Terakreditasi SINTA Peringkat 2
Surat Keputusan DirjenPenguatanRiset dan PengembanganRistekDikti No. 10/E/KPT/2019
masa berlakumulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(RekayasaSistem dan TeknologiInformasi)
Vol. 4 No. 4 (2020) 704 - 710 ISSN Media Elektronik: 2580-0760

Pemanfaatan Optical Character Recognition Dan Text Feature Extraction Untuk Membangun Basisdata Pengaduan Tenaga Kerja

Yan Puspitarani[1], Yenie Syukriyah[2]
[1,2] Informatika, Teknik, Universitas Widyatama
[1]yan.puspitarani@widyatama.ac.id, [2]yenie.syukriyah@widyatama.ac.id

Abstract

The examination of complaints of labor violations is part of the main activity of the labor inspection section within the Department of Manpower. Monitors will examine companies that are considered to have violated labor laws based on a letter of complaint sent by the relevant union organization or legal aid agency. The easy way to communicate at this time, making the submission of complaint letters can be directly sent in the form of images through electronic media such as whatsapp or email. This makes it difficult for administrative staff to recapitulate incoming complaints because they have to read and enter data manually into the system. Therefore, this research was conducted to create a system that utilizes OCR technology and text feature extraction to be able to input complaints data automatically. This research resulted in a prototype of letter input and a database of letter storage that can be further utilized for Data Mining and Business Intelligent. OCR implementation is done by using the Tesseract library while the text feature selection utilizes the Natural Language Toolkit (NLTK) library. The results of testing of the prototype showed an accuracy of 66.7% of the OCR results and 91.67% of the manually typed letters.

Keywords: OCR, text feature extraction, database

Abstrak

Pemeriksaan terhadap pengaduan pelanggaran ketenagakerjaan merupakan bagian dari aktivitas utama bagian pengawasan ketenagakerjaan di lingkungan Dinas Ketenagakerjaan. Pengawas akan memeriksa perusahaan yang dianggap melanggar undang-undang ketenagakerjaan berdasarkan surat pengaduan yang dikirim oleh organisasi serikat pekerja atau lembaga bantuan hukum terkait. Mudahnya cara berkomunikasi saat ini, membuat penyampaian surat pengaduan dapat langsung dikirimkan berupa gambar melalui media elektronik seperti *whatsapp* atau *email*. Hal ini membuat pegawai administrasi kesulitan dalam melakukan rekapitulasi pengaduan yang masuk karena harus membaca dan memasukkan data secara manual ke dalam sistem. Oleh karena itu, penelitian ini dilakukan untuk membuat sistem yang memanfaatkan teknologi *OCR* dan *text feature extraction* agar dapat melakukan input data pengaduan secara otomatis. Penelitian ini menghasilkan *prototipe* input surat dan basisdata penyimpanan surat yang dapat dimanfaatkan lebih lanjut untuk *Data Mining* dan *Business Intelligent*. Implementasi *OCR* dilakukan dengan memanfaatkan library *Tesseract* sedangkan *text feature selection* memanfaatkan library *Natural Language Toolkit (NLTK)*. Hasil pengujian terhadap prototipe menunjukkan akurasi sebesar 66.7% terhadap surat hasil *OCR* dan 91.67% terhadap surat yang diketik manual.

Kata kunci: OCR, text feature extraction, basisdata

1. Pendahuluan

Pengawasan ketenagakerjaan bertujuan untuk memastikan terlaksananya norma ketenagakerjaan di tempat kerja. Salah satu aktivitas yang dilakukan dalam pengawasan ketenagakerjaan ini berupa pemeriksaan terhadap perusahaan secara berkala atau berdasarkan pengaduan pekerja/masyarakat. Pengaduan ini dapat disampaikan melalui berbagai cara seperti pesan singkat melalui jejaring sosial, mengirimkan surat melalui

Diterima Redaksi: 29-06-2020 | Selesai Revisi: 24-07-2020 | Diterbitkan Online: 20-08-2020

email, dan menyampaikan langsung ke dinas tenaga kerja terdekat.

Kemudahan berkomunikasi membuat penyampaian pengaduan pelanggaran ketenagakerjaan ini semakin mudah. Surat dapat langsung dikirimkan berupa gambar melalui media pesan *online* agar lebih cepat ditanggapi. Informasi yang didapatkan dari pengaduan ini dapat dimanfaatkan untuk analisis kasus-kasus serupa beserta penanganannya agar dapat dilakukan usaha pencegahan demi meningkatkan ketenangan bekerja dan berusaha bagi para tenaga kerja dan pengusaha. Akan tetapi, secara administratif, pegawai ketenagakerjaan kesulitan memasukkan data pengaduan ke dalam database terstruktur dikarenakan pengaduan yang dikirimkan berbentuk gambar. Sehingga diperlukan teknologi yang tepat untuk menangani permasalahan ini.

Teknologi *OCR (Optical Character Recognition)* dapat digunakan untuk mengubah bentuk gambar menjadi text. Selanjutnya, teknologi ekstraksi fitur pun dimanfaatkan untuk membuat data berbentuk text menjadi lebih terstruktur agar dapat dimasukkan ke dalam basisdata.

OCR adalah konversi dari hasil pindai atau cetak gambar berisi text dan tulisan tangan menjadi text yang dapat diedit untuk pemrosesan selanjutnya. Teknologi ini menggunakan sebuah mesin untuk mengenali text secara otomatis. Jika dianalogikan dengan bagian tubuh manusia, *OCR* seperti kombinasi dari mata dan pikiran. Mata dapat melihat text pada gambar sedangkan otak memproses interpretasi text yang dibaca oleh mata[1].

Beragam penelitian dengan memanfaatkan teknologi *OCR* telah banyak dilakukan. *Library OCR* berbasis *open source* dan komersial pun banyak dihasilkan. Salah satu library berbasis *open source* yang banyak digunakan adalah *Tesseract*. Dalam penelitian [1], Patel membandingkan akurasi *Tesseract OCR* dan *Transym OCR* (aplikasi komersial) terhadap plat nomor kendaraan bermotor. Hasilnya menunjukkan bahwa *Tesseract OCR* menghasilkan rata-rata akurasi 70% untuk gambar *gray scale* sedangkan *Transym OCR* menghasilkan rata-rata akurasi sebesar 47%. Jika dilihat dari sisi rata-rata waktu pemrosesan, *Tesseract OCR* jauh lebih cepat, yaitu 0.82 detik dibandingkan *Transym OCR* yang memakan waktu 6.75 detik. Selain itu, Algun membandingkan dua aplikasi *open source*, yaitu *Tesseract* dan *Kraken*. Penelitian dilakukan terhadap gambar buku berbahasa Arab. Hasilnya menunjukkan bahwa *Tesseract* mencapai akurasi 100% pada gambar dengan hasil segmentasi *foreground* dan *background* yang sangat jernih [2]. Dari penelitian tersebut menunjukkan bahwa *Tesseract* merupakan *library open source* yang memiliki performansi yang sangat baik.

Setelah gambar dikonversi menjadi text, dilakukan ekstraksi terhadap informasi berupa text menjadi bagian-bagian kecil sehingga data yang dihasilkan lebih terstruktur. Proses ini menjadi hal yang sangat mendasar dan penting untuk proses *text mining* dan *information retrieval* [3]. Pada kasus ini, *Text Feature Extraction* atau lebih umum dikenal dengan *Information Extraction*, akan menentukan bagian-bagian dari teks berupa informasi yang lebih terstruktur seperti tanggal surat, nama organisasi, pengirim surat, dan pengaduan yang dilaporkan. Dengan kata lain, informasi faktual terstruktur dari teks yang tidak terstruktur dapat diperoleh dengan menggunakan *Information Extraction* [4].

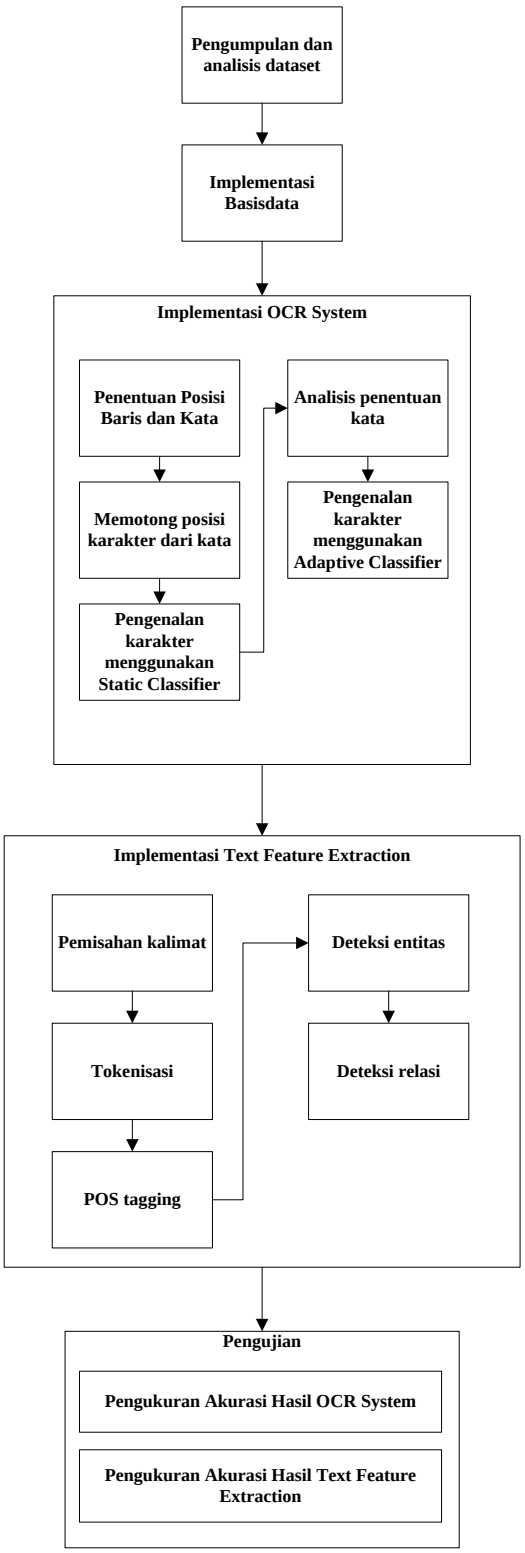
Salah satu modul *NLP* yang dapat dimanfaatkan untuk information extraction adalah *NLTK(Natural Language Toolkit)* [5]. Dalam penelitian yang pernah dilakukan, Christianti [6] melakukan tagging terhadap dokumen bahasa Indonesia menggunakan *Brill Tagger* milik *NLTK* dan menghasilkan akurasi sebesar 96.65% tanpa pemeriksaan aturan leksikal Bahasa Indonesia. Setelah dilakukan penyesuaian secara leksikal dan kontekstual, akurasi pun meningkat menjadi 99.75%. Selanjutnya, Yumusak [7] melakukan perbandingan akurasi POS tagging menggunakan *NLTK* terhadap beberapa jenis *Corpus*. Hasilnya menunjukkan bahwa rata-rata akurasi yang dihasilkan dari seluruh *corpus* yang dimiliki *NLTK* adalah 89.11%. Hal ini menunjukkan bahwa *library NLTK* menghasilkan *corpus* dan pemrosesan teks dengan akurasi yang sangat tinggi.

Penelitian serupa terkait pemanfaatan teknologi *OCR* dan *text feature extraction* juga pernah dilakukan Ismaya[8] dalam mendapatkan informasi dari Laporan Hasil Pemeriksaan (LHP) BPK RI menggunakan *rule based classifier* dan *POS tagging*. Dalam penelitiannya, dataset yang digunakan memiliki pola yang lebih terstruktur sehingga akurasi yang dihasilkan pun sangat tinggi, yaitu 98.27% untuk data training dan 89.77% untuk data testing. Pola dokumen Temuan Pemeriksaan pada penelitian tersebut telah tersusun berurutan mulai dari kondisi, kriteria, sebab, akibat, tanggapan dan saran. Akan tetapi, dataset pada penelitian tersebut telah memiliki pola yang lebih terstruktur sementara surat pengaduan ketenagakerjaan tidak terstruktur.

Oleh karena itu, penelitian ini dipusatkan untuk menerapkan teknologi *OCR* menggunakan library open source Tesseract dan ekstraksi fitur menggunakan modul *NLTK* pada penyampaian pengaduan pelanggaran ketenagakerjaan untuk menghasilkan basisdata pengaduan tenaga kerja terstruktur. Basisdata hasil dari penelitian ini dapat dikembangkan menjadi *Big Data analysis* untuk keperluan *Business Intelligence* dan *Data Mining*. Melalui teknologi ini, digitalisasi data pelaporan pelanggaran ketenagakerjaan dapat terwujud dan mempercepat Dinas Ketenagakerjaan dan Transmigrasi dalam menganalisis kasus-kasus yang terjadi di perusahaan.

2. Metode Penelitian

Penelitian ini bertujuan membangun basisdata pengaduan ketenagakerjaan dengan tahapan seperti pada gambar 1.



Gambar 1 Alur Pelaksanaan Penelitian

Pada tahap pertama dalam penelitian ini dilakukan pengumpulan surat pengaduan ketenagakerjaan yang dialamatkan kepada dinas ketenagakerjaan. Surat yang dikumpulkan dalam bentuk gambar printscreen, foto dan hasil pindai. Analisis informasi yang terkandung dalam dataset juga dilakukan pada tahap ini dan menjadi kebutuhan dalam melakukan implementasi basisdata untuk penyimpanan data hasil akhir dari implementasi text feature extraction pada tahap terakhir.

Setelah basisdata diimplementasikan, pemodelan dan pembangunan *OCR system* dilakukan. Gambar surat pada dataset akan menjadi input pada sistem kemudian dilakukan pencarian baris yang mengandung teks serta penentuan baseline sebagai batas gambar untuk teks dan membantu dalam menentukan huruf besar dan huruf kecil dalam sebuah kata pada gambar. Setelah itu gambar akan dipotong berdasarkan kata kemudian dipotong kembali menjadi kumpulan karakter. Gambar potongan karakter dan kata inilah yang akan menjadi masukkan bagi proses klasifikasi untuk menentukan huruf dan kata yang sesuai. Proses klasifikasi akan dilakukan dua tahap yaitu klasifikasi statis dan klasifikasi adaptif[9]. Perbedaannya hanya terletak dari data training. Tahap pertama pengenalan karakter menggunakan klasifikasi statis dengan training data yang sudah ditentukan. Setelah itu, data hasil klasifikasi statis akan dimasukkan ke dalam data trainig. Data training yang jumlahnya sering bertambah ini yang disebut sebagai klasifikasi adaptif. Dengan klasifikasi adaptif, kata yang tidak terklasifikasi pada tahap pertama akan terdeteksi pada tahap kedua. Output sistem berupa file berisi text hasil pengubahan bentuk gambar menjadi text.

Tahap selanjutnya, dengan menggunakan *text feature extraction*, isi file text hasil *OCR system* akan diubah strukturnya menjadi informasi surat berupa tanggal surat, organisasi pengirim surat, perihal surat, perusahaan yang diadukan, dan tempat dikeluarkannya surat. Informasi ini akan disimpan dalam basisdata yang telah dibangun. Dalam prosesnya, setelah file text didapatkan, dilakukan pemisahan kalimat pada text tersebut kemudian setiap kalimat akan dilakukan tokenisasi untuk memisahkan setiap kata dalam kalimat. Kata-kata ini akan diberi label melalui *POS tagging*. Hasil *POS tagging* akan menentukan kumpulan frase yang merupakan kata-kata yang berhubungan, misalnya kata *kucing* dengan tag kata benda akan berelasi dengan kata *putih* dengan tag kata sifat sehingga dapat menghasilkan frase *kucing putih*. Inilah yang disebut sebagai entitas. Setelah entitas-entitas ini ditemukan, kemudian dicari kembali relasi untuk setiap entitas untuk selanjutnya dilakukan klasifikasi penentuan nama dari relasi atau sering disebut *named entity recognition*.

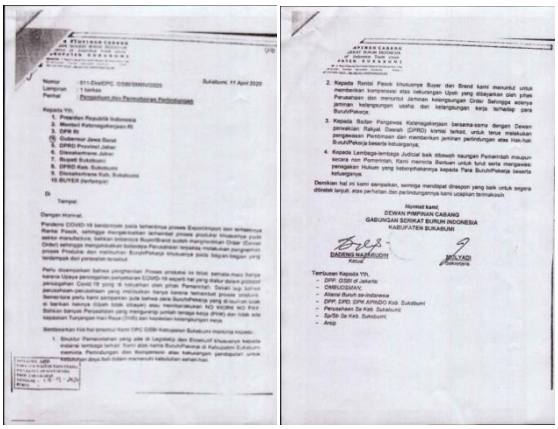
Pada tahapan terakhir penelitian ini, pengujian terhadap prototipe dilakukan untuk mengukur akurasi hasil OCR system dan akurasi hasil *Text Feature Extraction*. Akurasi hasil OCR System dihitung menggunakan tiga

jenis dataset, yaitu surat hasil pindai, surat hasil photo, dan surat hasil printscreen. Pengujian dilakukan dengan membandingkan text hasil OCR System dan text hasil ketik manual menggunakan *Cosine Similarity.*Selain itu, pengukuran akurasi hasil *Text Feature Extraction* dilakukan dengan membandingkan ketepatan hasil penentuan label dari text hasil OCR System dengan text hasil ketik manual.

3. Hasil dan Pembahasan

3.1. Dataset

Dataset yang digunakan pada penelitian ini berupa surat pengaduan tenaga kerja di lingkungan Dinas Tenaga Kerja Jawa Barat. Berdasarkan analisis surat tersebut, didapatkan informasi berupa kop surat yang selalu menandakan nama organisasi atau lembaga bantuan hukum yang mengirim surat, nama perusahaan, alamat perusahaan, tempat dan tanggal dikeluarkannya surat, nomor surat, perihal surat, tujuan surat, isi pengaduan surat, dan pengirim surat. Dataset surat yang digunakan juga dibagi menjadi 3 jenis berdasarkan cara mengambil gambar. Dataset 1 berisi gambar surat yang diambil menggunakan alat pindai yang dicontohkan pada Gambar 2, Dataset 2 berisi gambar surat yang diambil menggunakan kamera, dan Dataset 3 berisi gambar surat yang diambil menggunakan fitur printscreen.

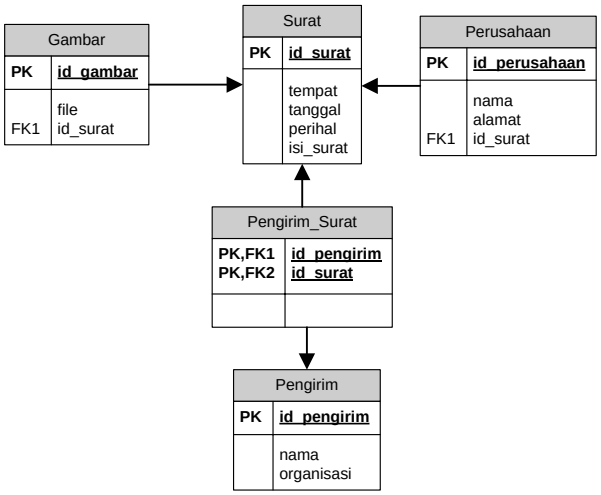


Gambar 2 Contoh Surat Pengaduan yang dipindai

Dari hasil analisis pun diketahui bahwa tujuan surat memungkinkan lebih dari satu, file gambar yang dikirim memungkinkan lebih dari satu, pengirim memungkinkan lebih dari satu, isi pengaduan pun berupa paragraf panjang yang memungkinkan untuk lebih dari satu paragraf, dan tidak semua surat terdapat tembusan.

3.2. Basisdata pengaduan ketenagakerjaan

Berdasarkan hasil analisis terhadap dataset, terdapat batasan-batasan yang dilakukan dalam membangun basisdata pengaduan ini. Batasan tersebut adalah tujuan surat tidak akan dimasukkan ke dalam basis data dikarenakan semua surat pengaduan, secara konteks ditujukan kepada dinas ketenagakerjaan dan isi pengaduan surat tidak akan dimasukkan ke dalam basisdata melainkan tetap disimpan dalam file teks untuk kegunaan penelitian berikutnya. Perancangan basisdata pengaduan ketenagakerjaan digambarkan menggunakan model konseptual pada Gambar 3.

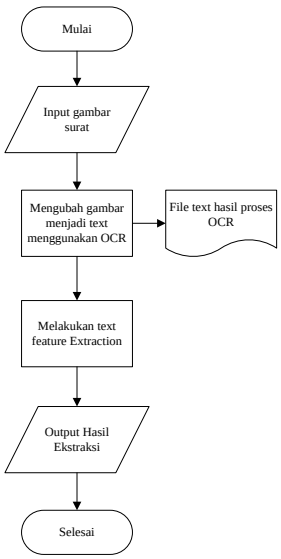


Gambar 3 Model Konseptual Basisdata Surat Pengaduan

Terdapat empat tabel yang dibangun, yaitu : Gambar, Surat, Perusahaan dan Pengirim. Rancangan ini diimplementasikan ke dalam *MySQL*. Hasil basisdata ini dapat dimanfaatkan untuk membuat Sistem Informasi, *Summarization*, *Data Mining*, dan *Business Intelligent*.

3.3. OCR System

Alur kerja sistem pada prototype ditunjukkan pada Gambar 4. Prototype akan menjalankan dua buah proses, yaitu *OCR System* dan *Text Feature Extraction*.



Gambar 4 Alur kerja sistem

Implementasi *OCR System* memanfaatkan *library open source* menggunakan *Python* bernama *Tesseract*. Dalam mengkonversi gambar menjadi file teks, ada beberapa tahapan yang harus dilakukan *OCR* menggunakan *Tesseract*. Tahap pertama yaitu mengubah file gambar

menjadi file biner berbentuk matriks. Kemudian dilakukan pencarian posisi yang mengandung baris dan posisi mengandung kata. Proses ini dilakukan untuk menentukan jika posisi huruf miring. Jika posisi telah ditemukan, dilakukan pemotongan posisi kata berdasarkan jarak spasi.

Setelah itu, tahap berikutnya yaitu dilakukan pemotongan posisi kata menjadi huruf. Kemudian, dilakukan pengenalan karakter menggunakan mesin learning dan membutuhkan data training yang berisi pola gambar dari setiap karakter. *Tesseract* versi terbaru telah menggunakan algoritma *Long Short Term Memory (LSTM)* yang lebih akurat.

Setelah karakter dikenali, analisis linguistik dilakukan menggunakan daftar string kata. *Tesseract* akan memilih kata dengan rating jarak terdekat yang artinya kata tersebut memiliki makna.

Tahap terakhir dari proses *OCR* ini adalah melakukan proses klasifikasi kembali menggunakan *adaptive classifier*. Proses ini digunakan untuk menentukan karakter berdasarkan font.

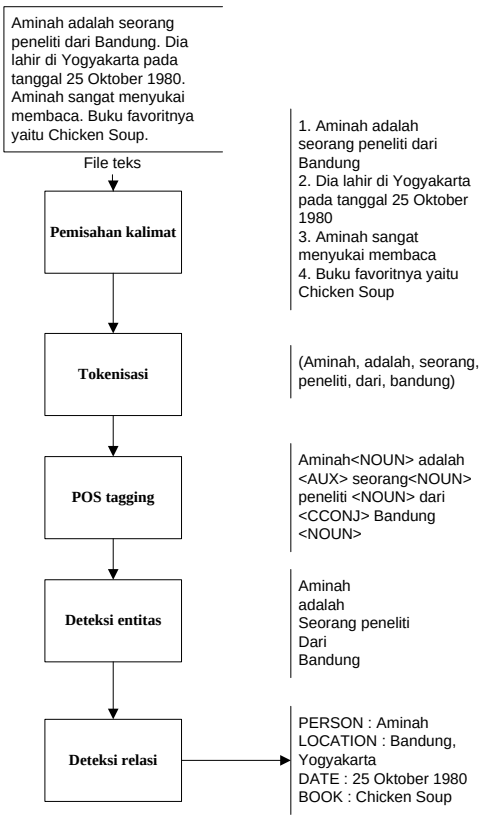
Setelah implementasi, dilakukan evaluasi terhadap sistem ini. Evaluasi dilakukan dengan mengetes akurasi hasil *OCR Tesseract* menggunakan 3 jenis surat, yaitu surat hasil scan (Dataset 1), surat hasil foto (Dataset 2), dan surat hasil printscreen (Dataset 3). Ketiga jenis surat ini dihitung akurasi ketepatan mendeteksi huruf dan kata menggunakan pengukuran *Cosine Similarity* yang dibandingkan dengan jenis surat yang diketik secara manual. Hasilnya menunjukkan bahwa surat hasil printscreen dan foto lebih akurat dibandingkan dataset yang lain. Seperti yang ditunjukkan pada Tabel 1, Nilai *similarity* hasil *OCR* terhadap printscreen dan foto paling besar di antara jenis surat yang lain. Sementara itu, waktu pemrosesan yang paling lama terjadi pada jenis surat berupa foto karena ukuran resolusi photo yang lebih besar. Hal ini menunjukkan bahwa sistem ini akan lebih akurat jika menggunakan dokumen surat yang kontras antara latar surat dan teksnya. Oleh karena itu, masih dibutuhkan penelitian lanjutan mengenai cara preproses gambar agar akurasinya lebih baik.

Tabel 1 Hasil pengujian OCR System

Jenis Surat	Similarity	Waktu
Dataset 1	0.849	20.83
Dataset 2	0.943	33.88
Dataset 3	0.946	5.37

3.4. Text Feature Extraction

File text hasil *OCR System* akan menjadi input bagi *Text Feature Extraction*. Tahapan yang dilakukan pada *Text Feature Extraction* ini dicontohkan pada gambar 5.



Gambar 5 Alur Proses Text Feature Extraction

Proses-proses pada gambar di atas diimplementasi menggunakan library *NLP NLTK(Natural Language Toolkit)* yang sudah tersedia. *NLTK* dirancang dengan beberapa tujuan utama, yaitu : *ease of use, consistency, extensibility, documentation, simplicity,* dan *modularity*. Selain itu, *NLTK* memiliki modul yang lengkap untuk pemrosesan teks, yaitu : *parsing, tagging, finite state automata, type checking, visualization* dan *text classification*[5].

Pada tahap pertama file teks akan dilakukan pemisahan kalimat atau segmentasi kalimat menggunakan fungsi `sent_tokenize(file_text)` pada modul nltk. Setelah kalimat-kalimat terpisah, dilanjutkan dengan melakukan tokenisasi atau pemecahan kalimat menjadi kata menggunakan fungsi `word_tokenize(word)` secara rekursif pada setiap kalimat hasil segmentasi di tahap pertama. Setelah itu dilakukan *POS tagging* menggunakan fungsi `pos_tag(word)` secara rekursif pada setiap kalimat. Tahap berikutnya dilakukan penentuan relasi setiap kata atau penentuan frasa menggunakan *chunking*. Proses *chunking* ini memerlukan daftar kosakata bahasa Indonesia untuk penentuan frasa menggunakan fungsi regular expression `RegexParser(grammar)` dan hasilnya berupa tree. Pada dasarnya, hasil *chunking* ini memiliki dua representasi, yaitu *tag* dan *tree*. Setelah frasa ditemukan, terkadang ada beberapa frasa yang memiliki tag yang sama tetapi berbeda jika dilihat berdasarkan konten. Untuk

mengatasi hal ini dilakukan classifier based tagger menggunakan fungsi klasifikasi pada nltk.

Pada tahap terakhir, ditentukan pengenalan pola informasi menggunakan *NER*. Pada nltk, fungsi yang digunakan adalah `ne_chunk()`. Pada proses ini juga dibutuhkan classifier untuk menentukan nama dari entitas yang ditemukan berdasarkan kelasnya, seperti *Person*, *Location*, *Date*, dan lain-lain. Pada penelitian ini, corpus data untuk pengenalan tag menggunakan tagging Bahasa Indonesia [10].

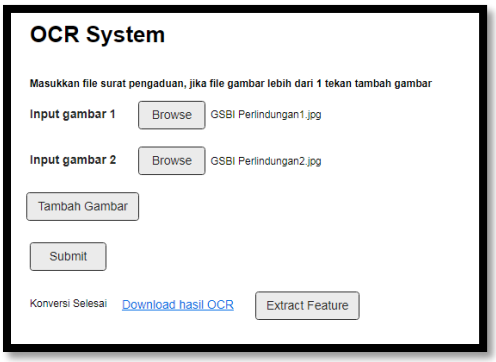
Berdasarkan hasil perbandingan antara hasil feature extraction dan isi surat yang asli, masih terdapat penentuan entitas yang kurang tepat. Salah satu penyebabnya adalah hasil konversi pada sistem *OCR* yang kurang tepat sehingga kata yang dihasilkan tidak bermakna. Selain itu, deteksi atribut yang dibaca berdasarkan paragraf masih ada beberapa yang kurang tepat seperti nama perusahaan atau alamat yang tidak terdaftar dalam *Corpus*. Sementara text yang berpola seperti nama kota dan tanggal dapat dikenali dengan akurat.

Hal ini menunjukkan bahwa masih diperlukan penelitian lanjutan mengenai algoritma yang sesuai untuk penentuan isi atribut dari text yang lebih akurat. Beberapa pilihan yang mungkin untuk dimasukkan ke dalam atribut pada basisdata sehingga atribut masih harus dipilih mana yang sesuai, contohnya untuk atribut tempat pada tabel surat memiliki beberapa opsi nama tempat yang terdeteksi, misalnya pada surat yang dimaksud tempat adalah tempat dikeluarkannya surat, tetapi hasil *text feature extraction* menghasilkan beberapa pilihan tempat yang dihasilkan misalnya Bandung, Yogyakarta dan sebagainya. Oleh karena itu, pada prototype aplikasi masih membutuhkan konfirmasi dari user untuk memeriksa kebenaran hasil *text feature extraction* dengan surat asli sebelum diinput ke basisdata.

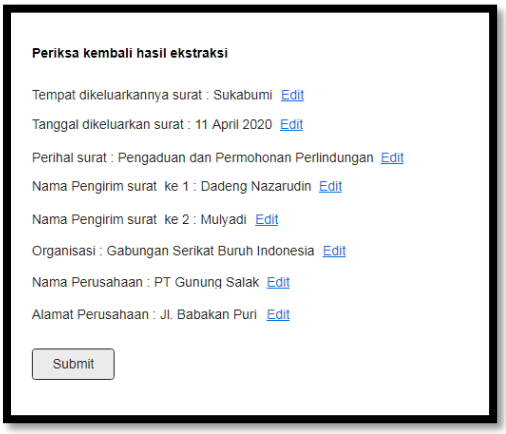
3.5. Prototipe aplikasi

Gambar 6 memperlihatkan prototype aplikasi yang digunakan. Prototype belum terintegrasi secara keseluruhan. Pada prosesnya masih memerlukan dua tahap yaitu, pengubahan file gambar menjadi file teks dengan *OCR* dan pembentukan feature dengan *text feature extraction.*

Prototype ini dibuat berbasis web. File gambar yang dimasukkan dapat lebih dari satu. Pada Gambar 6, prototype menggunakan file input berupa contoh surat pada Gambar 2. Setelah menekan tombol Submit maka proses *OCR* pun dijalankan. Jika sudah selesai, akan keluar pesan "Konversi Selesai" kemudian muncul link untuk mengunduh file teks dan tombol untuk melakukan text feature extraction. Setelah proses ekstraksi selesai akan muncul halaman seperti pada Gambar 7.



Gambar 6 Halaman OCR System dan Text Feature Extraction



Gambar 7 Halaman Konfirmasi Hasil Text Feature Extraction

Halaman pada gambar menunjukkan hasil *text feature extraction* dan *user* harus melakukan pengecekan sebelum data dimasukkan ke dalam basisdata. Ada kemungkinan kesalahan pengenalan karakter atau jumlah nama orang dan nama perusahaan yang terdeteksi lebih dari satu akibat kekurangan dari proses pengenalan. Selain itu, hasil *feature extraction* hanya dapat mengenali semua *tag* nama orang, organisasi, alamat, dan nama perusahaan bukan dari konteksnya. Oleh karena itu, user diminta untuk mengecek kembali data hasil ekstraksi agar data yang tersimpan dalam basisdata sesuai dengan isi surat.

Pengujian pada prototipe telah dilakukan menggunakan *black box testing*. Selain itu, dilakukan pula pengujian akurasi dari proses secara keseluruhan terhadap text surat hasil *OCR* dan file yang diketik secara manual. Yang dijadikan dasar pengukuran adalah ketepatan penentuan ekstraksi berdasarkan label-label yang ditemukan tanpa melihat konteksnya. Hasilnya ditunjukkan pada Tabel 2.

Tabel 2 Hasil pengujian prototype

Jenis Surat	Akurasi
Text Hasil OCR	66.67%
File hasil ketik	91.67%

Berdasarkan nilai akurasi pada Tabel 2, nilai akurasi untuk text hasil OCR lebih kecil dibandingkan hasil ketik. Hal ini dikarenakan akurasi hasil penentuan

feature extraction yang tidak dapat menentukan label dengan tepat sebagai efek proses segmentasi gambar pada OCR. Dalam prosesnya, OCR akan melakukan segmentasi berupa deteksi gambar logo yang biasanya terdapat nama perusahaan atau organisasi, gambar cap organisasi, kop surat, isi surat, tanda tangan pada surat, tembusan, dan terkadang terdapat surat dengan tulisan tangan. Hasil segmentasi tersebut memungkinkan kesalahan menentukan text dikarenakan struktur text hasil OCR menjadi acak. Hal ini berpengaruh terhadap proses *text feature extraction* yang menentukan label berdasarkan keterhubungan kata. Selain itu, adanya beberapa data surat hasil pindai yang tidak kontras antara background dan tulisan seperti pada Gambar 2 membuat deteksi teks yang kurang tepat. Hal ini terlihat dari nilai similarity hasil OCR dataset 1 pada Tabel 1 yang paling kecil dibandingkan dataset lainnya.

4. Kesimpulan

Penelitian ini menghasilkan suatu prototype input surat dan Basis data pengaduan ketenagakerjaan yang masih memerlukan penelitian lebih lanjut. Beberapa penelitian preproses gambar terkait hasil akurasi *OCR* dari dokumen gambar berupa hasil pindai jika background dan text tidak kontras masih diperlukan. Hal ini mempengaruhi hasil ekstraksi karena pengenalan karakter yang tidak tepat. Terbukti berdasarkan hasil pengujian prototype yang hanya menghasilkan akurasi 66.67%. Selain itu, pengenalan atribut dari teks berbentuk paragraf juga masih membutuhkan pencarian algoritma yang lebih akurat untuk teks berbahasa Indonesia.

UcapanTerimakasih

Penelitian ini dibiayai oleh Hibah Penelitian Dosen Pemula dari DIKTI dengan kontrak No. 021/SP2H/PPM/LL4/2020.

Daftar Rujukan :

[1] Patel, Chirag, "Optical Character Recognition by Open Source OCR Tool Tesseract : A Case Study," International Journal of Computer Applications, vol. 55, no. 10.

[2] Algun, Selcuk. (2018, December) Review for Tesseract and Kraken OCR for Text recognition. [Online]. HYPERLINK "https://medium.com/datadriveninvestor/review-for-tesseract-and-kraken-ocr-for-text-recognition-2e63c2adedd0" https://medium.com/datadriveninvestor/review-for-tesseract-and-kraken-ocr-for-text-recognition-2e63c2adedd0

[3] Liang, Hong, "Text feature extraction based on deep learning: a review," EURASIP Journal on Wireless Communications and Networking, 2017.

[4] Piskorski, Jacub and Yangarber, Roman, "Information Extraction Past, Present and Future," in Theory and Applications of Natural Language Processing. Berlin: Springer, 2013, pp. 23-49.

[5] Bird, Steven; Klein, Ewan; and Loper, Edward, Natural Language Processing with Python, 1st ed. USA: O'Reilly, 2009.

[6] M.,Viny Christianty; Pragantha, Jeany, and Purnamasari, Endah, "Implementasi Bill Tagger untuk memberikan POS Tagging pada Dokumen Bahasa Indonesia," Jurnal Teknik dan Ilmu Komputer, vol. I, no. 3.

[7] Yumusak, S; Dongdu, E.; and Kodaz, H., "Tagging Accuracy Analysis on Part-of-Speech Taggers," Journal of Computer and Communications, no. 2, pp. 157-162.

[8] Ismaya, Agny, "Algoritma Ekstraksi Informasi berbasis Aturan," JNTETI, vol. III, no. 4, 2014.

[9] Smith, R., "An Overview of the Tesseract OCR Engine," in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Parana, 2017.

[10] Rashel, Fam; Luthfi, Andri; Dinakaramani, Arawinda, and Manurung, Ruli, "Building an Indonesian Rule-Based Part-of-Speech Tagger," in International Conference on Asian Language Processing (IALP 2014), Kuching, 2014.