Available at https://jtsiskom.undip.ac.id (26 August 2020)
DOI:10.14710/jtsiskom.2020.13686

Jurnal Teknologi dan Sistem Komputer, 8(4), 2020, 263-269

Kombinasi metode NER-OCR untuk meningkatkan efisiensi pengambilan informasi di poster berbahasa Indonesia

Combining the NER-OCR methods to improve information retrieval efficiency in the Indonesian posters

Ahmad Syarif Rosidy, Tubagus Mohammad Akhriza*), Mochammad Husni

Program Studi Sistem Informasi, STMIK PPKIA Pradnya Paramita
Jalan Laksda Adi Sucipto No.249A, Malang, Jawa Timur, Indonesia 65125

Cara sitasi: A. S. Rosidy, T. M. Akhriza, and M. Husni, "Kombinasi metode NER-OCR untuk meningkatkan efisiensi pengambilan informasi di poster berbahasa Indonesia," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 4, pp. 263-269, 2020. doi: 10.14710/jtsiskom.2020.13686, [Online].

Abstract – Event organizers in Indonesia often use websites to disseminate information about these events through digital posters. However, manually processing for transferring information from posters to websites is constrained by time efficiency, given the increasing number of posters uploaded. Also, information retrieval methods, such as Named Entity Recognition (NER) for Indonesian posters, are still rarely discussed in the literature. In contrast, the NER method application to Indonesian corpus is challenged by accuracy improvement because Indonesian is a low-resource language that causes a lack of corpus availability as a reference. This study proposes a solution to improve the efficiency of information extraction time from digital posters. The proposed solution is a combination of the NER method with the Optical Character Recognition (OCR) method to recognize text on posters developed with the support of relevant training data corpus to improve accuracy. The experimental results show that the system can increase time efficiency by 94 % with 82-92 % accuracy for several extracted information entities from 50 testing digital posters.

Keywords *– digital posters; information retrieval; named entity recognition; optical character recognition*

Abstrak – Penyelenggara acara di Indonesia seringkali menggunakan situs web untuk menyebarkan informasi tentang acara tersebut melalui poster digital. Namun, proses mentransfer informasi dari poster ke situs web secara manual terkendala oleh efisiensi waktu, mengingat makin banyaknya poster yang diunggah. Di sisi lain, metode pengambilan informasi berbasis teknologi informasi, seperti Named Entity Recognition (NER), untuk poster berbahasa Indonesia masih jarang dibahas di literatur, sedangkan penerapan NER terhadap korpus berbahasa Indonesia ditantang dalam peningkatan akurasi karena bahasa Indonesia adalah bahasa dengan sumber daya rendah yang menyebabkan minimnya ketersediaan korpus sebagai referensi. Artikel ini mengusulkan solusi untuk meningkatkan efisiensi waktu ekstraksi informasi dari poster digital. Solusi yang diusulkan merupakan kombinasi antara metode NER dengan Optical Character Recognition (OCR) untuk mengenali teks di poster yang dikembangkan dengan dukungan korpus data latih yang relevan untuk meningkatkan akurasi. Hasil percobaan menunjukkan bahwa sistem mampu meningkatkan efisiensi waktu sebesar 94 % dengan akurasi 82-92 % untuk beberapa entitas informasi yang diekstraksi untuk 50 poster digital uji.

Kata kunci *– poster digital; pengambilan informasi; named entity recognition; optical character recognition*

I. PENDAHULUAN

Manusia mengadakan berbagai acara sebagai bentuk pemenuhan kebutuhan dasar berinteraksi sosial [1]. Suatu acara sering diumumkan melalui poster untuk menarik lebih banyak partisipan. Di era informasi digital saat ini, poster banyak dibagikan dalam format gambar digital melalui situs web, seperti HaiEvent.com, JadwalEvent.web.id, JadwalKajian.com, dan Kajian Muslim.com. Keberadaan situs web memudahkan pencarian acara berdasarkan entitas tertentu, seperti nama pembicara, tempat, kota, dan tanggal acara. Fitur pencarian dapat digunakan karena informasi acara yang terkandung di dalam poster telah diidentifikasi dan dimasukkan kembali ke situs web oleh pengelola situs.

Di balik kemudahan pencarian informasi yang disediakan oleh situs web, para pengelola masih menghadapi kendala dalam proses mentransfer informasi dari poster. Metode manual yang selama ini dilakukan masih kurang efisien dimana pemrosesannya membutuhkan waktu rata-rata 149,18 detik/poster, padahal jumlah poster yang diunggah makin banyak. Secara teknis, proses pada metode manual dimulai dengan membaca setiap susunan kata di dalam poster dengan beragam tata letak. Setelah itu, susunan kata tersebut diklasifikasi ke dalam entitas nama orang,

*) Penulis korespondensi (Tubagus M. Akhriza)
Email: akhriza@stimata.ac.id

Submitted: 27 February 2020; Revised: 9 July 2020; Accepted: 10 July 2020; Published: 31 October 2020

tempat, kota atau tanggal. Terakhir, informasi yang telah diperoleh harus diketik ulang ke dalam suatu formulir.

Pengambilan entitas informasi berbasis teknologi informasi banyak dilakukan menggunakan metode *Named Entity Recognition* (NER). NER pada penelitian sebelumnya mampu digunakan dalam pengambilan informasi di kartu nama [2], video tutorial [3], teks artikel [4], teks unggahan di media sosial [5]–[7], dan informasi entitas di dalam rekaman data BTS [8]. Saat ini, berbagai *tools* yang telah tersedia dapat digunakan untuk menerapkan metode ini. Kushol dkk. [2] telah berhasil mengambil informasi kontak dari kartu nama menggunakan *tools* Apache OpenNLP, sedangkan Baidwaik dkk. [3] berhasil mengidentifikasi topik relevan dari video tutorial menggunakan *tools* Stanford CoreNLP.

Namun demikian, performa NER-*tools* pada poster berbahasa Indonesia masih mengalami kendala akurasi [9]–[11] karena bahasa Indonesia termasuk kategori bahasa dengan sumber daya rendah (*low-resource language*), seperti halnya bahasa Bengali [4] dan Cina [12]. Salah satu penyebabnya adalah karena bahasa Indonesia bukan bahasa internasional seperti bahasa Inggris atau Prancis yang sudah memiliki korpus teks dengan jutaan perbendaharaan kata dan tersedia di Internet [9].

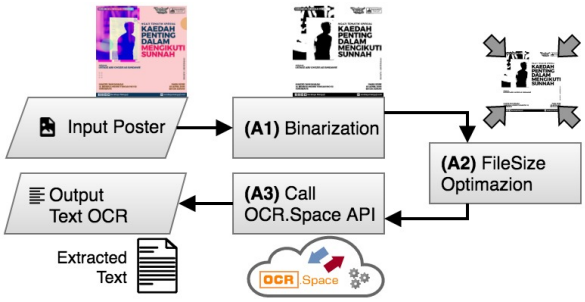
Beberapa literatur menyarankan pengembangan algoritme NER untuk cakupan yang lebih spesifik sebagai solusi atas keterbatasan performa NER-*tools* [5], [6]. Ritter dkk. [6] menjelaskan bahwa pengembangan dalam proses pelabelan *part-of-speech* (POS *tagging*), *chunking*, *kapitalisasi*, hingga segmentasi dan klasifikasi *Named-Entity* (NE) mampu meningkatkan performa F1 sebesar 25% dari dari *tools* Stanford NER terhadap data Twitter. Namun, data tersebut juga berbahasa Inggris.

Lebih jauh, studi literatur masih belum menemukan penelitian yang signifikan mengenai ekstraksi informasi dari poster digital berbahasa Indonesia. Kajian mengenai NER yang ada masih berfokus pada korpus teks dari internet seperti artikel berita daring [13] dan Wikipedia [9]–[11].

Artikel ini mengusulkan solusi untuk meningkatkan efisiensi waktu ekstraksi informasi dari poster digital melalui pengembangan metode pengambilan informasi dari poster digital berbahasa Indonesia berbasis NER yang dikombinasikan dengan metode *Optical Character Recognition* (OCR), suatu teknik membaca teks di dalam format digital [2], [3], [14], [15]. Metode NER yang dikembangkan didukung dengan korpus data latih yang relevan dengan poster berbahasa Indonesia untuk meningkatkan akurasi. Performa metode usulan tersebut dievaluasi dengan membandingkannya dengan metode manual dalam aspek efisiensi waktu pemrosesan dan akurasi hasil pengambilan informasi.

II. METODE PENELITIAN

Pengambilan informasi pada penelitian ini dibagi menjadi dua proses utama, yaitu proses OCR dan proses NER. Proses OCR berfungsi mengenali teks dari poster,



Gambar 1. *Pipeline* proses OCR

sedangkan NER berfungsi mengambil informasi dari teks hasil OCR.

Proses OCR merupakan tahap pra-pemrosesan bagi proses NER, yaitu mendapatkan teks dari poster. Masukan proses ini berupa poster dengan format gambar digital dan menghasilkan keluaran berupa teks hasil ekstraksi. Susunan *pipeline* dari masukan hingga keluaran disajikan pada Gambar 1.

Proses OCR dimulai dari tahap pra-pemrosesan *binarization* (A1) dengan masukan berupa poster mentah dengan format gambar digital. *Binarization* tersebut menghasilkan poster berlatar belakang putih dengan teks berwarna hitam.

Poster hasil *binarization* digunakan sebagai masukan pada tahap pra-pemrosesan berikutnya, yaitu optimasi *filesize* atau ukuran berkas (A2). Tahap ini dilakukan pemeriksaan ukuran berkas gambar hasil *binarization*. Jika berkas melebihi batas maksimal layanan OCR.Space API sebesar 1 MB [16], maka ukuran berkas gambar diperkecil.

Berkas gambar yang telah memenuhi syarat digunakan sebagai masukan proses pemanggilan layanan OCR.Space API (A3). Layanan ini memberikan respons dalam format JSON berisi teks hasil ekstraksi dari gambar digital [16]. Teks hasil ektraksi ini menjadi masukan pada *pipeline* proses NER.

Proses NER merupakan metode inti dalam penelitian ini. Proses ini mengolah masukan berupa teks hasil ekstraksi OCR untuk mendap47atkan keluaran akhir berupa entitas informasi nama pembicara, tempat, kota , dan tanggal acara. Kinerja NER-tools dapat ditingkatkan melalui normalisasi teks data uji dan menggunakan data latih dalam jumlah besar. Normalisasi dilakukan dengan melacak kata nonstandar untuk digantikan dengan kata standar yang sesuai [7], [17], sedangkan data latih dalam jumlah besar dapat diperoleh dengan mengolah artikel yang relevan dengan cakupan kasus [4], [9]–[11], [18], [19].

Berdasarkan kajian tersebut, tahapan proses NER disusun sebagai sebuah *pipeline* pada Gambar 2 yang dimulai dengan tahap normalisasi teks (B1). Tahap ini berfungsi mendeteksi kesalahan ekstraksi teks OCR dan mendeteksi kata-kata nonstandar untuk digantikan dengan kata dalam kamus. Kesalahan teks dideteksi menggunakan algoritme Levenshtein [20]. Algoritme Levenshtein menghitung tingkat perbedaan setiap kata hasil ekstraksi dengan kata pada kamus. Setiap kata hasil ekstraksi yang memiliki tingkat perbedaan kurang dari 2

digantikan dengan kata standar pada kamus. Proses normalisasi selanjutnya adalah *transform cases* dan *filter tokenize* untuk meminimalisir kata yang tidak berarti [21]. Contoh normalisasi ditunjukkan pada Tabel 1.

Teks yang telah melalui normalisasi digunakan sebagai masukan pada tahap tokenisasi kata (B2). Tokenisasi memecah kalimat menjadi potongan-potongan kata penyusunnya. Teks hasil normalisasi dipecah berdasarkan pemisah spasi dan tanda baca. Kumpulan kata hasil tokenisasi harus tetap memperhatikan urutan.

Kata-kata hasil tokenisasi kemudian diberi label (*tag*) dalam tahap POS *tagging* (B3). Label diberikan sesuai dengan daftar *tag* pada data latih yang relevan dengan poster acara berbahasa Indonesia untuk mendapatkan akurasi yang baik. Pada penelitian ini menggunakan 1367 kata berlabel yang diolah dari teks informasi yang tersedia pada situs web KajianMuslim.com.

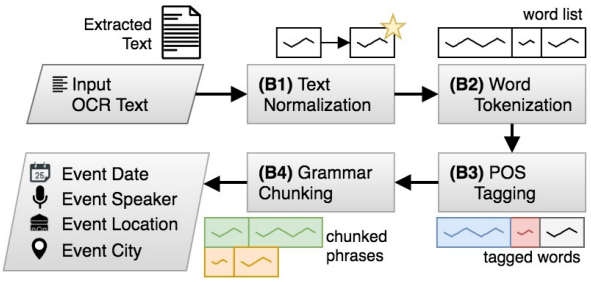
Setelah melewati tahap (B3), kata-kata pada poster mendapat label, yaitu sebagai berikut : {NUMx} berupa angka dengan x menunjukkan jumlah digit, {UP} berupa awalan nama orang, {UN} berupa kata penyusun nama orang, {US} berupa akhiran nama orang, {MN} berupa nama bulan, {MH} berupa nama bulan hijriah, {PP} berupa awalan nama tempat, {PN} kata penyusun nama tempat, {CP} berupa awalan nama kota, {CN} berupa kata penyusun nama kota, dan {CS} berupa akhiran nama kota. Contoh hasil pelabelan kata ditunjukkan pada Tabel 2.

Hasil pelabelan pada setiap kata menjadi masukan dalam tahap *grammar chunking* (B4). Tahap ini melakukan penguraian potongan frasa berdasarkan tata bahasa yang telah ditentukan dalam data latih. Data latih berupa aturan tata bahasa (*grammar rules*) merupakan susunan urutan kata berlabel beserta aturan kewajiban keberadaan kata berlabel tersebut. Penelitian ini menggunakan aturan kewajiban dengan tanda berikut: tanda bintang (*) artinya wajib ada satu kata berlabel, tanda plus (+) artinya wajib ada satu atau lebih kata berlabel, serta tanda tanya (?) atinya adanya kata berlabel bersifat opsional.

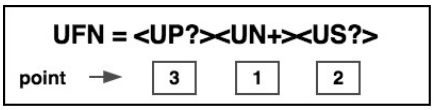
Contoh aturan tata bahasa ditunjukkan pada Gambar 3, dimana "UFN" merupakan contoh *grammar rule* untuk mendapatkan informasi nama pembicara. UFN memiliki aturan dengan urutan label kata "UP?", "UN+", dan kemudian "US?". Aturan "UP?" menunjukkan keberadaan kata berlabel "UP" di awal frasa bersifat opsional, "UN+" menunjukkan wajib ada minimal satu kata berlabel "UN", sedangkan "US?" menunjukkan bahwa keberadaan kata berlabel "US" bersifat opsional di akhir frasa.

Penguraian potongan frasa berdasarkan aturan "UFN" menghasilkan contoh frasa seperti ditunjukkan pada Gambar 4. Informasi nama pembicara diperoleh, yaitu "Dr Muhammad Nur Ihsan Lc Ma", yang tersusun dari satu kata berlabel "UP" di awal, tiga kata berlabel "UN" dan dua kata berlabel "US" di akhir.

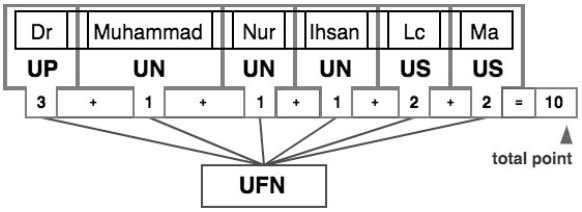
Grammar rule dalam penelitian ini memiliki poin penyusun yang berfungsi menentukan urutan prioritas



Gambar 2. *Pipeline* proses NER



Gambar 3. Contoh *grammar rule* nama pembicara



Gambar 4. Contoh hasil *grammar chunking*

Tabel 1. Contoh normalisasi teks

Asal Kata	Kata Pengganti
Sukdbumi	sukabumi
Mssjid	masjid
abduUah	abdullah
Pebruari	februari

Tabel 2. Contoh POS *tagging*

Kata	Label Kata (POS *tag*)
27	NUM2
April	MN
Dr	UP
Muhammad	UN

ketika ditemukan informasi multi-nilai dalam suatu poster. Sebagai contoh, Gambar 4 metampilkan perhitungan akumulasi poin informasi "Dr Muhammad Nur Ihsan Lc Ma" dan menghasilkan total poin 10. Jika dalam poster yang sama terdapat pula informasi nama pembicara "Muhammad Ali Lc" yang memiliki perhitungan poin 1+1+2=4, maka nama pembicara "Dr Muhammad Nur Ihsan Lc Ma" ditampilkan terlebih dahulu karena memiliki total poin lebih tinggi.

Eksperimen dilakukan untuk menguji performa metode yang diusulkan dibandingkan dengan dengan metode manual dalam aspek efisiensi waktu pemrosesan dan akurasi hasilnya untuk tiap poster. Perbandingan performa pada kedua metode dilakukan menggunakan sejumlah sampel poster yang sama. Jumlah minimal sampel yang diperlukan dalam penelitian eksperimental

menurut rekomendasi Fraenkel dkk. [22], yaitu sebanyak 30 sampel. Pada penelitian ini sampel poster uji yang diambil secara acak dari situs web KajianMuslim.com ditambah hingga sejumlah 50 agar semakin mewakili populasi poster.

Pengujian performa metode manual dilakukan oleh satu orang administrator portal web, yaitu dimulai dengan membaca setiap susunan kata dalam setiap poster uji. Setelah itu, susunan kata tersebut diklasifikasi ke dalam entitas nama orang, tempat, kota atau tanggal. Terakhir, informasi yang telah diperoleh diketik ulang ke dalam suatu formulir dalam situs web untuk disimpan di basisdata. Durasi pengujian metode manual dicatat sejak proses pembacaan kata dalam poster hingga informasi berhasil tersimpan. Informasi yang telah tersimpan digunakan sebagai acuan perhitungan akurasi hasil ekstraksi informasi menggunakan metode alternatif.

Algoritme metode alternatif yang diusulkan diimplementasi ke dalam aplikasi berbasis web yang digunakan sebagai alat pengujian performa. Aplikasi web dimaksud dikembangkan dengan bahasa pemrograman PHP dan dijalankan pada sistem operasi macOS High Sierra pada Laptop Intel Core i5 dengan spesifikasi 8 GB RAM dan kecepatan internet 10 Mbps. Korpus data latih untuk pengujian diperoleh dengan mengolah kata-kata di dalam keseluruhan populasi poster yang tersedia di situs web KajianMuslim.com. Pengujian akurasi dilakukan dengan menambah data latih untuk masing-masing kategori informasi secara bertahap, sedangkan durasi proses dicatat sejak poster diunggah oleh seorang administrator hingga informasi berhasil ditampilkan di aplikasi web.

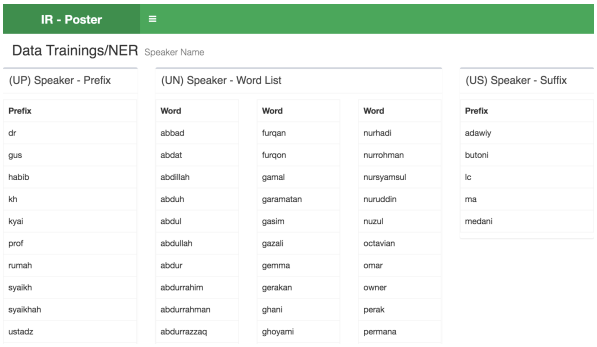
Akurasi hasil pengambilan informasi menggunakan metode alternatif pada setiap entitas nama pembicara, tempat, kota, dan tanggal acara ditunjukkan dengan persentase objek masukan yang dilabeli dengan benar [23]. Setiap poster yang diuji memiliki satu nilai pada setiap jenis entitas. Jumlah objek yang dilabeli dengan benar merupakan jumlah poster yang memiliki hasil ekstraksi informasi sesuai dengan hasil pengambilan informasi secara manual, sedangkan jumlah total objek merupakan jumlah total sampel poster uji. Rumus perhitungan akurasi dinyatakan pada Persamaan 1.

$$akurasi = \frac{Jumlah\ objek\ dilabeli\ benar}{Jumlah total\ objek} \quad (1)$$

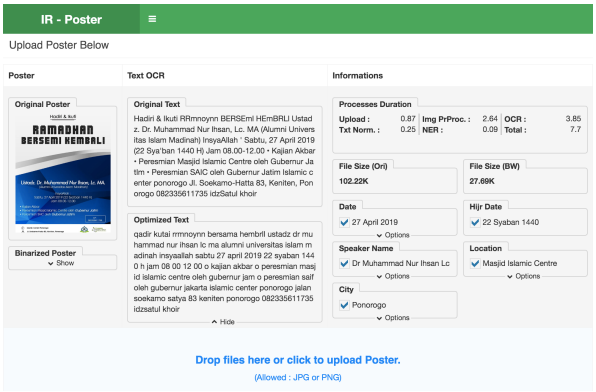
Aplikasi web yang dikembangkan dibagi menjadi dua halaman utama, yaitu halaman data latih dan halaman unggah poster. Contoh halaman data latih nama pembicara ditunjukkan pada Gambar 5 dan tampilan halaman unggah poster setelah informasi berhasil diambil ditunjukkan pada Gambar 6.

III. HASIL DAN PEMBAHASAN

Grafik perbandingan durasi pengambilan informasi menggunakan metode manual dan metode alternatif ditampilkan pada Gambar 7. Sejumlah 50 poster sebagai data uji memiliki ukuran berkas (*filesize*) bervariasi



Gambar 5. Halaman data latih nama pembicara



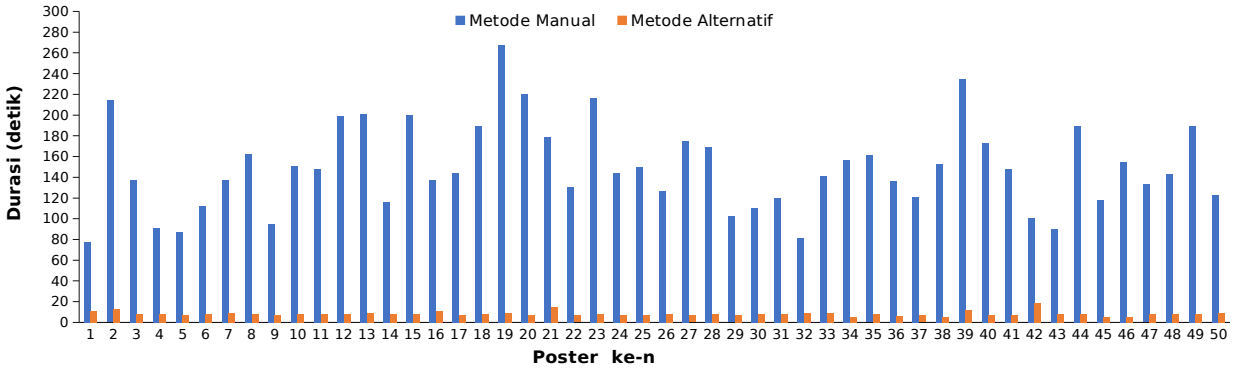
Gambar 6. Halaman unggah (*upload*) poster

antara 46-1197 KB dengan jumlah kata hasil ekstraksi OCR bervariasi antara 27-218 kata.

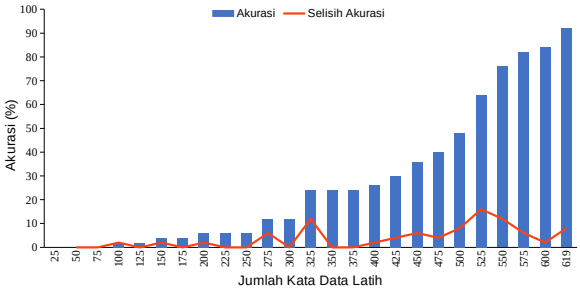
Berdasarkan data grafik perbandingan durasi pada Gambar 7, dapat diketahui bahwa pengambilan informasi pada poster menggunakan metode alternatif membutuhkan waktu antara 5,09-18,85 detik/poster dengan rata-rata 8,30 detik/poster. Dengan rincian rata-rata durasi proses OCR 8,04 detik/poster dan proses NER 0,26 detik/poster. Hal ini menunjukkan bahwa efisiensi waktu pengambilan dan pengenalan informasi dari poster dapat meningkat sebanyak 94 % dari metode manual yang membutuhkan waktu antara 81-268 detik/poster dengan rata-rata 149,18 detik/poster.

Hasil pengujian juga menunjukkan bahwa proses OCR menghabiskan lebih banyak waktu daripada proses NER. Hal ini dapat terjadi karena proses OCR melibatkan lebih banyak *library tools*, yaitu PHP GD untuk *binarization* dan optimasi *filesize*, serta layanan pihak ketiga OCR.Space API untuk ektraksi teks. Kecepatan proses layanan OCR.Space juga dipengaruhi kestabilan dan kecepatan jaringan internet, sedangkan di dalam proses NER hanya berisi pengolahan basis data, perulangan, dan algoritme pengolahan teks.

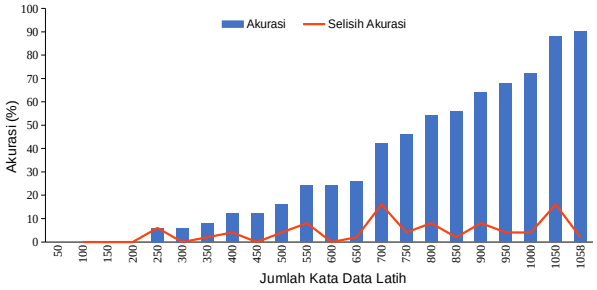
Hasil pengujian tersebut menunjukkan bahwa poster dengan variabel ukuran berkas dan jumlah kata yang bervariasi membutuhkan durasi pengambilan informasi yang bervariasi pula. Hubungan antar variabel tersebut diuji dengan menggunakan uji korelasi Pearson [24]. Uji korelasi menunjukkan koefisien sebagai berikut: ukuran berkas dengan durasi OCR 0,73; ukuran berkas dengan durasi NER 0,41; ukuran berkas dengan durasi total



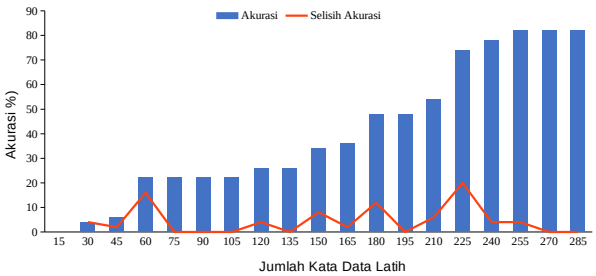
Gambar 7. Perbandingan durasi pengambilan informasi menggunakan metode manual dengan metode alternatif



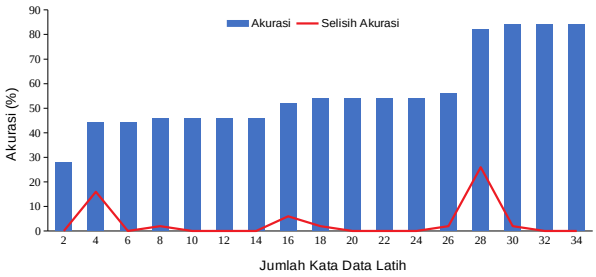
Gambar 8. Akurasi pengambilan informasi nama pembicara berdasarkan jumlah kata data latih



Gambar 9. Akurasi pengambilan informasi nama tempat berdasarkan jumlah kata data latih



Gambar 10. Akurasi pengambilan informasi nama kota berdasarkan jumlah kata data latih



Gambar 11. Akurasi pengambilan informasi tanggal berdasarkan jumlah kata data latih

0,73; jumlah kata dengan durasi OCR 0,54; jumlah kata dengan durasi NER 0,68; serta jumlah kata dengan durasi total 0,55.

Uji korelasi Pearson menunjukkan durasi proses OCR cenderung dipengaruhi besarnya ukuran berkas, sedangkan durasi proses NER cenderung dipengaruhi jumlah kata dalam poster. Hal ini menunjukkan bahwa proses OCR dapat dipercepat dengan mengoptimalkan ukuran berkas tanpa mengurangi kualitas piksel gambar. Pada penelitian selanjutnya dapat dilakukan perbandingan format berkas antara PNG dan JPG dari sisi optimasi ukuran berkas, sedangkan optimasi proses NER dapat dilakukan dengan mengurangi informasi-informasi yang tidak relevan dalam teks, misalnya dengan *stopword removal* dan *n-gram detection* [25].

Akurasi yang cukup signifikan juga berhasil dicapai oleh metode yang diusulkan. Pengujian akurasi dilakukan dengan menambah data latih untuk masing-masing entitas informasi secara bertahap. Tingkat akurasi dicatat dan divisualisasikan dalam grafik pada Gambar 8-Gambar 11.

Setiap tahap pengujian akurasi dalam pengambilan informasi nama pembicara dilakukan penambahan 25 kata berlabel UP, UN dan/atau US sebagai data latih. Setiap tahap menunjukkan tren positif pada grafik dengan tingkat kenaikan yang bervariasi antara 0-16%, sebagaimana ditunjukkan pada Gambar 8. Kenaikan akurasi yang cukup signifikan terjadi saat penambahan data latih dari 300 ke 325 (+12 %), dari 500 ke 525 (+16 %), serta dari 525 ke 550 (+12 %). Tahap akhir pengujian dengan 619 kata data latih berhasil mencapai akurasi 92 %.

Setiap tahap pengujian akurasi dalam pengambilan informasi nama tempat dilakukan penambahan 50 kata berlabel PP, PN, UN, MH dan/atau UP sebagai data latih. Setiap tahap juga menunjukkan tren positif pada grafik dengan tingkat kenaikan antara 0-16 %, sebagaimana ditunjukkan pada Gambar 9. Kenaikan akurasi yang cukup signifikan terjadi saat penambahan

data latih dari 600 ke 650 (+16 %), serta dari 1000 ke 1050 (+16 %). Tahap akhir pengujian dengan 1058 kata data latih berhasil mencapai akurasi 90 %.

Setiap tahap pengujian akurasi dalam pengambilan nama kota dengan penambahan 15 kata berlabel CP, CN dan/atau CS juga menunjukkan tren positif, seperti ditunjukkan pada Gambar 10. Setiap tahap pengujian hanya ditambahkan 15 kata data latih. Namun, sempat terjadi tingkat kenaikan akurasi yang lebih signifikan, yaitu +20 % pada 210 kata ke 225 kata. Tahap akhir pengujian dengan 285 data latih berhasil mencapai akurasi 82 %.

Berbeda dengan pengambilan entitas informasi lainnya yang membutuhkan data latih berjumlah ratusan, entitas tanggal acara hanya membutuhkan 34 kata data latih untuk mencapai akurasi 84 % (Gambar 11). Meskipun setiap tahap pengujian hanya ditambahkan data latih yang lebih sedikit, yaitu 2 kata, sempat terjadi tingkat kenaikan akurasi yang lebih signifikan, yaitu +26 % saat penambahan dari 26 ke 28 kata.

Semua grafik pengujian akurasi menunjukkan tren positif seiring penambahan data latih pada setiap entitas informasi yang diekstraksi. Hal ini senada dengan hasil penelitian [4], [9]–[11], [18], [19] yang menunjukkan bahwa kuantitas korpus data latih yang digunakan berperan pada akurasi hasil NER. Namun, relevansi data latih juga perlu diperhatikan selain kuantitas. Hal ini ditunjukkan pada variasi besarnya nilai peningkatan akurasi pada setiap tahap pengujian meskipun jumlah data latih yang ditambahkan sama. Sebagai contoh, pada entitas tanggal kajian penambahan 2 kata data latih pada 4-6, 8-14, 20-24 serta 32-34 sama sekali tidak memberi peningkatan akurasi, sedangkan pada 26-28 kata data latih peningkatan yang terjadi dapat mencapai 26 %, yaitu dari 56 % ke 82 %.

Berdasarkan analisis grafik tersebut, dapat diketahui adanya peran relevansi data latih terhadap akurasi NER. Oleh karena itu, penentuan sumber korpus data latih yang akan ditambahkan perlu dipertimbangkan dalam penelitian selanjutnya untuk mendapatkan akurasi tinggi. Jika aplikasi akan digunakan dalam pengambilan informasi poster acara keagamaan, maka menjadi kurang relevan ketika mengambil sumber korpus dari situs web atau artikel acara musik.

Korpus data latih memegang peran penting dari sisi kuantitas maupun relevansi terhadap akurasi. Namun, meskipun telah memiliki korpus data latih dengan kuantitas dan relevansi yang cukup baik, tidak serta merta akurasi 100 % dapat dicapai. Salah satu faktor yang menyebabkan kesalahan adalah adanya kesalahan ekstraksi teks dengan OCR. Pada penelitian selanjutnya dapat dilakukan optimasi pra-pemrosesan gambar yang lebih handal dan membandingkan hasil ektraksi teks menggunakan layanan OCR-*tools* lainnya.

IV. KESIMPULAN

Metode pengambilan informasi dari poster digital berbahasa Indonesia menggunakan NER dan OCR mampu menjadi solusi untuk meningkatkan efisiensi waktu pengambilan informasi sebanyak 94 % dari proses metode manual yang rata-rata membutuhkan 149,18 detik/poster menjadi 8,30 detik/poster. Akurasi hasil pengambilan informasi 82-92 % pada setiap entitas informasi yang diekstraksi boleh dikatakan signifikan.

Namun demikian, korpus data latih yang digunakan masih berasal dari satu sumber situs web dan poster data uji juga berasal dari sumber yang sama. Penelitian selanjutnya perlu mengembangkannya dengan menambah data latih dalam jumlah besar yang berasal dari berbagai sumber yang lebih bervariasi untuk peningkatan akurasi. Di sisi pengujian, perlu dilakukan validasi kinerja dengan poster uji yang lebih bervariasi pula. Lebih jauh, pada sisi optimasi kecepatan OCR dapat dilakukan perbandingan ukuran berkas poster antara PNG dan JPG, sedangkan untuk kecepatan proses NER dapat dioptimasi dengan mengurangi informasi-informasi yang tidak relevan dalam teks, misalnya dengan *stopword removal* dan *n-gram detection.*

DAFTAR PUSTAKA

[1] J. Armbrecht, E. Lundberg, and T. D. Andersson, *A research agenda for event management*. Edward Elgar Publishing, 2019. doi: 10.4337/9781788114363

[2] R. Kushol, I. Ahsan, and M. N. Raihan, "An Android-based useful text extraction framework using image and natural language processing," *International Journal of Computer Theory and Engineering*, vol. 10, no. 3, pp. 77-83, 2018. doi: 10.7763/IJCTE.2018.V10.1203

[3] K. Badwaik, K. Mahmood, and A. Raza, "Towards applying OCR and semantic web to achieve optimal learning experience," in *IEEE 13th International Symposium on Autonomous Decentralized Systems*, Bangkok, Thailand, Mar. 2017, pp. 262-267. doi: 10.1109/ISADS.2017.40

[4] A. Das, D. Ganguly, and U. Garain, "Named entity recognition with word embeddings and wikipedia categories for a low-resource language," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 16, no. 3:18, 2017. doi: 10.1145/3015467

[5] L. Derczynski et al., "Analysis of named entity recognition and linking for tweets," *Information Processing and Management*, vol. 51, no. 2, pp. 32-49, 2015. doi: 10.1016/j.ipm.2014.10.006

[6] A. Ritter, C. Sam, Mausam, and O. Etzioni, "Named entity recognition in tweets: An experimental study," in *Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, Jul. 2011, pp. 1524-1534.

[7] Li and Y. Liu, "Improving named entity recognition in tweets via detecting non-standard words," in *7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Beijing, China, Jul. 2015, pp. 929-938. doi: 10.3115/v1/P15-1090

[8] T. M. Akhriza, H. Y. Sahaduta, and A. D. Susilo, "Improving mobility of base transceiver station locating method using telegram's application," *International Journal of Technology*, vol. 8, no. 1, pp. 175-183, 2017. doi: 10.14716/ijtech.v8i1.6012

[9] I. Alfina, R. Manurung, and M. I. Fanany, "DBpedia entities expansion in automatically building dataset for Indonesian NER," in *2016 International Conference on Advanced Computer Science and Information Systems*, Malang, Indonesia, Oct. 2017, pp. 335-340. doi: 10.1109/ICACSIS.2016.7872784

[10] R. A. Leonandya, B. Distiawan, and N. H. Praptono, "A Semi-supervised algorithm for indonesian named entity recognition," in *3rd International Symposium on Computational and Business Intelligence*, Bali, Indonesia, Dec. 2015, pp. 45-50. doi: 10.1109/ISCBI.2015.15

[11] A. Luthfi, B. Distiawan, and R. Manurung, "Building an Indonesian named entity recognizer using Wikipedia and DBPedia," in *International Conference on Asian Language Processing*, Kuching, Malaysia, Oct. 2014, pp. 19-22. doi: 10.1109/IALP.2014.6973520

[12] N. Peng and M. Dredze, "Improving named entity recognition for Chinese social media with word segmentation representation learning," in *Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, Aug. 2016, pp. 149-155. doi: 10.18653/v1/P16-2025

[13] I. Budi and S. Bressan, "Application of association rules mining to Named Entity Recognition and co-reference resolution for the Indonesian language," *International Journal of Business Intelligence and Data Mining*, vol. 2, no. 4, pp. 426-446, 2007. doi: 10.1504/IJBIDM.2007.016382

[14] A. S. Agbenemu, J. Yankey, and E. O., "An automatic number plate recognition system using Opencv and Tesseract OCR engine," *International Journal of Computer Applications*, vol. 180, no. 43, pp. 1-5, 2018. doi: 10.5120/ijca2018917150

[15] A. E. Utami, O. D. Nurhayati, and K. T. Martono, "Aplikasi penerjemah bahasa Inggris - Indonesia dengan optical character recognition berbasis android," *Jurnal Teknologi dan Sistem Komputer*, vol. 4, no. 1, pp. 167-177, 2016. doi: 10.14710/jtsiskom.4.1.2016.167-177

[16] OCR.Space, "Free OCR API," 2019. [online]. Available: https://ocr.space/ocrapi

[17] Ç. Sönmez and A. Özgü, "A graph-based approach for contextual text normalization," in *Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, Oct. 2014, pp. 313-324. doi: 10.3115/v1/D14-1037

[18] H. D. M. Alfarohmi and M. A. Bijaksana, "Building the Indonesian NE dataset using Wikipedia and DBpedia with entities expansion method on DBpedia," in *International Conference on Asian Language Processing*, Bandung, Indonesia, Nov. 2018, pp. 334-339. doi: 10.1109/IALP.2018.8629117

[19] J. Ni and R. Florian, "Improving multilingual named entity recognition with wikipedia entity type mapping," in *Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Nov. 2016, pp. 1275-1284. doi: 10.18653/v1/D16-1135

[20] H. N. Abdulkhudhur, I. Q. Habeeb, Y. Yusof, and S. A. M. Yusof, "Implementation of improved Levenshtein algorithm for spelling correction word candidate list generation," *Journal of Theoretical and Applied Information Technology*, vol. 88, no. 3, pp. 449-455, 2016.

[21] A. T. J. Harjanta, "Preprocessing text untuk meminimalisir kata yang tidak berarti dalam proses text mining," *Jurnal Informatika UPGRIS*, vol. 1, no. 1, pp. 1-9, 2015.

[22] J. R. Fraenkel, N. E. Wallen, and H. H. Hyun, *How to design and evaluate research in education, 8th edition*. New York: McGraw-Hill, 2012.

[23] N. Quoc Viet Hung, N. T. Tam, L. N. Tran, and K. Aberer, "An evaluation of aggregation techniques in crowdsourcing," in *International Conference on Web Information Systems Engineering*, Nanjing, China, Oct. 2013, pp. 1-15. doi: 10.1007/978-3-642-41154-0_1

[24] J. Sarwono, *Metode penelitian kuantitatif dan kualitatif*. Yogyakarta: Graha Ilmu, 2006.

[25] H. Najjichah, A. Syukur, and H. Subagyo, "Pengaruh text preprocessing dan kombinasinya," *Jurnal Teknologi Informasi*, vol. 15, no. 1, pp. 1-11, 2019.