Pembangunan Kamus Bahasa Indonesia sebagai Sumber Daya Natural Language Processing (NLP) Bahasa Indonesia

Rizki Hilmi Fauzi[1]
Teknik Informatika
Universitas Indonesia
Jalan Dipatiukur No 114-116 Bandung 40132
rizkihill@gmail.com

Ken Kinanti Purnamasari[2]
Teknik Informatika
Universitas Indonesia
Jalan Dipatiukur No 114-116 Bandung 40132
Guinev_gs@yahoo.co.nz

***Abstrak** - Kamus adalah buku yang memuat kata-kata beserta keterangan maknanya, pemakaiannya atau terjemahannya. Kamus Besar Bahasa Indonesia terdiri dari lema, label kamus dan arti kata. Pembangunan kamus bahasa Indonesia sudah diteliti sebelumnya dalam "Pembangunan Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia" telah mendeteksi sebanyak 38.870 lema. Kemudian dilakukan pengembangan dalam penelitian berjudul "Pengembangan Pendeteksian Kamus Jenis Kata Sebagai Sebagai Sumber Daya NLP Bahasa Indonesia" telah mendeteksi sebanyak 43.060 lema. Selanjutnya dilakukan pengembangan dengan mendeteksi lema berupa kata majemuk dalam penelitian berjudul "Pengembangan Kamus Jenis Kata yang dilengkapi Kata Majemuk Sebagai Sumber Daya NLP Bahasa Indonesia" telah mendeteksi sebanyak 51.147 lema. Masalah dalam penelitian sebelumnya belum terdeteksinya label kamus selain jenis kata dan arti kata.*

Dari hasil analisis yang telah dilakukan, maka untuk dapat mendeteksi lema, label kamus dan arti kata dilakukan beberapa tahapan terdiri dari tahapan penyusunan data masukan yang bertujuan untuk memperbaiki data masukan agar memiliki pola entri yang sama dan tahapan pendeteksian lema, label kamus dan arti kata yang bertujuan untuk mendeteksi elemen-elemen kamus.

Hasil dari penelitian ini telah berhasil mendeteksi elemen kamus berupa lema, label kamus dan arti kata dan berhasil mendapatkan hasil lema sebanyak 51.972 lema dengan persentase sebesar 57.72% dari total jumlah 90.049 lema pada Kamus Besar Bahasa Indonesia edisi ke empat yang digunakan sebagai data masukan. Berdasarkan hasil analisis, peningkatan hasil pendeteksian terjadi karena proses pendeteksian menggunakan penanda tag html strong untuk lema dan tag html em untuk label kamus, sehingga dapat mendeteksi seluruh lema yang terdapat didata masukan.

Kata kunci : Kamus bahasa indonesia, kamus online, lema, Wordnet.

I. PENDAHULUAN

Kamus merupakan sumber rujukan yang andal dalam memahami makna kata suatu bahasa karena kamus memuat pembendaharaan kata suatu bahasa, yang secara ideal tidak terbatas jumlahnya. Kamus disusun dalam bentuk entri yang diurutkan berdasarkan abjad, disertai dengan arti kata dalam bentuk deskripsi dan contoh kalimat. Entri dalam kamus disertai dengan label kamus yang menjelaskan penggunaan suatu kata dan artinya. Bahasa Indonesia memiliki beberapa ragam kata yang diklasifikasikan berdasarkan bentuknya, diantaranya adalah kata dasar, kata turunan, kata majemuk, dan kata ulang.

Sumber daya kamus bahasa Indonesia sangat dibutuhkan untuk penelitian dibidang NLP (Natural Language Processing) bahasa Indonesia seperti tokenisasi awal pada POS Tag, Wordnet dan question answering. Saat ini kamus bahasa Indonesia tidak hanya berupa hardcopy, tetapi telah dilakukan pengembangan kamus bahasa Indonesia dalam bentuk digital berupa website antara lain; kbbidaring.info, kateglo.com, kamusbahasaindonesia.org dan lain-lain, tetapi saat ini pengembangan kamus secara digital tidak bersifat open resource. Hal ini mengakibatkan pengguna hanya bisa melakukan pencarian kata tanpa bisa mengunduh sumber daya kamus bahasa Indonesia, sehingga sumber daya kamus bahasa Indonesia yang sudah ada tidak bisa digunakan untuk penelitian dibidang NLP bahasa Indonesia.

Kamus Besar Bahasa Indonesia pada edisi ke empat yang diterbitkan pada tahun 2008 memuat 90.049 lema terdiri dari 41.250 kata dasar, 24.607 kata berimbuhan dan 23.536 gabungan kata. Pada penelitian pertama mengenai kamus bahasa Indonesia yang berjudul "Pembangunan Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia" yang dilakukan oleh C G Ceppy Efraim Bolly telah berhasil mendeteksi lema di dalam kamus Bahasa Indonesia dengan persentasi 43,17% [1]. Kemudian dilakukan pengembangan dengan menganalisis ulang tahap-tahapan yang dilakukan untuk dapat mendeteksi kata dan jenis kata oleh Arief Adiguna Putra dengan penelitian yang berjudul "Pengembangan

Pendeteksian Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia" telah berhasil mendeteksi lema dengan persentasi 47,82% [2]. Selanjutnya dilakukan pengembangan dengan menambahkan pendeteksian kata majemuk untuk meningkatkan jumlah kata yang terdeteksi dilakukan oleh Yoppy Yansyah dengan penelitian yang berjudul "Pengembangan Kamus Jenis Kata yang Dilengkapi Kata Majemuk Sebagai Sumber Daya NLP Bahasa Indonesia" dapat mendeteksi lema dengan persentase 57,80% [3]. Penelitian sebelumnya berfokus pada pendeteksian kata dan label kamus berupa jenis kata, sedangkan label kamus selain jenis kata dan arti kata tidak dideteksi. Hal ini mengakibatkan lema yang tidak memiliki jenis kata tidak terdeteksi, sehingga mempengaruhi jumlah lema yang berhasil terdeteksi. Selain itu, dikarenakan tidak mendeteksi arti kata hasil akhir kamus bahasa Indonesia tidak bisa digunakan untuk penelitian NLP bahasa Indonesia seperti tokenisasi awal pada POS Tag, Wordnet, question answering.

Maka pada penelitian ini akan dilakukan pendeteksi elemen kamus berupa lema, label kamus dan arti kata untuk meningkatkan jumlah lema yang berhasil terdeteksi. Selain itu, akan dilakukan analisis terhadap proses-proses yang dilakukan pada penelitian sebelumnya untuk mencari solusi terhadap permasalah yang terjadi pada penelitian sebelumnya.

II. ISI PENELITIAN

Berdasarkan pada latar belakang yang telah dipaparkan, bahwa pada penelitian sebelumnya tidak mendeteksi elemen kamus berupa label kamus selain jenis kata dan elemen kamus berupa arti kata, sehingga pada penelitian ini akan dilakukan pengembangan dengan mendeteksi kamus bahasa Indonesia yang disertai lema, label kamus dan arti kata.

A. Kamus

Secara etimologi, kata kamus berasal dari kata dalam bahasa Arab, yaitu *qamus* (bentuk jamaknya *qawamus*). Bahasa Arab menyerap kata kamus dari kata dalam bahasa Yunani kuno, okeanos yang berarti lautan. [4]

Kridalaksana menyebutkan bahwa kamus adalah buku referensi yang memuat daftar kata atau gabungan kata dengan keterangan mengenai berbagai segi maknanya dan penggunaannya dalam bahasa, kamus biasanya disusun menurut abjad. [4] Berdasarkan pengertian kamus yang dikemukakan oleh beberapa ahli di atas, dapat disimpulkan beberapa hal sebagai berikut.

1. Kamus termasuk buku referensi yang berisi kata-kata atau gabungan kata dari suatu bahasa;
2. Kata-kata tersebut disusun secara alfabetis;
3. Kata-kata tersebut diberi keterangan tentang makna dan penggunaannya;
4. Kata itu selain diberi keterangan maknanya, juga diberi keterangan tentang ucapannya, ejaannya, dan pelbagai hal lain;
5. Keterangan tentang makna itu diberikan juga dalam bahasa lain; Pada penelitian ini kamus digunakan sumber data masukan untuk memperoleh kata dan jenis kata.

Pada penelitian ini kamus digunakan sebagai sumber daya masukan untuk memperoleh kata, label kamus dan arti kata.

B. Analisis Masalah

Masalah dalam penelitian sebelumnya yang mempengaruhi jumlah lema yang berhasil dideteksi adalah penelitian sebelumnya hanya mendeteksi lema dan jenis kata, sehingga lema yang memiliki label kamus selain jenis kata, sebagai contoh label kamus ragam bahasa ark (arkais) tidak terdeteksi. Selain itu, lema yang memiliki arti kata lebih dari satu (homonim) hanya terdeteksi 1 lema dikarenakan tidak dilengkapi arti kata yang membedakan lema satu dengan lema lainnya, sebagai contoh lema 'A' berupa abjad pertama dalam bahasa Indonesia dengan 'a' berupa Ampere dalam satuan ukuran arus listrik. Selanjutnya adakan dilakukan analisis pada penelitian sebelumnya, sehingga dapat menyelesaikan permasalahan yang terjadi pada penelitian sebelumnya.

Tabel 1. Proses–Proses Penelitian Sebelumnya

No	**Penelitian Ceppy Bolly [1]**	**Penelitian Arief Adiguna Putra [3]**	**Penelitian Yoppy Yansyah [4]**
1	Penghilangan Digit Diawal Kalimat	Penghilangan Kalimat Dengan Delimeter Digit	-
2	-	-	Penghapusan Nomor Makna
3	Penghilangan Spasi Kosong Diawal Kalimat	Penghilangan Spasi Kosong Diawal Kalimat	-
4	Penghilangan Spasi Ganda	Penghilangan Spasi Ganda	-
5	-	-	Penghapusan Karakter Entitas *Html*
6	Penghilangan *Blank Line*	Penghilangan Baris Kosong	Penghapusan Baris Kosong
7	Penghilangan Baris Di awali Simbol	Penghilangan Baris Di awali Simbol	-
8	-	-	Penyamaan Simbol Kata Majemuk
9	-	-	Pengisian Kata dan Jenis Kata Untuk Kata Majemuk
10	-	-	Penyatuan Kata yang Terpisah
11	-	-	Pemisahan Kata Majemuk

12	Penghilangan Simbol Kecuali Strip	Penghilangan Simbol Kecuali Strip	Pembersihan
13	Pengambilan Kata dan Jenis Kata	Pengambilan Kata dan Jenis Kata	Pengambilan Kata dan Jenis Kata
14	Penghilangan Baris yang kurang dari dua kata	Penghilangan Baris yang kurang dari dua kata	-
15	Pengecekan Kata Perbaris	Pengecekan Kata Perbaris	Pemisahan Kata Dasar Dengan Kata Turunannya
16	Pengurutan	Pengurutan	-
17	-	Pengecekan Kata Duplikasi	-
18	-	-	Penghapusan Label Kecuali Label Jenis Kata
19	-	-	Pengkategorian

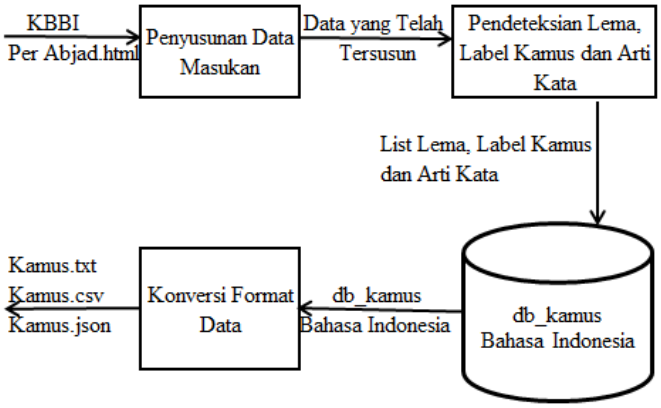
III. ANALISIS SOLUSI

Analisis solusi merupakan tahapan pemahaman terhadap sistem atau aplikasi yang dibuat dengan mengindentifikasikan permasalahan yang ada dan kebutuhan-kebutuhan yang diperlukan agar dapat membuat sistem yang lebih baik.

Pada penelitian proses pendeteksian lema, label kamus dan arti kata akan menggunakan metode *role-based*, dengan cara menentukan tahapan-tahapan yang diperlukan hingga mendapatkan hasil akhir berupa lema, label kamus dan arti kata. Pendeteksian lema, label kamus dan arti kata menggunakan fungsi dari *regular expression* atau yang sering disebut *regex* dengan cara mencari dan mengganti suatu kata berdasarkan pola kemunculannya di dalam kamus.

Pada penelitian ini akan menggunakan sumber masukan berformat *html* dikarenakan dapat mendeteksi penulisan cetak tebal (*bold*) dan cetak miring (*italic*). Didalam Kamus Besar Bahasa Indonesia penulisan cetak tebal digunakan untuk menulis lema, sedangkan penulisan cetak miring digunakan untuk menulis label kamus dan pribahasa. Sumber masukan berformat *html* akan mempermudah proses pendeteksian lema, label kamus dan arti kata sehingga mendapatkan hasil yang lebih baik.

Pada analisis solusi sistem yang akan dijalankan memiliki 2 tahapan yaitu tahapan penyusunan data masukan dan pendeteksian lema, label kamus dan arti kata, kemudian dimasukan kedalam database, selanjutnya data dapat dikonversi kedalam beberapa format data keluaran. Penjelasan setiap tahap dapat dilihat pada gambar berikut.



Gambar 1 Tahapan Umum Sistem

Pada gambar di atas Gambaran Tahapan Umum Sistem akan dijelaskan sebagai berikut.

1. Pada tahap awal memasukan sumber data yang merupakan Kamus Besar bahasa Indonesia yang sudah dibagi perabjad dari A sampai Z dengan format *.html* ke dalam sistem.
2. Pada tahap penyusunan data masukan akan dilakukan proses penghapusan karakter entitas *html*, penyatuan kata majemuk yang terpenggal, penghapusan *drop cap*, penghapusan tika, penyatuan arti kata dan penghapusan tanda cara baca. Tahapan ini bertujuan untuk menyusun pola data masukan yang sebelumnya terpenggal oleh spasi dan enter, sehingga proses pendeteksian mendapatkan hasil akhir yang lebih baik.
3. Selanjutnya, tahap pendeteksian kata, label kamus dan arti kata dilakukan proses pemisahaan kata sinonim, penggantian simbol majemuk, penggantian nomor arti kata dan pendeteksian elemen entri. Hasil akhir dari tahapan ini berupa list kata, label kamus dan arti kata.
4. Selanjutnya list kata, label kamus dan arti kata disimpan kedalam *database*. Dalam penelitian ini *database* yang digunakan adalah MySQL.
   Tahap terakhir yang dilakukan adalah konversi format data menjadi beberapa format yaitu *.txt*, *.csv*, *.json*.

A. Analisis Data Masukan

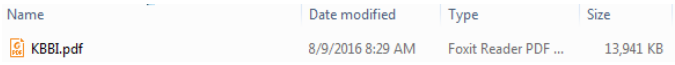
Analisis data masukan pada sistem pembangunan kamus bahasa Indonesia yaitu menjelaskan proses data masukan berupa data Kamus Besar Bahasa Indonesia dengan format *html* yang akan diproses oleh sistem. Data masukan Kamus Besar Bahasa Indonesia berformat *pdf* akan di konversi menjadi format *html* dengan 2 tahap, yaitu :

1. Pembagian File
   Data masukan yang pada awalnya berupa sebuah file dibagi menjadi perabjad dari A sampai Z dengan menggunakan program Nitro Pro 10. Proses ini bertujuan untuk menghindari gangguan teknis, saat sistem sedang berjalan dikarenakan banyaknya jumlah data yang

diproses secara bersamaan. Untuk lebih jelasnya digambarkan dalam bentuk blok diagram berikut.

Gambar 2 Blok Diagram Pembagian File

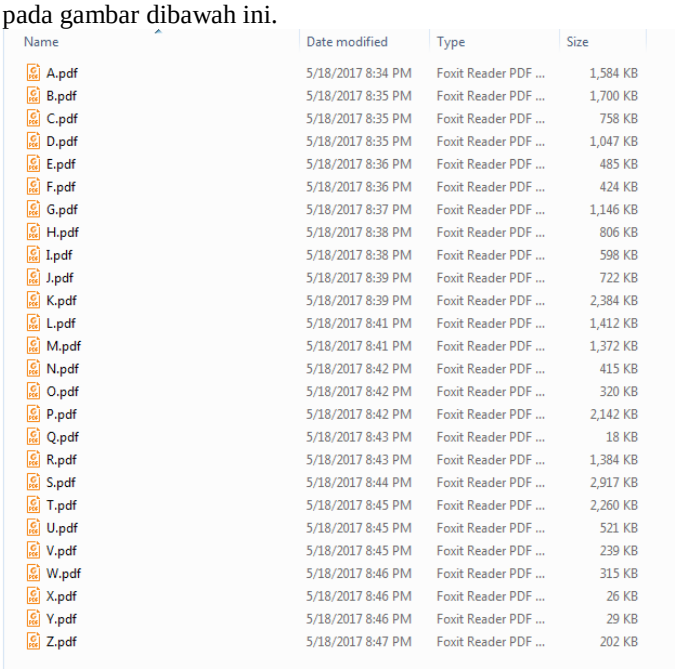
Berikut ini adalah contoh data masukan awal sebelum proses pembagian file yang dapat dilihat pada gambar berikut ini.



Name	Date modified	Type	Size
KBBI.pdf	8/9/2016 8:29 AM	Foxit Reader PDF ...	13,941 KB

Gambar 3 KBBI berformat *pdf*

Berikut ini adalah contoh data masukan setelah proses pembagian file menjadi KBBI perabjad yang dapat dilihat pada gambar dibawah ini.

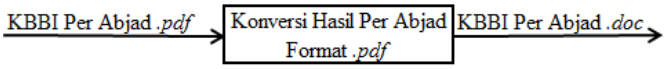


Name	Date modified	Type	Size
A.pdf	5/18/2017 8:34 PM	Foxit Reader PDF ...	1,584 KB
B.pdf	5/18/2017 8:35 PM	Foxit Reader PDF ...	1,700 KB
C.pdf	5/18/2017 8:35 PM	Foxit Reader PDF ...	758 KB
D.pdf	5/18/2017 8:35 PM	Foxit Reader PDF ...	1,047 KB
E.pdf	5/18/2017 8:36 PM	Foxit Reader PDF ...	485 KB
F.pdf	5/18/2017 8:36 PM	Foxit Reader PDF ...	424 KB
G.pdf	5/18/2017 8:37 PM	Foxit Reader PDF ...	1,146 KB
H.pdf	5/18/2017 8:38 PM	Foxit Reader PDF ...	806 KB
I.pdf	5/18/2017 8:38 PM	Foxit Reader PDF ...	598 KB
J.pdf	5/18/2017 8:39 PM	Foxit Reader PDF ...	722 KB
K.pdf	5/18/2017 8:39 PM	Foxit Reader PDF ...	2,384 KB
L.pdf	5/18/2017 8:41 PM	Foxit Reader PDF ...	1,412 KB
M.pdf	5/18/2017 8:41 PM	Foxit Reader PDF ...	1,372 KB
N.pdf	5/18/2017 8:42 PM	Foxit Reader PDF ...	415 KB
O.pdf	5/18/2017 8:42 PM	Foxit Reader PDF ...	320 KB
P.pdf	5/18/2017 8:42 PM	Foxit Reader PDF ...	2,142 KB
Q.pdf	5/18/2017 8:43 PM	Foxit Reader PDF ...	18 KB
R.pdf	5/18/2017 8:43 PM	Foxit Reader PDF ...	1,384 KB
S.pdf	5/18/2017 8:44 PM	Foxit Reader PDF ...	2,917 KB
T.pdf	5/18/2017 8:45 PM	Foxit Reader PDF ...	2,260 KB
U.pdf	5/18/2017 8:45 PM	Foxit Reader PDF ...	521 KB
V.pdf	5/18/2017 8:45 PM	Foxit Reader PDF ...	239 KB
W.pdf	5/18/2017 8:46 PM	Foxit Reader PDF ...	315 KB
X.pdf	5/18/2017 8:46 PM	Foxit Reader PDF ...	26 KB
Y.pdf	5/18/2017 8:46 PM	Foxit Reader PDF ...	29 KB
Z.pdf	5/18/2017 8:47 PM	Foxit Reader PDF ...	202 KB

Gambar 4 KBBI per abjad berformat *pdf*

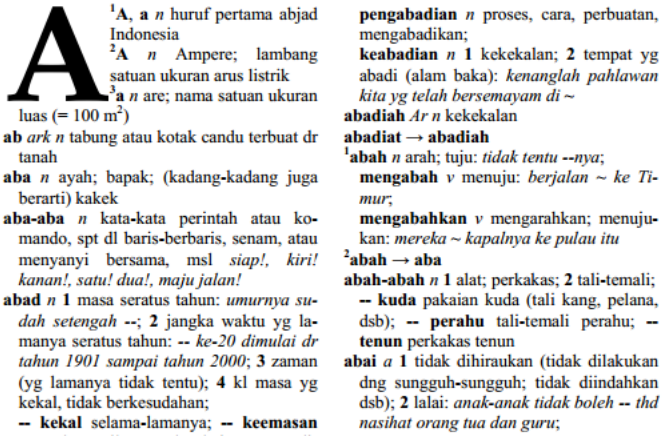
2. Konversi Data

Proses konversi data masukan dilakukan sebanyak 2 kali, pertama data masukan berformat *.pdf* di konversi menjadi berformat *.doc* dengan menggunakan program Nitro Pro 10. Proses ini dilakukan agar proses konversi menjadi format *.html* mendapatkan hasil yang lebih baik. Untuk lebih jelasnya digambarkan dalam bentuk blok diagram berikut.



Gambar 5 Blok Diagram Konversi Data Masukan Berformat *Doc*

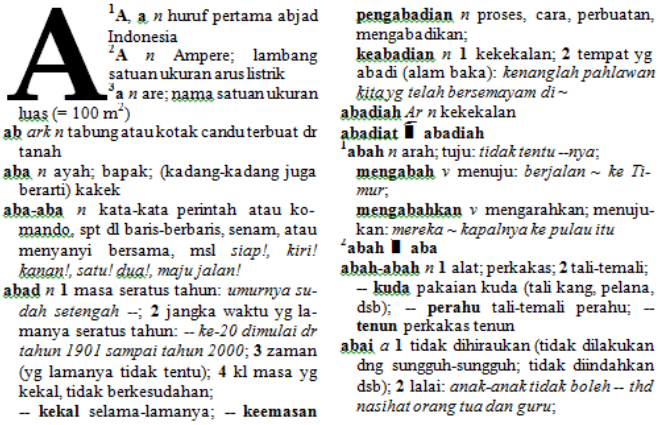
Berikut ini adalah contoh data masukan sebelum proses konversi data yang dapat dilihat pada gambar berikut



A, **a** *n* huruf pertama abjad Indonesia
A *n* Ampere; lambang satuan ukuran arus listrik
a *n* are; nama satuan ukuran luas (= 100 m^2)
ab *ark n* tabung atau kotak candu terbuat dr tanah
aba *n* ayah; bapak; (kadang-kadang juga berarti) kakek
aba-aba *n* kata-kata perintah atau komando, spt dl baris-berbaris, senam, atau menyanyi bersama, msl *siap!, kiri! kanan!, satu! dua!, maju jalan!*
abad *n* **1** masa seratus tahun: *umurnya sudah setengah* **--**; **2** jangka waktu yg lamanya seratus tahun: **--** *ke-20 dimulai dr tahun 1901 sampai tahun 2000*; **3** zaman (yg lamanya tidak tentu); **4** kl masa yg kekal, tidak berkesudahan;
-- kekal selama-lamanya; **-- keemasan** **pengabadian** *n* proses, cara, perbuatan, mengabadikan;
keabadian *n* **1** kekekalan; **2** tempat yg abadi (alam baka): *kenanglah pahlawan kita yg telah bersemayam di ~*
abadiah *Ar n* kekekalan
abadiat → abadiah
abah *n* arah; tuju: *tidak tentu --nya*;
mengabah *v* menuju: *berjalan ~ ke Timur*;
mengabahkan *v* mengarahkan; menujukan: *mereka ~ kapalnya ke pulau itu*
abah → aba
abah-abah *n* **1** alat; perkakas; **2** tali-temali;
-- kuda pakaian kuda (tali kang, pelana, dsb); **-- perahu** tali-temali perahu; **-- tenun** perkakas tenun
abai *a* **1** tidak dihiraukan (tidak dilakukan dng sungguh-sungguh; tidak diindahkan dsb); **2** lalai: *anak-anak tidak boleh -- thd nasihat orang tua dan guru*;

Gambar 7 Data Masukan berformat *pdf*

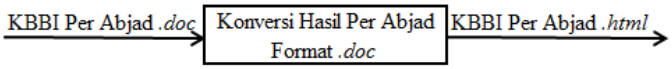
Berikut ini hasil konversi data masukan setelah proses konversi menjadi format *.doc* dengan menggunakan program Nitro Pro 10. Dapat dilihat pada gambar berikut.



A, **a** *n* huruf pertama abjad Indonesia
A *n* Ampere; lambang satuan ukuran arus listrik
a *n* are; nama satuan ukuran luas (= 100 m^2)
ab *ark n* tabung atau kotak candu terbuat dr tanah
aba *n* ayah; bapak; (kadang-kadang juga berarti) kakek
aba-aba *n* kata-kata perintah atau komando, spt dl baris-berbaris, senam, atau menyanyi bersama, msl *siap!, kiri! kanan!, satu! dua!, maju jalan!*
abad *n* **1** masa seratus tahun: *umurnya sudah setengah* --; **2** jangka waktu yg lamanya seratus tahun: -- *ke-20 dimulai dr tahun 1901 sampai tahun 2000*; **3** zaman (yg lamanya tidak tentu); **4** kl masa yg kekal, tidak berkesudahan;
-- **kekal** selama-lamanya; -- **keemasan** **pengabadian** *n* proses, cara, perbuatan, mengabadikan;
keabadian *n* **1** kekekalan; **2** tempat yg abadi (alam baka): *kenanglah pahlawan kita yg telah bersemayam di ~*
abadiah *Ar n* kekekalan
abadiat → abadiah
abah *n* arah; tuju: *tidak tentu --nya*;
mengabah *v* menuju: *berjalan ~ ke Timur*;
mengabahkan *v* mengarahkan; menujukan: *mereka ~ kapalnya ke pulau itu*
abah → aba
abah-abah *n* **1** alat; perkakas; **2** tali-temali;
-- **kuda** pakaian kuda (tali kang, pelana, dsb); -- **perahu** tali-temali perahu; -- **tenun** perkakas tenun
abai *a* **1** tidak dihiraukan (tidak dilakukan dng sungguh-sungguh; tidak diindahkan dsb); **2** lalai: *anak-anak tidak boleh -- thd nasihat orang tua dan guru*;

Gambar 7 Data Masukan berformat *Doc*

Selanjutnya proses konversi data masukan dari format *.doc* menjadi berformat *.html* dengan menggunakan *html cleaner*. Untuk lebih jelasnya digambarkan dalam bentuk diagram blok berikut.



Gambar 8 Blok Diagram Konversi Data Masukan Berformat *html*

Berikut ini adalah contoh data masukan setelah proses konversi data yang dapat dilihat pada gambar berikut ini.

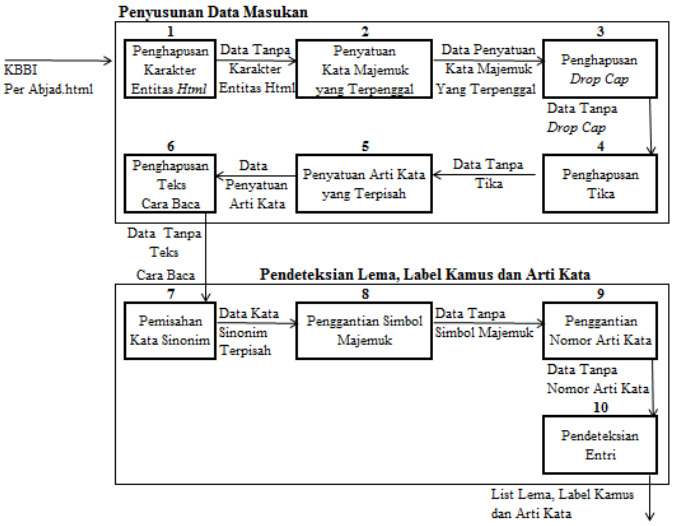
```
<table>
<tbody>
<tr>
<td width="84">
<table width="100%">
<tbody>
<tr>
<td>
<p><strong>A</strong></p>
</td>
</tr>
</tbody>
</table>
 </td>
</tr>
</tbody>
</table>
<p><strong>ab </strong><em>ark n </em>tabung atau kotak candu terbuat dr tanah</p>
<p><strong>abadiah </strong><em>Ar n </em>kekekalan</p>
<p><strong>abadiat </strong><strong>1 abadiah</strong></p>
<p><strong>abaimana </strong><em>ark n </em><strong>1 </strong>lubang pelepasan; dubur;</p>
<p><strong>2 </strong>kemaluan</p>
<p><strong>abc </strong>/ab&eacute;c&eacute;/ <em>n </em><strong>1 </strong>abjad Latin: <em>tidak tahu --</em>, tidak tahu
<p><strong>1acah </strong><em>v</em>, <strong>beracah-acah </strong>bermain-main; ber- pura-pura;</p>
<p><strong>2acah </strong><em>v, </em><strong>mengacah </strong>melanggar: <em>murid yg ~</em></p>
<p><em>akan dihukum</em></p>
<p><strong>acu </strong><em>v</em>, <strong>mengacu </strong><em>v </em><strong>1 </strong>mengangkat atau mengacungkan (ti
<p><strong>acuan </strong><em>n </em>yg diacu: <em>buku ~</em>, menjadi pe- doman (sumber) dl karangan dsb;</p>
<p><strong>~ sepatu </strong>alat untuk mengukur bentuk se- patu; </p>
<p><strong>asap </strong><em>n </em>gas yg tampak keluar dr sesuatu yg panas, mendidih, atau dibakar;</p>
<p><em>belum dipanjat </em>-- <em>kemenyan, pb </em>belum kawin;  <em>menggantang  </em>--,  <em>pb
<p>perbuatan yg sia-sia; berangan-angan yg</p>
<p>hampa;</p>
<p><strong>-- air </strong>gas yg keluar dr air yg mendidih; uap; -- <strong>api </strong>asap dr api;</p>
```

Gambar 9 Data Masukan berformat *html*

B. Analisis Proses

Analisis proses terdiri dari beberapa tahap yaitu, penyusunan data masukan dan pendeteksian kata, label kamus dan arti kata.

Tahapan penyusunan data masukan terdiri dari 6 proses yaitu proses penghapus karakter entitas *html*, penyatuan kata majemuk yang terpenggal, penghapusan *drop cap*, penghapusan tika, penyatuan arti kata yang terpisah dan penghapusan teks cara baca. Tahapan pendeteksian kata, label kamus dan arti kata terdiri dari 4 proses yaitu proses pemisahan kata sinonim, penggantian simbol majemuk, penggantian nomor arti kata, dan pendeteksian entri. Untuk lebih jelas dapat dilihat pada gambar berikut.



Gambar 10 Skema Proses

1. Tahap Penyusunan Data Masukan

Tahap penyusunan data masukan dilakukan untuk memperbaiki data masukan agar memiliki pola entri yang sama, yaitu lema, label kamus dan arti kata. Sehingga pada proses pendeteksian didapatkan hasil yang lebih baik.

Tahap penyusunan data masukan terdiri dari 6 proses yaitu penghapusan karakter entitas *html*, penyatuan kata majemuk yang terpenggal, penghapusan *drop cap*, penghapusan tika, penyatuan arti kata yang terpisah dan penghapusan teks cata baca.

Berikut ini contoh data masukan sebelum dan sesudah dilakukan tahap penyusunan data masukan.

Tabel 2. Contoh Tahap Penyusunan Data Masukan

Sebelum
abadiah Ar n kekekalan
abc /abécé/ n 1 abjad Latin: tidak tahu --, tidak tahu membaca huruf Latin; 2 ki hal- hal pokok yg paling pertama harus diketahui dr suatu keadaan atau perkara: belum tahu -- kehidupan
1acah v, beracah-acah bermain-main; ber- pura-pura;
2acah v, mengacah melanggar: murid yg ~
akan dihukum
asap n gas yg tampak keluar dr sesuatu yg panas, mendidih, atau dibakar;
belum dipanjat -- kemenyan, pb belum kawin; menggantang --, pb melakukan
perbuatan yg sia-sia; berangan-angan yg
hampa;
-- air gas yg keluar dr air yg mendidih; uap; -- api asap dr api;
mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan, mengha-rumkan, dsb) dng asap;
~ daging memanggang daging; ~ pakai-
an mengharumkan pakaian dng asap ra- tus dsb; ~ nyamuk mengusir nyamuk dng asap; mengasapi v 1 memberi asap;
2 memasak (mengharumkan dsb) dng asap;
Sesudah
abadiah Ar n kekekalan
abc n 1 abjad Latin: tidak tahu --, tidak tahu membaca huruf Latin; 2 ki hal- hal pokok yg paling pertama harus diketahui dr suatu keadaan atau perkara: belum tahu -- kehidupan
acah v, beracah-acah bermain-main; ber- pura-pura;
acah v, mengacah melanggar: murid yg ~ akan dihukum
asap n gas yg tampak keluar dr sesuatu yg panas, mendidih, atau dibakar; belum dipanjat -- kemenyan, pb belum kawin; menggantang --, pb melakukan perbuatan yg sia-sia; berangan-angan yg hampa;
-- air gas yg keluar dr air yg mendidih; uap; -- api asap dr api;
mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan, mengha-

rumkan, dsb) dng asap;

~ daging memanggang daging; ~ pakaian mengharumkan pakaian dng asap ra- tus dsb ~ nyamuk mengusir nyamuk dng asap; mengasapi v 1 memberi asap;

2 memasak (mengharumkan dsb) dng asap;

2. Tahap Pendeteksian Lema, Label Kamus Dan Arti Kata

Tahap pendeteksian lema, label kamus dan arti kata bertujuan untuk mengisi simbol majemuk dan nomor makna, setelah itu mendeteksi elemen entri yaitu lema, label kamus dan arti kata.

Tahap pendeteksian lema, label kamus dan arti kata terdiri dari 4 proses yaitu, proses pemisahan kata sinonim, proses penggantian simbol majemuk, proses penggantian nomor arti kata, dan pendeteksian entri.

Berikut ini contoh data sebelum dan sesudah proses pendeteksian lema, label kamus dan arti kata.

Tabel 3. Contoh Tahap Pendeteksian Lema, Label Kamus Dan Arti Kata

Sebelum

abadiah Ar n kekekalan

abc n 1 abjad Latin: tidak tahu --, tidak tahu membaca huruf Latin; 2 ki hal- hal pokok yg paling pertama harus diketahui dr suatu keadaan atau perkara: belum tahu -- kehidupan

acah v, beracah-acah bermain-main; ber- pura-pura;

acah v, mengacah melanggar: murid yg ~ akan dihukum

asap n gas yg tampak keluar dr sesuatu yg panas, mendidih, atau dibakar; belum dipanjat -- kemenyan, pb belum kawin; menggantang --, pb melakukan perbuatan yg sia-sia; berangan-angan yg hampa;

-- air gas yg keluar dr air yg mendidih; uap; -- api asap dr api;

mengasap v 1 menjadi asap; 2 memasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;

~ daging memanggang daging; ~ pakaian mengharumkan pakaian dng asap ra- tus dsb ~ nyamuk mengusir nyamuk dng asap; mengasapi v 1 memberi asap;

2 memasak (mengharumkan dsb) dng asap;

Sesudah

abadiah Ar n kekekalan

abc n abjad Latin: tidak tahu abc, tidak tahu membaca huruf Latin;

abc n ki hal- hal pokok yg paling pertama harus diketahui dr suatu keadaan atau perkara: belum tahu abc kehidupan

acah vbermain-main; ber- pura-pura;

beracah-acah vbermain-main; ber- pura-pura;

acah v melanggar: murid yg ~ akan dihukum

mengacah v melanggar: murid yg ~ akan dihukum

asap n gas yg tampak keluar dr sesuatu yg panas, mendidih, atau dibakar; belum dipanjat asap kemenyan, pb belum kawin; menggantang asap, pb melakukan perbuatan yg sia-sia; berangan-angan yg hampa;

asap air n gas yg keluar dr air yg mendidih; uap; asap api asap dr api;

mengasap v menjadi asap; asapmemasak (mengeringkan, membersihkan, mengha- rumkan, dsb) dng asap;

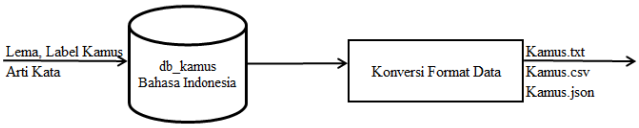
mengasap daging v memanggang daging; mengasap pakaian mengharumkan pakaian dng asap ra- tus dsb mengasap nyamuk mengusir nyamuk dng asap; mengasapi v memberi asap;

mengasapi v memasak (mengharumkan dsb) dng asap;

Pada pengujian akurasi akan dihitung nilai akurasinya dari jumlah lema yang didapat saat proses pendeteksian lema, label kamus dan arti kata. Jumlah lema yang didapat sebanyak 51.972 lema dari total jumlah lema pada Kamus Besar Bahasa

3. Konversi Format Output

Data hasil proses Pendeteksian kata, label kamus dan arti kata akan disimpan ke dalam *database* yang menghasilkan format *.sql*. selanjutnya akan dilakukan konversi data menjadi beberapa format data seperti : *.txt*, *.json*, *.csv*. Untuk lebih jelasnya digambarkan dalam bentuk blok diagram berikut ini.



Gambar 11 Blok Diagram Konversi Format Data Keluaran

IV. PENGUJIAN SISTEM

Tahap pengujian sistem bertujuan untuk menemukan kesalahan – kesalahan atau kekurangan – kekurangan pada sistem yang diuji. Pengujian bermaksud untuk mengetahui apakah program kamus bahasa Indonesia yang dibuat sesuai dengan tujuan penelitian.

A. Pengujianan Hasil Pendeteksian

Indonesia (KBBI) sebanyak 90.049 lema. Maka hasil perhitungannya sebagai berikut.

Jumlah total lema : 90.049

Total lema yang didapat : 51.972

Hasil Pengujian Pendeteksian :

= (Jumlah Lema Yang Didapat / Jumlah Lema KBBI) * 100%
= (51.972 / 90.049) * 100%
= 57.72%
Maka pada penelitian ini menghasilkan nilai akurasi sebesar 57.72%, dengan menghasilkan peningkatan sebanyak 825 lema dengan persentase sebesar 0.92%.

V. KESIMPULAN DAN SARAN

Pada bagian ini akan menjelaskan kesimpulan dan saran dari penelitian yang dilakukan.

A. Kesimpulan

Dalam hal ini Pembangunan Kamus Bahasa Indonesia Sebagai Sumber Daya Natural Language Processing Bahasa Indonesia telah mendapatkan hasil lema sebanyak 51.972 lema dengan persentase sebanyak 57.715% dari total lema didalam Kamus Besar Bahasa Indonesia (KBBI). Berdasarkan hasil analisis, peningkatan hasil pendeteksian terjadi karena pada proses pendeteksian memanfaatkan sumber masukan dengan format *html* yang dapat mendeteksi penulisan cetak tebal dan cetak miring. Namun peningkatan pendeteksian sekitar 825 lema dengan persentase sebesar 0.92% maka, perlu dilakukan pengecekan manual terhadap data masukan yang digunakan dalam penelitian ini.

B. Saran

Saran untuk penelitian berikutnya adalah meggunakan sumber masukan kamus terbaru, sehingga dapat meningkatkan lema yang berhasil terdeteksi. Pengembangan selanjutnya dilakukan dengan memisahkan ungkapan kata dan pribahasa yang terdapat didalam kamus.

REFERENSI

[1] Ceppy C G Efraim Bolly, "Pembangunan Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia," *Pembangunan Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia*, vol. I, Agustus 2016.

[2] Arief Adiguna Putra, "Pengembangan Pendeteksian Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia," *Pengembangan Pendeteksian Kamus Jenis Kata Sebagai Sumber Daya NLP Bahasa Indonesia*, vol. II, April 2017.

[3] Yoppy Yansyah, "Pengembangan Kamus Jenis Kata yang Dilengkapi Kata Majemuk Sebagai Sumber Daya NLP Bahasa Indonesia," *Pengembangan Kamus Jenis Kata yang Dilengkapi Kata Majemuk Sebagai Sumber Daya NLP Bahasa Indonesia*, April 2017.

[4] Badan Bahasa. Badan Pengembangan dan Pembinaan Bahas. [online]. https:/badanbahasa.kemdikbud.go.id/