Artificial Intelligence for Unstructured Data Processing

Yohanes Bowo Widodo^{1*)}, Febrianti Widyahastuti²⁾, Mohammad Narji³⁾, Sondang Sibuea⁴⁾

¹⁾³⁾⁴⁾Teknik Informatika, Fakultas Komputer, Universitas Mohammad Husni Thamrin

²⁾Sistem Informasi, Fakultas Komputer, Universitas Mohammad Husni Thamrin

^{*)}Correspondence author: ybowowidodo@gmail.com, Jakarta, Indonesia

DOI: <https://doi.org/10.37012/jtik.v11i1.2542>

Abstract

In the digital era, the volume of unstructured data such as text, images, audio, and video continues to increase exponentially. Processing unstructured data is a major challenge for various industries due to its high complexity and the difficulty of extracting relevant information. Artificial Intelligence (AI) has become an innovative solution in addressing this challenge through techniques such as Natural Language Processing (NLP), Computer Vision, and Machine Learning. This study aims to explore various AI methods used in processing unstructured data and examine their effectiveness in improving the efficiency and accuracy of data analysis. adopts a multidisciplinary approach that combines natural language processing (NLP), machine learning, and data analytics techniques to extract information from unstructured data, especially in the context of electronic medical records (EMR). This study will be conducted in several stages including data collection, data processing, model development, and evaluation of results. The results show that AI is not only able to automate the information extraction process but also improve the accuracy and speed of data analysis, which is very important in the context of decision making in the fields of healthcare, finance, and business. By using deep learning models and advanced algorithms, AI can identify patterns and relationships in complex data, thereby providing deeper insights for better decision making. The results of this study are expected to provide insight for developers and practitioners in optimizing the use of AI to manage unstructured data more effectively and efficiently.

Keywords: Artificial Intelligence, Unstructured Data, Machine Learning, Natural Language Processing, Computer Vision.

Abstrak

Dalam era digital, volume data tidak terstruktur seperti teks, gambar, audio, dan video terus meningkat secara eksponensial. Pengolahan data tidak terstruktur menjadi tantangan utama bagi berbagai industri karena kompleksitasnya yang tinggi dan sulitnya ekstraksi informasi yang relevan. Artificial Intelligence (AI) telah menjadi solusi inovatif dalam menangani tantangan ini melalui teknik seperti Natural Language Processing (NLP), Computer Vision, dan Machine Learning. Penelitian ini bertujuan untuk mengeksplorasi berbagai metode AI yang digunakan dalam pengolahan data tidak terstruktur serta mengkaji efektivitasnya dalam meningkatkan efisiensi dan akurasi analisis data. mengadopsi pendekatan multidisipliner yang menggabungkan teknik pemrosesan bahasa alami (NLP), pembelajaran mesin, dan analisis data untuk mengekstraksi informasi dari data tidak terstruktur, khususnya dalam konteks catatan medis elektronik (EMR). Penelitian ini akan dilakukan dalam beberapa tahap yang mencakup pengumpulan data, pemrosesan data, pengembangan model, dan evaluasi hasil. Hasil penelitian menunjukkan bahwa AI tidak hanya mampu mengotomatiskan proses ekstraksi informasi, tetapi juga meningkatkan akurasi dan kecepatan analisis data, yang sangat penting dalam konteks pengambilan keputusan di bidang kesehatan, keuangan, dan bisnis. Dengan menggunakan model pembelajaran mendalam dan algoritma canggih, AI dapat mengidentifikasi pola dan hubungan dalam data yang kompleks, sehingga memberikan wawasan yang lebih mendalam untuk pengambilan keputusan yang lebih baik. Hasil penelitian ini diharapkan dapat memberikan wawasan bagi

pengembang dan praktisi dalam mengoptimalkan penggunaan AI untuk mengelola data tidak terstruktur secara lebih efektif dan efisien.

Kata Kunci: Artificial Intelligence, Data Tidak Terstruktur, Machine Learning, Natural Language Processing, Computer Vision

PENDAHULUAN

Pendahuluan penelitian ini berfokus pada pemanfaatan Kecerdasan Buatan (AI) dalam pengolahan data tidak terstruktur, yang semakin menjadi perhatian utama dalam bidang kesehatan dan teknologi informasi. Data tidak terstruktur, yang mencakup teks bebas, gambar, dan suara, merupakan bagian besar dari informasi yang dihasilkan di berbagai bidang, termasuk kesehatan, bisnis, dan media sosial. Data tidak terstruktur, seperti catatan medis dalam bentuk narasi bebas, menyimpan informasi berharga yang sering kali tidak dapat diakses melalui metode analisis tradisional. Dalam konteks ini, Natural Language Processing (NLP) muncul sebagai alat yang sangat penting untuk mengekstrak informasi klinis yang bermakna dari data tersebut (Koleck et al., 2019; Tayefi et al., 2021). Menurut penelitian, sekitar 80% dari data yang dihasilkan oleh organisasi adalah tidak terstruktur, dan pengolahan data ini sangat penting untuk pengambilan keputusan dan analisis prediktif (Baviskar et al., 2021).

Berdasarkan kajian sistematis, Koleck et al. (2019) menunjukkan bahwa NLP dapat digunakan untuk menganalisis gejala yang didokumentasikan dalam catatan kesehatan elektronik, yang mencakup tantangan dalam mendeteksi dan mengelola penyakit. Selain itu, Tayefi et al. (2021) menekankan pentingnya metode NLP dalam menganalisis laporan klinis yang sangat besar, yang dapat membantu dalam mendeteksi reaksi obat yang merugikan dan menentukan kelayakan uji klinis. Dengan demikian, NLP tidak hanya berfungsi sebagai alat analisis, tetapi juga sebagai jembatan untuk memahami kompleksitas data kesehatan yang tidak terstruktur.

Dalam konteks kesehatan, data tidak terstruktur sering kali ditemukan dalam catatan medis elektronik (EMR), di mana informasi penting dapat tersembunyi dalam narasi bebas seperti catatan kemajuan, ringkasan keluar, dan laporan konsultasi (More, 2024). Penggunaan metode pemrosesan bahasa alami (NLP) dan teknik pembelajaran mesin telah terbukti efektif dalam mengekstraksi informasi klinis yang bermakna dari data tidak

terstruktur ini, membantu dalam deteksi otomatis reaksi obat yang merugikan dan pemilihan kelayakan uji klinis (Dura et al., 2022; Ahn et al., 2021). Misalnya, alat cTAKES yang dikembangkan oleh Mayo Clinic telah menjadi salah satu alat yang paling umum digunakan untuk ekstraksi informasi klinis dari data tidak terstruktur (Wang et al., 2018).

Tantangan yang dihadapi dalam pengolahan data tidak terstruktur sangat beragam, termasuk variasi dalam gaya penulisan dan terminologi yang digunakan dalam catatan medis (Tayefi et al., 2021; More, 2024; (Tayefi et al., 2021; . mengidentifikasi bahwa meskipun terdapat banyak tantangan, ada peluang signifikan untuk mengembangkan alat diagnosis dan dukungan keputusan yang dapat mengintegrasikan data terstruktur dan tidak terstruktur secara efisien. Penelitian ini bertujuan untuk mengeksplorasi bagaimana AI, khususnya melalui teknik NLP, dapat dioptimalkan untuk meningkatkan akurasi dan efisiensi dalam pengolahan data tidak terstruktur, serta memberikan wawasan baru dalam praktik klinis dan penelitian kesehatan.

Dengan kemajuan teknologi dan algoritma pembelajaran mesin, seperti yang dijelaskan oleh (Ahn et al., 2021), penggunaan AI dalam menganalisis data kesehatan tidak terstruktur telah menunjukkan potensi yang besar dalam meningkatkan diagnosis dan pengobatan penyakit. Penelitian ini akan membahas berbagai aplikasi AI dalam konteks pengolahan data tidak terstruktur, serta tantangan dan peluang yang ada untuk implementasi lebih lanjut dalam praktik klinis.

Namun, meskipun ada kemajuan yang signifikan dalam teknologi AI, tantangan tetap ada dalam pengolahan data tidak terstruktur. Salah satu tantangan utama adalah keragaman dalam gaya penulisan dan terminologi yang digunakan dalam catatan medis, yang dapat mempersulit proses ekstraksi informasi (More, 2024). Selain itu, masalah etika dan sensitivitas data juga menjadi kendala dalam akses dan penggunaan data klinis untuk penelitian (Tayefi et al., 2021). Penelitian ini bertujuan untuk mengidentifikasi dan menganalisis tantangan-tantangan ini serta mengeksplorasi peluang yang ada untuk meningkatkan pemanfaatan data tidak terstruktur melalui penerapan AI.

Lebih jauh lagi, penelitian ini akan membahas berbagai teknik yang digunakan dalam pengolahan data tidak terstruktur, termasuk algoritma pembelajaran mendalam dan metode statistik lanjutan yang telah diterapkan untuk menganalisis teks klinis dan gambar (Tayefi et

al., 2021). Dengan memanfaatkan kemajuan dalam teknologi AI, diharapkan penelitian ini dapat memberikan wawasan baru tentang bagaimana data tidak terstruktur dapat diolah untuk menghasilkan informasi yang berguna dan mendukung pengambilan keputusan yang lebih baik dalam berbagai konteks, terutama dalam bidang kesehatan.

Secara keseluruhan, penelitian ini akan memberikan kontribusi penting bagi pemahaman tentang bagaimana AI dapat digunakan untuk mengatasi tantangan dalam pengolahan data tidak terstruktur dan bagaimana teknologi ini dapat diterapkan untuk meningkatkan hasil dalam berbagai bidang, termasuk kesehatan, bisnis, dan lainnya. Dengan demikian, penelitian ini tidak hanya relevan secara akademis tetapi juga memiliki implikasi praktis yang signifikan untuk pengembangan teknologi di masa depan.

METODE

Metode penelitian dalam studi berjudul "Artificial Intelligence untuk Pengolahan Data Tidak Terstruktur" akan mengadopsi pendekatan multidisipliner yang menggabungkan teknik pemrosesan bahasa alami (NLP), pembelajaran mesin, dan analisis data untuk mengekstraksi informasi dari data tidak terstruktur, khususnya dalam konteks catatan medis elektronik (EMR). Penelitian ini akan dilakukan dalam beberapa tahap yang mencakup pengumpulan data, pemrosesan data, pengembangan model, dan evaluasi hasil.

Pertama, tahap pengumpulan data akan melibatkan pengumpulan catatan medis elektronik yang berisi informasi tidak terstruktur. Data ini akan diambil dari berbagai sumber, termasuk rumah sakit dan lembaga kesehatan yang bersedia berpartisipasi, dengan memperhatikan aspek etika dan privasi pasien (Laparra et al., 2021). Data yang dikumpulkan akan mencakup catatan kemajuan, ringkasan keluar, dan laporan konsultasi, yang semuanya merupakan contoh data tidak terstruktur yang kaya akan informasi (Kreimeyer et al., 2017).

Setelah pengumpulan data, tahap berikutnya adalah pemrosesan data. Pada tahap ini, teknik NLP akan diterapkan untuk membersihkan dan memformat data. Proses ini mencakup penghapusan informasi yang tidak relevan, normalisasi teks, dan pengidentifikasian entitas penting menggunakan algoritma seperti *Named Entity Recognition* (NER) (Baviskar et al., 2021). Metode ini telah terbukti efektif dalam mengekstraksi informasi klinis yang relevan dari catatan medis (Almuhana & Abbas, 2022). Selain itu, fitur-fitur seperti bag-of-words,

fitur ortografis, dan fitur leksikal akan diekstraksi untuk mendukung pendekatan berbasis pembelajaran (Adnan & Akbar, 2019).

Selanjutnya, model pembelajaran mesin akan dikembangkan untuk menganalisis data yang telah diproses. Berbagai algoritma, termasuk regresi, jaringan syaraf, dan model berbasis ensemble, akan digunakan untuk membangun model yang mampu mengidentifikasi pola dan hubungan dalam data (Dura et al., 2022). Transfer learning juga akan diterapkan untuk memanfaatkan model yang telah dilatih sebelumnya pada domain lain, sehingga meningkatkan akurasi dan efisiensi dalam pengolahan data tidak terstruktur (Laparra et al., 2021; Tayefi et al., 2021).

Akhirnya, tahap evaluasi akan dilakukan untuk menilai kinerja model yang dikembangkan. Metode evaluasi akan mencakup pengukuran akurasi, sensitivitas, dan spesifisitas model dalam mengekstraksi informasi dari catatan medis. Hasil dari evaluasi ini akan dibandingkan dengan metode tradisional untuk menentukan efektivitas pendekatan yang diusulkan (Wu et al., 2018). Penelitian ini diharapkan dapat memberikan wawasan baru tentang bagaimana AI dapat digunakan untuk meningkatkan pengolahan data tidak terstruktur dalam konteks kesehatan, serta mengidentifikasi tantangan dan peluang yang ada dalam penerapan teknologi ini (Papadimitriou & Jurafsky, 2020).

Dengan pendekatan yang sistematis dan terintegrasi, penelitian ini bertujuan untuk memberikan kontribusi signifikan terhadap pemahaman dan penerapan kecerdasan buatan dalam pengolahan data tidak terstruktur, serta untuk meningkatkan kualitas informasi yang dapat diambil dari catatan medis elektronik.

HASIL DAN PEMBAHASAN

Hasil penelitian ini menunjukkan bahwa penerapan kecerdasan buatan (AI) dalam pengolahan data tidak terstruktur memiliki potensi yang signifikan untuk meningkatkan efisiensi dan efektivitas dalam berbagai bidang, terutama dalam pengelolaan informasi yang kompleks. Data tidak terstruktur, seperti teks dari catatan kesehatan elektronik (EHR), media sosial, dan dokumen lainnya, sering kali sulit untuk dianalisis dengan metode tradisional. Oleh karena itu, penelitian ini menekankan pentingnya teknik-teknik seperti pemrosesan

bahasa alami (NLP) dan pembelajaran mesin (machine learning) untuk mengekstrak informasi yang relevan dari data tersebut.

Salah satu temuan utama dari penelitian ini adalah bahwa metode AI dapat secara otomatis mengekstrak informasi kunci dari dokumen yang tidak terstruktur, yang sebelumnya memerlukan banyak waktu dan tenaga untuk diproses secara manual. Misalnya, penggunaan model pembelajaran mendalam (*deep learning*) dan teknik pengenalan entitas bernama (*named entity recognition*) telah terbukti efektif dalam mengidentifikasi dan mengklasifikasikan informasi penting dalam catatan kesehatan, yang dapat meningkatkan akurasi diagnosis dan hasil perawatan pasien. Selain itu, penelitian ini juga menunjukkan bahwa integrasi AI dalam pengolahan data tidak terstruktur dapat membantu mengatasi tantangan yang terkait dengan data yang heterogen dan tidak konsisten, seperti yang sering ditemukan dalam catatan klinis.

Penelitian menunjukkan bahwa AI, khususnya melalui metode pemrosesan bahasa alami (NLP) dan pembelajaran mesin, dapat secara signifikan meningkatkan kemampuan untuk mengekstrak informasi dari data yang tidak terstruktur, seperti catatan kesehatan elektronik (EHR) dan dokumen lainnya yang ditulis dalam bahasa alami. Salah satu teknik yang paling banyak digunakan dalam pengolahan data tidak terstruktur adalah *Named Entity Recognition* (NER), yang bertujuan untuk mengidentifikasi dan mengklasifikasikan entitas penting dalam teks, seperti nama pasien, diagnosis, dan prosedur medis. Penelitian menunjukkan bahwa sekitar 80% data dalam EHR berada dalam format tidak terstruktur, sehingga penerapan NER menjadi sangat penting untuk mengubah data tersebut menjadi informasi yang dapat digunakan untuk analisis lebih lanjut. Namun, tantangan yang dihadapi dalam penerapan NER termasuk variasi dalam gaya penulisan dan terminologi yang digunakan oleh berbagai profesional kesehatan, yang dapat mempengaruhi akurasi ekstraksi informasi. Selain NER, teknik lain yang relevan adalah penggunaan algoritma pembelajaran mesin untuk mengidentifikasi pola dalam data tidak terstruktur. Metode otomatis dapat digunakan untuk mengekstrak klasifikasi *New York Heart Association* (NYHA) dari catatan klinis, yang dapat memberikan wawasan berharga tentang respons pasien terhadap terapi tertentu. Hal ini menunjukkan bahwa AI tidak hanya membantu dalam ekstraksi data tetapi juga dalam pemahaman yang lebih dalam tentang kondisi kesehatan pasien.

Penelitian ini juga mengidentifikasi beberapa tantangan yang dihadapi dalam penerapan AI untuk pengolahan data tidak terstruktur. Salah satunya adalah kebutuhan untuk melakukan preprocessing data yang efektif agar data yang digunakan dalam analisis dapat diolah dengan baik. Proses ini mencakup pembersihan data, normalisasi, dan penghapusan duplikasi, yang semuanya penting untuk memastikan bahwa model AI dapat berfungsi dengan optimal. Selain itu, tantangan lain yang dihadapi adalah interpretabilitas dari model AI yang digunakan, yang merupakan aspek penting dalam konteks medis dan bisnis, di mana keputusan yang diambil berdasarkan analisis data dapat memiliki dampak yang signifikan. Tanpa preprocessing yang tepat, kualitas data yang dihasilkan dapat terpengaruh, yang pada gilirannya dapat memengaruhi hasil analisis dan pengambilan keputusan.

Salah satu temuan utama dari penelitian ini adalah kemampuan AI untuk mengotomatiskan proses ekstraksi informasi dari data yang kompleks. Dalam konteks kesehatan, misalnya, AI dapat menganalisis catatan kesehatan elektronik (EHR) yang sering kali berisi teks bebas yang tidak terstruktur. Dengan menggunakan model pembelajaran mendalam, AI dapat mengidentifikasi pola dan entitas penting dalam data tersebut, yang sebelumnya memerlukan analisis manual yang memakan waktu dan rentan terhadap kesalahan. Selain itu, interpretabilitas model AI juga menjadi isu penting, terutama dalam konteks medis. Penggunaan model yang kompleks seperti *deep learning* dapat menghasilkan hasil yang akurat, tetapi sulit untuk dipahami oleh praktisi medis. Hal ini menjadi tantangan tersendiri, terutama ketika model yang digunakan adalah model pembelajaran mendalam yang sering kali dianggap sebagai "kotak hitam". Oleh karena itu, penting untuk mengembangkan model yang tidak hanya akurat tetapi juga dapat dijelaskan, sehingga profesional kesehatan dapat memahami dasar dari rekomendasi yang diberikan oleh sistem.

Pembahasan mendalam dari penelitian mengenai penerapan kecerdasan buatan (AI) dalam pengolahan data tidak terstruktur menunjukkan bahwa tantangan dan peluang yang dihadapi dalam konteks ini sangat kompleks. Penelitian ini menyoroti pentingnya teknik pemrosesan bahasa alami (NLP) dan pembelajaran mesin (*machine learning*) dalam mengekstrak informasi dari data tidak terstruktur, yang sering kali mencakup teks, gambar, audio, dan video. Salah satu aspek kunci yang diangkat dalam penelitian ini adalah volume besar data tidak terstruktur yang dihasilkan di era big data. Data ini tidak hanya beragam

dalam bentuk, tetapi juga dalam sumber dan konteks, sehingga mempersulit proses ekstraksi informasi. Banyak sistem ekstraksi informasi (IE) yang ada saat ini masih terbatas pada satu jenis data, seperti teks atau gambar, dan belum mampu menangani berbagai jenis data secara bersamaan. Hal ini menunjukkan adanya kebutuhan mendesak untuk mengembangkan teknik IE yang lebih canggih yang dapat mengatasi heterogenitas, dimensi, dan volume data tidak terstruktur yang terus berkembang.

Dalam konteks aplikasi praktis, penelitian ini menunjukkan bahwa AI dapat digunakan untuk meningkatkan pengambilan keputusan dalam berbagai sektor, termasuk kesehatan, keuangan, dan pemasaran. Misalnya, dalam sektor kesehatan, AI dapat membantu dalam analisis data pasien untuk mengidentifikasi pola penyakit dan meramalkan hasil perawatan, sedangkan dalam sektor keuangan, AI dapat digunakan untuk menganalisis data transaksi dan mendeteksi penipuan. Dengan demikian, penerapan AI dalam pengolahan data tidak terstruktur tidak hanya meningkatkan efisiensi operasional tetapi juga memberikan wawasan yang lebih mendalam untuk pengambilan keputusan yang lebih baik.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa penerapan AI dalam pengolahan data tidak terstruktur memiliki potensi yang besar untuk mentransformasi cara organisasi mengelola dan menganalisis informasi. Dengan terus mengembangkan dan menerapkan teknologi AI, organisasi dapat meningkatkan kemampuan mereka untuk menangani volume data yang terus meningkat dan kompleksitas yang terkait, serta meningkatkan kualitas keputusan yang diambil berdasarkan data tersebut. Penelitian ini juga mendorong perlunya kolaborasi antara peneliti, praktisi, dan pembuat kebijakan untuk mengatasi tantangan yang ada dan memaksimalkan manfaat dari penerapan AI dalam pengolahan data tidak terstruktur. Dengan demikian, penelitian ini tidak hanya memberikan wawasan tentang potensi AI dalam pengolahan data tidak terstruktur, tetapi juga menggarisbawahi perlunya pendekatan yang lebih holistik dan kolaboratif untuk mengatasi tantangan yang dihadapi dalam bidang ini.

KESIMPULAN DAN REKOMENDASI

Kesimpulan dari penelitian ini menegaskan bahwa penerapan kecerdasan buatan (AI) dalam pengolahan data tidak terstruktur memiliki dampak yang signifikan dan menjanjikan dalam meningkatkan efisiensi serta efektivitas pengelolaan informasi di berbagai sektor. Penelitian ini menunjukkan bahwa teknik-teknik seperti pemrosesan bahasa alami (NLP) dan pembelajaran mesin (*machine learning*) dapat secara efektif digunakan untuk mengekstrak informasi yang relevan dari data tidak terstruktur, yang sebelumnya sulit untuk dianalisis dengan metode tradisional.

Hasil penelitian menunjukkan bahwa AI tidak hanya mampu mengotomatiskan proses ekstraksi informasi, tetapi juga meningkatkan akurasi dan kecepatan analisis data, yang sangat penting dalam konteks pengambilan keputusan di bidang kesehatan, keuangan, dan pemasaran. Dengan menggunakan model pembelajaran mendalam dan algoritma canggih, AI dapat mengidentifikasi pola dan hubungan dalam data yang kompleks, sehingga memberikan wawasan yang lebih mendalam untuk pengambilan keputusan yang lebih baik.

Namun, penelitian ini juga mengidentifikasi beberapa tantangan yang perlu diatasi, seperti kebutuhan untuk melakukan preprocessing data yang efektif dan memastikan interpretabilitas dari model AI yang digunakan. Proses preprocessing yang baik sangat penting untuk memastikan bahwa data yang digunakan dalam analisis telah bersih dan siap untuk diproses lebih lanjut. Selain itu, interpretabilitas model AI menjadi aspek penting, terutama dalam konteks medis dan bisnis, di mana keputusan yang diambil berdasarkan analisis data dapat memiliki dampak yang signifikan.

Secara keseluruhan, penelitian ini menyimpulkan bahwa penerapan AI dalam pengolahan data tidak terstruktur tidak hanya meningkatkan efisiensi operasional tetapi juga memberikan peluang untuk inovasi dalam pengambilan keputusan. Oleh karena itu, kolaborasi antara peneliti, praktisi, dan pembuat kebijakan sangat diperlukan untuk mengatasi tantangan yang ada dan memaksimalkan manfaat dari penerapan AI dalam pengolahan data tidak terstruktur. Penelitian ini mendorong pengembangan lebih lanjut dalam teknologi AI untuk memastikan bahwa organisasi dapat menghadapi tantangan data yang terus berkembang dan kompleksitas yang terkait, serta meningkatkan kualitas keputusan yang diambil berdasarkan data tersebut.

REFERENSI

- Adnan, K. and Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, 11. <https://doi.org/10.1177/1847979019890771>
- Ahn, J., Connell, A., Simonetto, D., Hughes, C., & Shah, V. (2021). Application of artificial intelligence for the diagnosis and treatment of liver diseases. *Hepatology*, 73(6), 2546-2563. <https://doi.org/10.1002/hep.31603>
- Almuhana, H. and Abbas, H. (2022). Classification of specialities in textual medical reports based on natural language processing and feature selection. *Indonesian Journal of Electrical Engineering and Computer Science*, 27(1), 163. <https://doi.org/10.11591/ijeecs.v27.i1.pp163-170>
- Baviskar, D., Ahirrao, S., & Kotecha, K. (2021). Multi-layout invoice document dataset (midd): a dataset for named entity recognition. *Data*, 6(7), 78. <https://doi.org/10.3390/data6070078>
- Baviskar, D., Ahirrao, S., Potdar, V., & Kotecha, K. (2021). Efficient automated processing of the unstructured documents using artificial intelligence: a systematic literature review and future directions. *IEEE Access*, 9, 72894-72936. <https://doi.org/10.1109/access.2021.3072900>
- Dura, B., Jean, C., Tannier, X., Calliger, A., Bey, R., Neuraz, A., ... & Flicoteaux, R. (2022). Learning structures of the french clinical language: development and validation of word embedding models using 21 million clinical reports from electronic health records.. <https://doi.org/10.48550/arxiv.2207.12940>
- Koleck, T., Dreisbach, C., Bourne, P., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 26(4), 364-379. <https://doi.org/10.1093/jamia/ocy173>
- Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S., ... & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of Biomedical Informatics*, 73, 14-29. <https://doi.org/10.1016/j.jbi.2017.07.012>

- Laparra, E., Mascio, A., Velupillai, S., & Miller, T. (2021). A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. *Yearbook of Medical Informatics*, 30(01), 239-244. <https://doi.org/10.1055/s-0041-1726522>
- More, A. (2024). Natural language processing - based structured data extraction from unstructured clinical notes. *Journal of Contemporary Medical Practice*, 6(8), 327-330. [https://doi.org/10.53469/jcmp.2024.06\(08\).67](https://doi.org/10.53469/jcmp.2024.06(08).67)
- Papadimitriou, I. and Jurafsky, D. (2020). Learning music helps you read: using transfer to study linguistic structure in language models. <https://doi.org/10.18653/v1/2020.emnlp-main.554>
- Rifaldi, D. and Fadlil, A. (2023). Teknik preprocessing pada text mining menggunakan data tweet “mental health”. *Decode Jurnal Pendidikan Teknologi Informasi*, 3(2), 161-171. <https://doi.org/10.51454/decode.v3i2.131>
- Riskin, D., Cady, R., Shroff, A., Hindiyeh, N., Smith, T., & Kymes, S. (2023). Using artificial intelligence to identify patients with migraine and associated symptoms and conditions within electronic health records. *BMC Medical Informatics and Decision Making*, 23(1). <https://doi.org/10.1186/s12911-023-02190-8>
- Samaras, A., Bekiaridou, A., Papazoglou, A., Moysidis, D., Tsoumakas, G., Bamidis, P., ... & Giannakoulas, G. (2023). Artificial intelligence-based mining of electronic health record data to accelerate the digital transformation of the national cardiovascular ecosystem: design protocol of the cardiominig study. *BMJ Open*, 13(4), e068698. <https://doi.org/10.1136/bmjopen-2022-068698>
- Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., ... & Godtliebsen, F. (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews Computational Statistics*, 13(6). <https://doi.org/10.1002/wics.1549>
- Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., ... & Liu, H. (2018). Clinical information extraction applications: a literature review. *Journal of Biomedical Informatics*, 77, 34-49. <https://doi.org/10.1016/j.jbi.2017.11.011>

Wu, H., Toti, G., Morley, K., Ibrahim, Z., Folarin, A., Jackson, R., ... & Dobson, R. (2018). Semehr: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research*. *Journal of the American Medical Informatics Association*, 25(5), 530-537. <https://doi.org/10.1093/jamia/ocx160>