Unsupervised Heuristic-Based Hierarchical Method for Name Disambiguation in Bibliographic Citations', 61(May), pp. 1853–1870. doi: 10.1002/asi.

Dawoud, A., Shahristani, S. and Raun, C. (2019) 'Dimensionality Reduction for Network Anomalies Detection: A Deep Learning Approach', in *Advances in Intelligent Systems and Computing*. Springer Verlag, pp. 957–965. doi: 10.1007/978-3-030-15035-8_94.

Ferreira, A. A. *et al.* (2010) 'Effective self-training author name disambiguation in scholarly digital libraries', *Proceedings of the ACM International Conference on Digital Libraries*, pp. 39–48. doi: 10.1145/1816123.1816130.

Ferreira, A. A., Gonçalves, M. A. and Laender, A. H. F. (2012) 'A brief survey of automatic methods for author name disambiguation', *ACM SIGMOD Record*, 41(2), p. 15. doi: 10.1145/2350036.2350040.

Ferreira, V. O. *et al.* (2015) 'A model for anomaly classification in intrusion detection systems', in *Journal of Physics: Conference Series*. doi: 10.1088/1742-6596/633/1/012124.

Firdaus, F. (2018) 'Improving Data Integrity of Individual-based Bibliographic Repository Using Clustering Techniques', *Computer Engineering and Applications Journal*, 7(1), pp. 49–56. doi: 10.18495/comengapp.v7i1.223.

Fugate, M. and Gattiker, J. R. (2002) 'Anomaly detection enhanced classification in computer intrusion detection', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 186–197. doi: 10.1007/3-540-45665-1_15.

Gong, Z. and Chen, H. (2016) 'Model-based oversampling for imbalanced

sequence classification', *International Conference on Information and Knowledge Management, Proceedings*, 24-28-October-2016, pp. 1009–1018. doi: 10.1145/2983323.2983784.

Gu, S. *et al.* (2016) 'Name Disambiguation Method Based on Multi-step Clustering', *Procedia Computer Science*, 83(Ant), pp. 488–495. doi: 10.1016/j.procs.2016.04.237.

Han, H. *et al.* (2004) 'Two supervised learning approaches for name disambiguation in author citations', *Proceedings of the ACM IEEE International Conference on Digital Libraries, JCDL 2004*, pp. 296–305.

Hazra, R. *et al.* (2016) 'An efficient technique for author name disambiguation', *2016 IEEE International Conference on Current Trends in Advanced Computing, ICCTAC 2016*. doi: 10.1109/ICCTAC.2016.7567344.

Hussain, I. and Asghar, S. (2017) 'A survey of author name disambiguation techniques: 2010–2016', *The Knowledge Engineering Review*, 32, p. e22. doi: 10.1017/S0269888917000182.

Hussain, I. and Asghar, S. (2018a) 'Author Name Disambiguation by Exploiting Graph Structural Clustering and Hybrid Similarity', *Arabian Journal for Science and Engineering*. Springer Berlin Heidelberg, 43(12), pp. 7421–7437. doi: 10.1007/s13369-018-3099-0.

Hussain, I. and Asghar, S. (2018b) 'LUCID: Author name disambiguation using graph Structural Clustering', *2017 Intelligent Systems Conference, IntelliSys 2017*, 2018-Janua(September), pp. 406–413. doi: 10.1109/IntelliSys.2017.8324326.

John, H. and Naaz, S. (2019) 'Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest International Journal of Computer Sciences and Engineering Open Access Credit Card Fraud Detection using Local Outlier Factor and Isolation', (September). doi: 10.26438/ijcse/v7i4.10601064.

Kharitonov, A. and Zimmermann, A. (2019) 'Intrusion detection using growing hierarchical self-organizing maps and comparison with other intrusion detection techniques', *CPSS 2019 - Proceedings of the 5th ACM Cyber-Physical System Security Workshop, co-located with AsiaCCS 2019*, pp. 13–23. doi: 10.1145/3327961.3329531.

Kim, J. (2018) 'Evaluating author name disambiguation for digital libraries: a case of DBLP', *Scientometrics*. Springer International Publishing, 116(3), pp. 1867–1886. doi: 10.1007/s11192-018-2824-5.

Kim, Jinseok and Kim, Jenna (2018) 'The impact of imbalanced training data on machine learning for author name disambiguation', *Scientometrics*. Springer International Publishing, 117(1), pp. 511–526. doi: 10.1007/s11192-018-2865-9.

Kunang, Y. N. *et al.* (2019) 'Automatic Features Extraction Using Autoencoder in Intrusion Detection System', *Proceedings of 2018 International Conference on Electrical Engineering and Computer Science, ICECOS 2018*. IEEE, (June 2019), pp. 219–224. doi: 10.1109/ICECOS.2018.8605181.

Leevy, J. L. *et al.* (2018) 'A survey on addressing high-class imbalance in big data', *Journal of Big Data*. Springer International Publishing, 5(1). doi: 10.1186/s40537-018-0151-6.

Li, N. and Han, J. (2017) 'The application of naive bayes classifier in name disambiguation', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10603 LNCS, pp. 611–618. doi: 10.1007/978-3-319-68542-7_52.

Lu, J. *et al.* (no date) 'Tutorial Proposal : Synergy of Database Techniques and Machine Learning Models for String Similarity Search and Join'.

Luque, A. *et al.* (2019) 'The impact of class imbalance in classification performance metrics based on the binary confusion matrix', *Pattern Recognition*, 91, pp. 216–231. doi: 10.1016/j.patcog.2019.02.023.

Milojević, S. (2013) 'Accuracy of simple, initials-based methods for author name disambiguation', *Journal of Informetrics*, 7(4), pp. 767–773. doi: 10.1016/j.joi.2013.06.006.

Mitra, P. *et al.* (2007) 'Are your citations clean?', *Communications of the ACM*, 50(12), pp. 33–38. doi: 10.1145/1323688.1323690.

Mozafari, F. and Tahayori, H. (2019) 'Emotion Detection by Using Similarity Techniques', *2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems, CFIS 2019*. IEEE, pp. 1–5. doi: 10.1109/CFIS.2019.8692152.

Müller, M. C., Reitz, F. and Roy, N. (2017) 'Data sets for author name disambiguation: an empirical analysis and a new resource', *Scientometrics*, 111(3), pp. 1467–1500. doi: 10.1007/s11192-017-2363-5.

Naway, A. and Li, Y. (2019) 'Android Malware Detection Using Autoencoder', pp. 1–9.

Nicholson, S. W. and Bennett, T. B. (2016) 'Dissemination and Discovery of Diverse Data: Do Libraries Promote Their Unique Research Data Collections?', *International Information and Library Review*, 48(2), pp. 85–93. doi: 10.1080/10572317.2016.1176448.

On, B. W., Lee, I. and Lee, D. (2012) 'Scalable clustering methods for the name disambiguation problem', *Knowledge and Information Systems*, 31(1), pp. 129–151. doi: 10.1007/s10115-011-0397-1.

Palfrey, J. (2016) ‘Design choices for libraries in the digital-plus era’, *Daedalus*, 145(1), pp. 79–86. doi: 10.1162/DAED_a_00367.

Qin, Y. and Lou, Y. (2019) ‘Hydrological time series anomaly pattern detection based on isolation forest’, *Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2019*. IEEE, (Itnec), pp. 1706–1710. doi: 10.1109/ITNEC.2019.8729405.

Rettig, L. *et al.* (2015) ‘Online anomaly detection over Big Data streams’, *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 1113–1122. doi: 10.1109/BigData.2015.7363865.

Shah, N. B., Balakrishnan, S. and Wainwright, M. J. (2016) ‘Feeling the bern: Adaptive estimators for Bernoulli probabilities of pairwise comparisons’, *IEEE International Symposium on Information Theory - Proceedings*, 2016-Augus, pp. 1153–1157. doi: 10.1109/ISIT.2016.7541480.

Shen, Q. *et al.* (2017) ‘NameClarifier: A Visual Analytics System for Author Name Disambiguation’, *IEEE Transactions on Visualization and Computer Graphics*, 23(1), pp. 141–150. doi: 10.1109/TVCG.2016.2598465.

Song, M., Kim, E. H. J. and Kim, H. J. (2015) ‘Exploring author name disambiguation on PubMed-scale’, *Journal of Informetrics*. Elsevier Ltd, 9(4), pp. 924–941. doi: 10.1016/j.joi.2015.08.004.

Tran, H. N., Huynh, T. and Do, T. (2014) ‘Author name disambiguation by using deep neural network’, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8397 LNAI(PART 1), pp. 123–132. doi: 10.1007/978-3-319-05476-6_13.

Vluymans, S. (2019) ‘Dealing with imbalanced and weakly labelled data in machine learning using fuzzy and rough set methods’, *Studies in Computational Intelligence*, 807, pp. 1–249. doi: 10.1007/978-3-030-04663-7_1.

Wang, J.-P. *et al.* (2013) ‘Effective string processing and matching for author disambiguation’, 15, pp. 1–9. doi: 10.1145/2517288.2517295.

Weiss, A. (2016) ‘Examining Massive Digital Libraries (MDLs) and Their Impact on Reference Services’, *Reference Librarian*, 57(4), pp. 286–306. doi: 10.1080/02763877.2016.1145614.

Wressnegger, C. *et al.* (2013) ‘A close look on n-grams in intrusion detection: Anomaly detection vs. classification’, in *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 67–76. doi: 10.1145/2517312.2517316.

Wu, H. *et al.* (2014) ‘Unsupervised author disambiguation using Dempster–Shafer theory’, *Scientometrics*, 101(3), pp. 1955–1972. doi: 10.1007/s11192-014-1283-x.

Yang, X. *et al.* (2019) ‘A fast and efficient local outlier detection in data streams’, *ACM International Conference Proceeding Series*, Part F1477, pp. 111–116. doi: 10.1145/3317640.3317653.

Yao, C. *et al.* (2019) ‘Distribution Forest: An Anomaly Detection Method Based on Isolation Forest’, in, pp. 135–147. doi: 10.1007/978-3-030-29611-7_11.

Zhang, T., Wang, E. and Zhang, D. (2019) ‘Predicting failures in hard drivers based on isolation forest algorithm using sliding window’, *Journal of Physics: Conference Series*, 1187(4). doi: 10.1088/1742-6596/1187/4/042084.

Zhao, J., Wang, P. and Huang, K. (2013) ‘A semi-supervised approach for author disambiguation in KDD CUP 2013’, in *Proceedings of the 2013 KDD Cup 2013 Workshop on - KDD Cup ’13*. doi: 10.1145/2517288.2517298.

Zhu, Y. and Li, Q. (2013) ‘Enhancing object distinction utilizing probabilistic topic model’, *Proceedings - 2013 International Conference on Cloud Computing and Big Data, CLOUDCOM-ASIA 2013*, pp. 177–182. doi: 10.1109/CLOUDCOM-ASIA.2013.61.