Algoritma Ekstraksi Informasi Berbasis Aturan

Agny Ismaya[1]

***Abstract*— The information in in the audit report of local government financial statement (LHP LKPD) was not managed digitally. The information in LHP from 33 provinces has just accumulate in a place without next process to take its main information. The absence of information searching application inhibit the learning process of the existing reports in advance. Therefore, an application can extract information from a set of LHP documents are needed to get main information, called criteria, consequence, cause, response, and audit advice.**

This research creates a tool to extract the information in the audit report of local government financial statement (LHP LKPD). Information extraction method that used in this research is rule-based classification and pre-processing method that used is POS Tagging. The objective of information extraction in this research finds some sections in audit finding (Temuan Pemeriksaan-TP) that are criteria, consequence, cause, response, and audit advice.

The accuracy of training and test data are 98,27% and 89,77%. Decrease accuracy caused by usage of pdf2text that do not give a convertible identical between the input and output data, and usage of wordmatch method for classification.

***Intisari*— Informasi yang terdapat dalam Laporan Hasil Pemeriksaan (LHP) BPK RI saat ini belum di-manaje dengan baik secara digital. Informasi yang terdapat dalam LHP yang terdapat pada 33 provinsi sampai saat ini hanya terkumpul di satu tempat tanpa ada proses selanjutnya untuk mengambil informasi inti dalam laporan tersebut. Ketiadaan aplikasi pencarian informasi menghambat proses pembelajaran terhadap laporan-laporan yang ada sebelumnya. Oleh karena itu, diperlukan sebuah aplikasi yang mampu melakukan ekstraksi informasi dari sekumpulan dokumen LHP untuk dapat mengambil data inti dari dokumen tersebut, yaitu criteria, akibat, sebab, tanggapan, dan saran pemeriksaan.**

Penelitian ini membuat sebuah tools untuk melakukan ekstraksi informasi pada dokumen Laporan Hasil Pemeriksaan (LHP) atas Laporan Keuangan Pemerintah Daerah (LKPD). Metode yang digunakan dalam penelitian ini adalah rule-based classification, dan metode preprocessing yang digunakan adalah POS Tagging. Tujuan ekstraksi informasi dalam penelitian ini adalah mendapatkan informasi beberapa bagian penyusun TP dalam dokumen LHP LKPD, yaitu kriteria, akibat, sebab, tanggapan, dan saran pemeriksaan.

Akurasi dari data latih dan data uji masing-masing adalah 98,27% dan 89,77%. Penurunan akurasi antara lain disebabkan oleh penggunaan aplikasi pdf2text yang tidak memberikan hasil konversi yang identik antara data input dan output, serta penggunaan metode wordmatch untuk klasifikasi.

***Kata kunci*— LHP LKPD, ekstraksi informasi, klasifikasi, POS Tagging, rule-based classification.**

[1]*Staff, Badan Pemeriksa Keuangan RI Perwakilan Provinsi NTB Jl. Udayana No. 22 Mataram Nusa Tenggara Barat 83123 Telp. (0370)6163333,Fax. (0370)6162999*

I. PENDAHULUAN

Informasi yang terdapat dalam Laporan Hasil Pemeriksaan (LHP) BPK RI saat ini belum dikelola dengan baik secara digital. Informasi yang terdapat dalam LHP pada 33 provinsi sampai saat ini hanya terkumpul di satu tempat tanpa ada proses selanjutnya untuk mengambil informasi inti dalam laporan tersebut. Ketiadaan proses ekstraksi informasi ini mengakibatkan adanya kesulitan jika sewaktu-waktu diperlukan informasi tertentu dari sejumlah besar dokumen LHP dan menghambat proses pembelajaran terhadap laporan-laporan yang ada sebelumnya. Oleh karena itu, diperlukan sebuah aplikasi yang mampu melakukan ekstraksi informasi dari sekumpulan dokumen LHP untuk dapat mengambil data inti dari dokumen tersebut.

Penelitian ini akan melakukan klasifikasi bagian dalam Temuan Pemeriksaan (TP) yang merupakan bagian dari LHP. Setiap TP terdiri dari 6 bagian, yaitu kondisi, kriteria, sebab, akibat, tanggapan, dan saran. Klasifikasi dalam penelitian ini dilakukan dengan mengembangkan sebuah Sistem Informasi yang mampu mengambil informasi dari 5 bagian, yaitu kriteria, sebab, akibat, tanggapan, dan saran dari temuan pemeriksaan. Kriteria menjelaskan peraturan yang dilanggar dalam setiap TP. Sebab menjelaskan latar belakang terjadinya kasus yang dicantumkan dalam TP. Tanggapan berisi penjelasan entitas (pihak yang diperiksa) atas adanya TP. Saran mencantumkan saran yang diberikan oleh BPK kepada entitas atas adanya TP.

Saat ini, Temuan Pemeriksaan (TP) hanya dikumpulkan secara elektronis berupa file pdf. Dari kumpulan file pdf tersebut, tidak ada proses yang dilakukan lebih lanjut untuk mendapatkan suatu informasi sebagai dasar pengambilan keputusan bagi “top level management” ataupun pengambilan keputusan untuk mempertimbangkan penyusunan LHP bagi kasus serupa di lokasi yang berbeda. Pengambilan informasi yang selama ini dilakukan hanya bersifat manual sehingga sangat memungkinkan terjadinya perbedaaan informasi yang diambil untuk kasus yang semestinya serupa. Oleh karena itu diperlukan suatu aplikasi yang mampu mengidentifikasi atau mengklasifikasi temuan – temuan pemeriksaan yang telah ada pada tahun-tahun sebelumnya, sehingga diharapkan dapat membantu proses pengambilan keputusan.

Penelitian ini akan menggunakan salah satu cabang dari Pemrosesan Bahasa Alami (PBA), yaitu Ekstraksi Informasi. Ekstraksi Informasi dapat diartikan sebagai sebuah proses untuk mendapatkan informasi inti dari dokumen tidak terstruktur ataupun dokumen semi terstruktur. Dalam penelitian ini, ekstraksi informasi dibagi dalam dua proses, yaitu preprocessing dan klasifikasi. Preprocessing dilakukan untuk menyiapkan dokumen agar dapat diproses pada tahap berikutnya, yaitu klasifikasi, serta untuk mendapatkan akurasi yang lebih tinggi dari hasil klasifikasi. Sedangkan klasifikasi

merupakan proses inti dalam penelitian ini, yang bertujuan untuk mengelompokkan isi dokumen ke dalam beberapa klasifikasi.

Penelitian mengenai Pemrosesan Bahasa Alami ataupun part of speech tagger untuk bahasa Indonesia pernah dilakukan oleh beberapa peneliti. Salah satu diantaranya menggunakan metode Conditional Random Fields dan Transformation Based Learning [1]. Di samping itu, terdapat Penelitian serupa memanfaatkan algoritma GLR (Generalized Left-to-right Rightmost) dengan kalimat yang dibangun bersifat rule-based untuk membangun pengurai kalimat Bahasa Indonesia.

Penelitian lain menggunakan metode Hidden Markov Model (HMM) untuk membangun POS Tagger Bahasa Indonesia (HMM Based Part-of-Speech Tagger for Bahasa Indonesia) [2].

Dari beberapa penelitian tersebut, POSTagger dengan metode HMM memiliki akurasi tertinggi, yaitu 96,50% , dimana 99,40% untuk kata-kata yang ada dalam kosakata bahasa Indonesia, dan 80,40% untuk kata-kata di luar kosakata bahasa Indonesia (OOV-out of vocabulary).

Penelitian ini akan memanfaatkan aplikasi IPOSTagger yang dibangun dengan metode HMM. Dokumen LHP akan menjadi input pada aplikasi IPOSTagger, dan selanjutnya proses klasifikasi akan dilakukan terhadap output dari aplikasi ini.

Informasi yang diharapkan dari penelitian ini adalah berupa pengelompokkan komponen temuan pemeriksaan dalam LHP LKPD yang berupa kriteria, sebab, akibat, tanggapan, dan saran.

Tujuan dari penelitian ini adalah membuat tools untuk dapat meng-ekstrak informasi yang terdapat dalam Laporan Hasil Pemeriksaan (LHP) atas Laporan Keuangan Pemerintah Daerah (LKPD). Informasi dalam Temuan Pemeriksaan (TP) LHP LKPD yang ingin diperoleh dari penelitian ini adalah kriteria, sebab, akibat, tanggapan, dan saran pemeriksaan.

II. EKSTRAKSI INFORMASI

Ekstraksi Informasi bisa diartikan sebagai sebuah proses untuk menemukan informasi terstruktur dari dokumen tidak terstruktur atau semi terstruktur. Ekstraksi informasi merupakan salah satu bagian dari Pemrosesan Bahasa Alami (Natural Language Processing) [3].

Terdapat beberapa metode yang dapat digunakan dalam ekstraksi informasi, di antaranya rule-based methods, statistical methods,dan knowledge methods.

Pada penelitian ini, ekstraksi informasi akan diterapkan pada LHP atas LKPD. LHP atas LKPD merupakan hasil dari pemeriksaan keuangan pada pemerintah daerah.

Pemeriksaan keuangan adalah pemeriksaan atas laporan keuangan pemerintah (Pusat, Daerah, BUMN, dan BUMD) dengan tujuan pemeriksaan menilai kewajaran laporan keuangan dan memberikan pernyataan pendapat/opini tentang tingkat kewajaran informasi yang disajikan dalam laporan keuangan pemerintah pusat/daerah.

A. Praprocessing

Preprocessing adalah proses ataupun metode yang perlu dilakukan agar data dapat digunakan dalam proses inti data mining. Sebagian besar permasalahan pada text mining dapat diselesaikan dengan beberapa algoritma yang berbeda. Setiap teknik yang digunakan dalam preprocessing dimulai dengan mengamati dokumen terstruktur dan memprosesnya untuk mendapatkan lebih banyak pola dengan cara menyaring kemiripan yang ada dan menambahkan ciri yang lain.

Di akhir fase preprocessing, pola yang paling mendekati kemiripan dengan dokumen yang diteliti akan digunakan untuk text mining. Metode – metode yang dapat digunakan dalam preprocessing antara lain POS Tagging, stemming, full parsing, shallow parsing [4].

Penelitian ini menggunakan metode POS Tagging karena metode ini dianggap sesuai dan dapat digabungkan dengan rule-based classification untuk mendapatkan pola yang diharapkan. POSTagging yang digunakan dalam penelitian ini memanfaatkan metode Hidden Markov Model (HMM).

HMM adalah peluasan dari rantai Markov dimana state-nya tidak dapat diamati secara langsung (tersembunyi), tetapi hanya dapat diobservasi melalui suatu himpunan pengamatan lain [5].

Markov Chain bermanfaat untuk menghitung probabilitas urutan kejadian yang dapat diamati. HHM berguna untuk mendapatkan urutan kejadian yang ingin diketahui tetapi tidak dapat diamati. Salah satu contoh adalah kasus part-of speech tagging (POS Tagging).

Pada POS Tagging, urutan tag tidak dapat diamati secara langsung. Pengamatan secara langsung hanya dapat dilakukan terhadap urutan kata. Dari urutan kata tersebut harus dicari urutan tag yang paling tepat. Dalam kasus ini, tag adalah bagian yang tersembunyi.

Dalam HMM, bagian yang dapat diamati disebut observed state, sedangkan bagian yang tersembunyi disebut hidden state. HMM memungkinkan pemodelan system yang mengandung observed state dan hidden state yang saling terkait. Pada kasus POS Tagging, observed state adalah urutan kata sedangkan hidden state adalah urutan tag.

B. Klasifikasi

Klasifikasi merupakan proses pembelajaran sebuah fungsi atau model terhadap sekumpulan data latih, sehingga model tersebut dapat digunakan untuk memprediksi klasifikasi dari data uji [6].

Beberapa metode yang dapat digunakan dalam klasifikasi antara lain decision tree, rule-based classifiers, Bayesian classifier, Support Vector Machines, Artificial Neural Networks, Lazy Learners, dan ensemble methods.

Decision tree merupakan sebuah metode pembelajaran dengan menggunakan data latih yang telah dikelompokkan berdasarkan kelas-kelas tertentu dalam pohon keputusan. Rule-based classifiers merupakan salah satu teknik klasifikasi dengan menggunakan aturan “if ... then ... else...”.

Bayesian classifiers menggunakan metode statistik dan berdasarkan pada teori Bayes. Artificial Neural Networks merupakan metode klasifikasi dengan menggunakan perhitungan yang mengadaptasi cara kerja otak manusia. Ensemble methods membangun classifier dari data latih dan menghasilkan prediksi klasifikasi yang dibentuk dari masing-masing classifier.

Metode klasifikasi yang digunakan dalam penelitian ini adalah rule-based classification. Metode ini dipilih karena

dinilai paling sesuai dengan struktur dokumen LHP yang mejadi obyek penelitian.

Dalam penelitian ini, proses klasifikasi dilakukan setelah preprocessing dengan metode POSTagging. Setiap kata yang dikenali dari proses POS Tagging akan digunakan dalam klasifikasi. Proses pertama yang dilakukan dalam klasifikasi adalah pencarian kondisi berhenti tiap bagian dokumen. Setelah kondisi berhenti untuk masing-masing bagian dokumen diketahui, proses dapat dilanjutkan dengan mencari isi masing-masing bagian dokumen dan menghitung jumlahnya.

Jenis dokumen yang digunakan dalam penelitian ini adalah dokumen semi terstruktur. Menurut Feldman, dokumen semi terstruktur bisa diartikan sebagai dokumen yang memiliki elemen atau format yang konsisten dimana setiap tipe dari bagian dokumen tersebut dapat dikenali dengan mudah. Contoh dari dokumen semi terstruktur adalah HTML, file pdf, dan file word yang memiliki template atau batasan-batasan pada style-sheet.

Dokumen yang menjadi obyek dalam penelitian ini dikategorikan semi terstruktur karena setiap dokumen LHP terdiri atas beberapa Temuan Pemeriksaan (TP), dan setiap TP terdiri dari beberapa bagian, yaitu kondisi, kriteria, akibat, sebab, tanggapan, dan saran.

III. METODOLOGI

Proses klasifikasi dokumen LHP akan dilakukan dengan membangun system yang akan melakukan klasifikasi dengan metode rule-based.

Langkah pertama adalah melakukan pemrosesan awal terhadap LHP LKPD dengan melakukan pencarian pola setiap komponen (kriteria, sebab, akibat, dan saran) dalam temuan pemeriksaan berdasarkan kata kunci. Hasil dari proses ini adalah pola setiap bagian TP.

Setelah mendapatkan pola, proses selanjutnya adalah mencari posisi bagian tersebut dalam dokumen. Jika posisi dari tiap bagian tersebut telah diketahui (berdasarkan kata kunci yang menjadi penanda awal dan akhir suatu bagian TP), maka isi dari suatu bagian dapat diperoleh.

IV. PEMBAHASAN

Proses ekstraksi informasi yang dilakukan dalam penelitian ini adalah pencarian informasi inti dari dokumen Laporan Hasil Pemeriksaan (LHP) atas Laporan Keuangan Pemerintah Daerah (LKPD) pada BPK RI Perwakilan Provinsi NusaTenggara Barat.

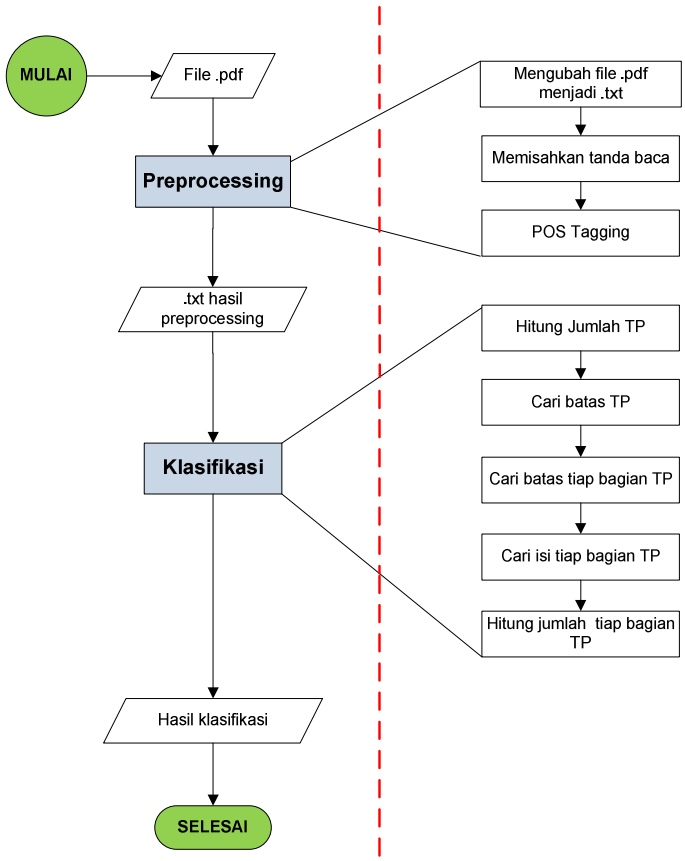
LHP LKPD yang menjadi data latih dalam penelitian adalah hasil pemeriksaan Tahun 2012. Penelitian ini dilakukan terhadap hasil pemeriksaan dari 10 entitas dari 11 entitas yang diperiksa oleh BPK RI Perwakilan Provinsi NTB.

Dokumen pemeriksaan dari masing-masing entitas yang digunakan dalam penelitian ini adalah 10 dokumen LHP Sistem Pengendalian Intern dan 10 LHP Kepatuhan atas Peraturan Perundangan.

Secara keseluruhan, dokumen yang digunakan sebagai data latih dalam penelitian ini terdiri dari 184 Temuan Pemeriksaan yang masing-masing terdiri dari 6 bagian, yaitu kondisi, kriteria, akibat, sebab, tanggapan, dan saran/rekomendasi.

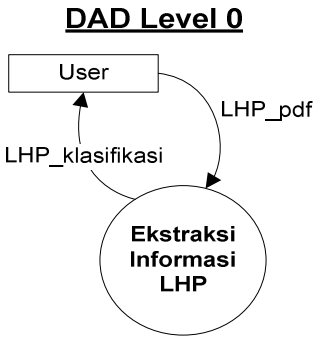
Sedangkan data uji yang digunakan dalam penelitian ini adalah LHP SPI dan LHP Kepatuhan Tahun 2011 sebanyak 8 dokumen LHP yang terdiri dari 57 Temuan Pemeriksaan.

Adapaun proses yang dilakukan dalam penelitian ini dijelaskan pada Gbr. 1.

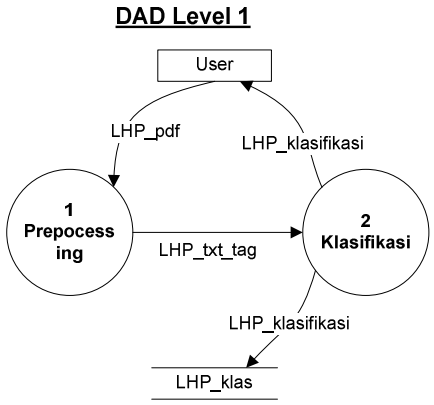


Gbr. 1. Alur Proses Ekstraksi Informasi

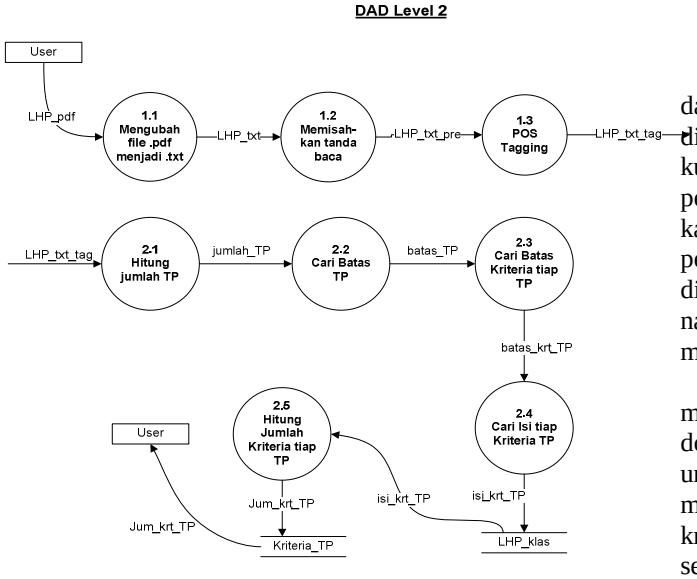
Sedangkan Digram Arus Data (DAD) dalam penelitian ini adalah tercantum pada Gbr. 2, Gbr. 3, dan Gbr. 4.



Gbr. 2 DAD Level 0



Gbr. 3 DAD Level 1



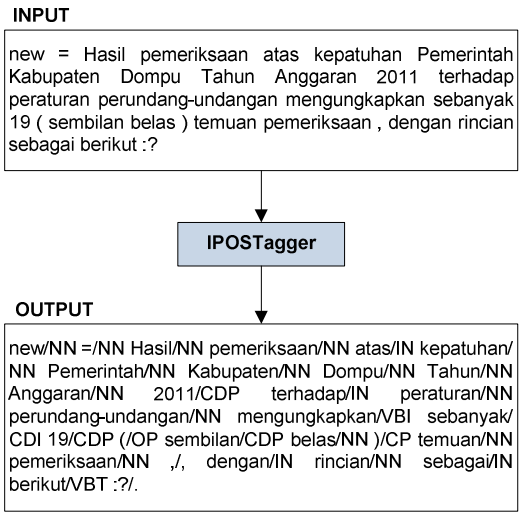
Gbr. 4 DAD Level 2

Pada proses 1.1 dilakukan perubahan file hasil pemeriksaan yang bertipe .pdf menjadi file .txt. proses ini dilakukan karena IPOSTagger tidak dapat menerima masukan berupa file .pdf. Kemudian dilakukan preprocessing terhadap file .txt pada proses 1.2 untuk memisahkan setiap tanda baca yang menyertai kata.

Proses 1.2 perlu dilakukan untuk mengurangi kesalahan yang terjadi ketika dilakukan POSTagger atau mengenali jenis dari setiap kata ataupun tanda baca. Setiap tanda baca yang terhubung dengan kata (tidak dipisahkan oleh spasi), maka hanya tanda baca tersebut yang akan dikenali oleh IPOSTagger.

Proses 1.3 adalah melakukan tagging kata untuk mengenali jenis dari setiap kata (kata kerja, kata benda, kata sifat, dll). Proses ini menggunakan aplikasi IPOSTagger. Output dari IPOSTagger adalah berupa file .txt yang berisi kata ataupun tanda baca beserta jenisnya.

Setiap kalimat dan setiap kata beserta hasil tag dalam file output disimpan dalam bentuk array. Setiap kalimat dipisahkan oleh karakter titik (.). Contoh dokumen input dan output IPOSTagger dijelaskan pada Gbr. 5.



Gambar 5. Contoh Input dan Output IPOSTagger

Dokumen output dari IPOSTagger akan menjadi input dalam proses klasifikasi. Proses klasifikasi dokumen dilakukan dengan mencari setiap bagian berdasarkan kata kunci untuk mengetahui salah satu jenis – jenis bagian temuan pemeriksaan. Proses pencarian ini memanfaatkan susunan kata hasil output yang tersimpan dalam array, sehingga pencarian dapat dilakukan dengan lebih mudah. Jika telah diketahui masing-masing bagian dari temuan pemeriksaan nantinya dapat digunakan sebagai acuan ketika akan menyusun temuan pemeriksaan yang serupa.

Langkah pertama yang dilakukan dalam klasifikasi adalah mencari jumlah Temuan Pemeriksaan (TP) dalam setiap dokumen LHP. Berdasarkan hasil percobaan, cara yang tepat untuk mendapatkan jumlah TP secara akurat adalah dengan mencari berdasarkan kata kunci yang sesuai dengan kategori kriteria pemeriksaan atau peraturan yang dilanggar dalam setiap TP. Kata kunci yang digunakan dalam kategori kriteria antara lain tercantum dalam Tabel I.

TABEL I
KATA KUNCI KRITERIA

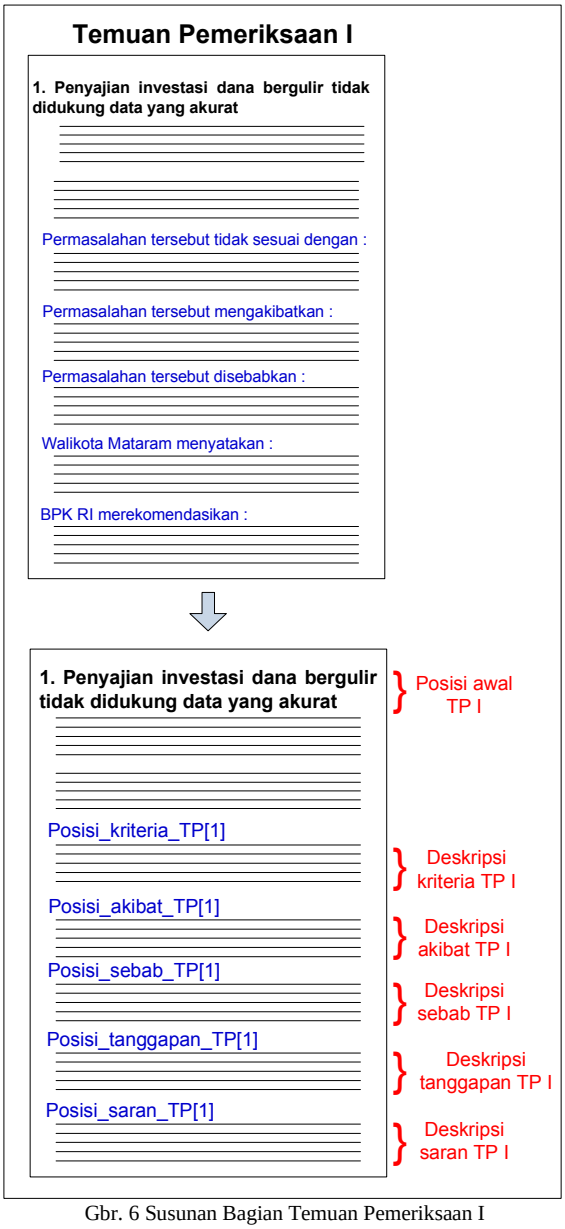
No	Kata Kunci Kriteria
1	Peraturan Presiden
2	Peraturan Menteri
3	Peraturan Gubernur
4	Peraturan Bupati
5	Peraturan Walikota
6	Undang-undang
7	Keputusan Presiden
8	Keputusan Bersama
9	Keputusan Menteri
10	Keputusan Walikota
11	Keputusan Bupati
12	Peraturan Daerah
13	Buletin Teknis
14	Perjanjian Kontrak
15	Surat Perintah Kerja

Setelah mendapatkan jumlah TP, langkah selanjutnya adalah mencari batas dari masing-masing TP. Setiap TP dibatasi oleh nomor beserta judul TP. Berdasarkan hasil penelitian, setiap TP diakhiri oleh karakter angka yang

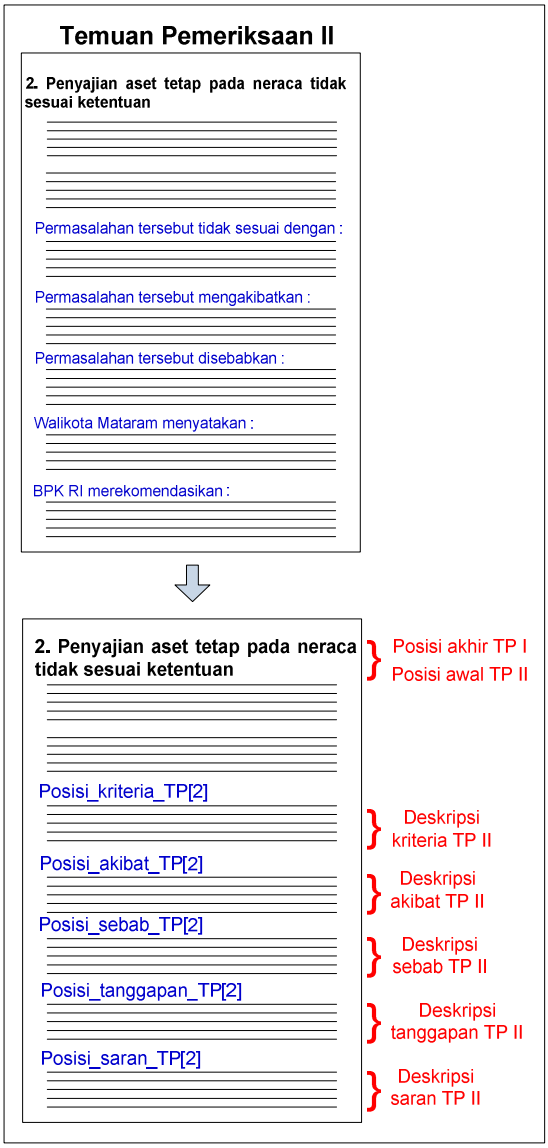
merupakan nomor TP selanjutnya, kemudian diikuti oleh karakter titik (.), dan kemudian judul TP berikutnya sabagai mana diperlihatkan Gbr. 6 dan Gbr.7.

Setelah mendapatkan batas dari masing-masing TP, langkah selanjutnya adalah mencari batas dari tiap bagian dalam TP. Batas setiap bagian ditentukan dengan mencari kata kunci dari setiap bagian, sebagaimana tercantum pada TabelII. Berdasarkan penelitian, secara umum susunan bagian dari masing-masing TP diperlihatkan oleh Tabel II.

Berdasarkan susunan tersebut, maka bagian TP yang pertama kali dicari adalah kriteria. Setiap kriteria dicari dengan mencocokkan kata dalam dokumen dengan kata kunci yang sesuai dengan kriteria, seperti pada Tabel 2. Jika ditemukan kata yang sesuai dengan kata kunci, maka kata tersebut beserta posisinya akan disimpan berdasarkan kategorinya. Dengan menggunakan kata kunci seperti ini, dapat diketahui pula batas dari setiap bagian dalam dokumen, sehingga akan mempermudah proses selanjutnya. Proses ini dilakukan untuk setiap bagian TP.



Gbr. 6 Susunan Bagian Temuan Pemeriksaan I



Gbr. 7 Susunan Bagian Temuan Pemeriksaan II

TABEL II
URUTAN BAGIAN TP

Urutan pada TP	Kategori
1	Kondisi
2	Kriteria
3	Akibat
4	Sebab
5	Tanggapan
6	Saran

Setelah batas dari setiap bagian dokumen diketahui, proses berikutnya adalah mencari isi atau deskripsi dari setiap bagian dokumen.

Untuk menghitung akurasi dari data latih dan data uji, cara yang digunakan adalah dengan membandingkan secara manual hasil dari sistem dengan dokumen LHP. Hal ini karena saat ini belum ada suatu sistem yang dapat menghitung setiap bagian dalam dokumen LHP.

Hasil klasifikasi terhadap data latih dan data uji tercantum pada Tabel III dan Tabel IV.

TABEL III
URUTAN BAGIAN TP

Keterangan	Sensitivity (%)	PPV (%)	Specivicity (%)	NPV (%)	Akurasi (%)
Seluruh Bagian TP	98.66	98.01	97.85	98.56	98.27

TABEL IV
URUTAN BAGIAN TP

Keterangan	Sensitivity (%)	PPV (%)	Specivicity (%)	NPV (%)	Akurasi (%)
Seluruh Bagian TP	85.96	92.52	93.39	87.48	89.77

Menurut Chaudhary [7], perhitungan sensitivity dan Positive Predictive Value (PPV) adalah sebagai berikut :

Sensitivity = TP / (TP+FN)
Specificity = TN / (TN+FP)
PPV = TP / (TP+FP)
NPV = TN / (TN/FN)
Akurasi = (TrP+TN) / (TrP+TN+FP+FN)

TP : True Positive, bagian yang seharusnya diklasifikasikan dalam suatu kelas (missal A), dan memang dikenali sebagai kelas A

FP : False Positive, bagian yang seharusnya tidak diklasifikasikan dalam kelas A, dan akan tetapi dikenali sebagai kelas A.

TN : True Negative, bagian yang seharusnya dikenali sebagai selain kelas A, dan memang dikenali sebagai selain A.

FN : False Negative, bagian yang seharusnya diklasifikasikan dalam kelas A, dan akan tetapi tidak dikenali sebagai kelas A.

Dari hasil penelitian diketahui bahwa hasil klasifikasi pada data uji mengalami penurunan. Penurunan akurasi data uji antara lain disebabkan oleh Pencarian membandingkan antara kata yang ditemukan dalam dokumen dengan kata kunci secara identik, sehingga apabila terjadi sedikit perbedaan, akan terjadi kesalahan dalam klasifikasi, Output dari pdf2text tidak selalu sesuai dengan input, sehingga mempengaruhi proses pencarian dan hasil klasifikasi. Hal ini karena batas-batas dari masing-masing kelas adalah kata atau rangkaian kata tertentu, Jumlah hasil klasifikasi untuk kategori akibat, sebab, tanggapan, dan saran pemeriksaan adalah dengan dihitung berdasarkan jumlah kalimat output dalam tiap kategori. Perhitungan jumlah kalimat adalah dengan menghitung karakter titik (.) , titik dua (:), dan titik koma (;) dalam setiap kategori, karena karakter tersebut dianggap sebagai penanda akhir kalimat.

V. KESIMPULAN

Ektraksi Informasi pada dokumen LHP dapat dilakukan dengan menggunakan rule-based classification dengan akurasi data latih sebesar 98,27% dan akurasi data uji sebesar 89,77%. Penurunan akurasi ini antara lain disebabkan oleh kesalahan konversi pdf2text dan metode wordmatch yang digunakan dalam klasifikasi.

Metode yang digunakan dalam preprosessing sangat mempengaruhi hasil klasifikasi dan tingkat akurasinya. Penggunaan aplikasi pdf2text yang tidak menghasilkan file output yang identik dengan file input dapat menurunkan akurasi klasifikasi.

Rule-based classification dengan cara wordmatch dapat menurunkan akurasi. Cara yang dapat digunakan untuk meningkatkan akurasi adalah dengan menambahkan aturan-aturan dalam klasifikasi.

Engine yang dikembangkan dalam penelitian ini telah mengakomodir fungsi-fungsi ekstraksi informasi, yaitu *preprocessing* (pemisahan kata dan tanda baca, POS Tagger), pemilihan kata, dan ekstraksi dalam Temuan Pemeriksaan pada dokumen Laporan Hasil Pemeriksaan.

Tampilan aplikasi ini masih berupa *command prompt*, belum menggunakan *Graphical User Interface* (GUI).

VI. SARAN

Permasalahan yang menjadi dasar penurunan akurasi pada penelitian ini adalah penentuan pola dokumen dengan cara membandingkan bagian-bagian dokumen dengan kata kunci secara identik (*wordmatch*). Permasalahan ini kemungkinan dapat diselesaikan jika metode klasifikasi yang digunakan adalah membandingkan data dengan kata kunci tidak secara identik, tetapi dari segi kemiripannya (metode *stemming* dan *clustering*).

Konversi file .pdf menjadi .txt dengan menggunakan pdf2text tidak selalu menghasilkan output yang identik dengan file input. Penelitian berikutnya bisa dikembangkan dengan menggunakan metode atau tools konversi selain pdf2text untuk mendapatkan hasil *preprocessing* yang lebih akurat, akan lebih baik jika sebelum menentukan metode konversi yang akan digunakan, dilakukan perbandingan terlebih dahulu mengenai kelebihan dan kekurangan dari masing-masing metode konversi file .pdf menjadi file .txt.

Penelitian ini dapat dikembangkan dengan memanfaatkan metode *parsing* kalimat utnuk mendapatkan informasi pada LHP LKPD secara lebih detail (dapat diketahui masing-masing bagian kalimat : subyek, predikat, obyek, keterangan).

Penelitian ini dapat dikembangkan dengan membuat tampilan GUI untuk mempermudah interaksi dengan pengguna.

REFERENSI

[1] Chandrawati, T. 2008. Pengembangan Part of Speech Tagger untuk Bahasa Indoensia Berdasarkan Metode Conditional Random Fields dan Transformation Based Learning. *Universitas Indonesia*.

[2] Wicaksono, A. F., and A. Purwarianti. 2010. HMM Based Part-of-Speech Tagger for Bahasa Indonesia. *Institut Teknologi Bandung*.

[3] Jiang, J. 2012. Information Extraction from Text. In *Mining Text Data*, edited by C. C. Aggarwal and C. Zhai: Springer, 11-41.

[4] Feldman, R., and J. Sanger. 2006. *The Text Mining Handbook, Advances Approaches in Analyzing Unstructured Data*. Cambridge: Cambridge University Press.

[5] Firdaniza, N. Gusriani, and Akmal. Hidden Markov Model. Bandung: Universitas Padjadjaran.

[6] Palanisamy, S. K. 2006. Association Rule Based Classification, COmputer Science, Worcester Polytechnic Institute.

[7] Chaudhary, U. K., I. Papapanagiotou, and M. Devetsikiotis. Flow Classification Using Clustering and Association Rule Mining. *North Carolina State University*.