# EKSTRAKSI KATA KUNCI OTOMATIS UNTUK DOKUMEN BAHASA INDONESIA STUDI KASUS: ARTIKEL JURNAL ILMIAH KOLEKSI PDII LIPI

Diana Permata Sari[1]* dan Ayu Purwarianti[2]

[1]Lembaga Ilmu Pengetahuan Indonesia
[2]Institut Teknologi Bandung

*Korespondensi: dianapermata@ymail.com

## ABSTRACT

Keyword determination by using controlled vocabulary is not a difficult task for information analysts. However, specify keywords for hundreds or even thousands of articles will take time and effort of the analysts. To ease the work, it needs to be made a system of automatic keyword extraction. The construction of this system passes the stages of preprocessing, translating, and pinpointing keyword candidates with a list of keywords. The research was carried out by using 33 articles taken from PDII LIPI journal collections. This research employed 3 weighing methods, namely TF, TF x IDF and WIDF. The best result was obtained from TF x IDF method. To refine the result, the author carried out fixing the keywords results and using levensthein algorithm.

## ABSTRAK

Penentuan kata kunci dengan menggunakan kosakata terbatas *(controlled vocabulary)* merupakan pekerjaan yang tidak sulit bagi para analis informasi. Namun, menentukan kata kunci untuk ratusan bahkan ribuan artikel, akan memakan waktu dan tenaga yang tidak sedikit bagi para analis informasi. Untuk meringankan pekerjaan tersebut, dibangun sebuah sistem ekstraksi kata kunci otomatis. Pembangunan sistem melewati tahapan praproses, translasi, dan pencocokan kandidat kata kunci dengan daftar kata kunci. Eksperimen dilakukan dengan menggunakan 33 data artikel yang diambil dari kumpulan jurnal koleksi PDII LIPI. Eksperimen ini menggunakan 3 metode pembobotan, yaitu TF, TFxIDF dan WIDF. Hasil terbaik didapat dari pembobotan TFxIDF. Untuk menyempurnakan hasil eksperimen, dilakukan perbaikan pada daftar hasil kata kunci dan penggunaan algoritma levensthein.

**Keywords:** Automatic indexing; Controlled vocabularies; Keywords searching

## 1. PENDAHULUAN

Kata kunci merupakan satu atau beberapa istilah yang dianggap penting dan digunakan untuk kemudahan temu kembali dokumen yang diambil/diekstrak dari judul, abstrak atau isi artikel (Hartinah, 2002). Terdapat dua cara yang bisa digunakan untuk menentukan kata kunci. Pertama, penulis dapat memilih secara bebas kata kunci yang menggambarkan isi dari dokumen (Gazendam, 2010). Kedua, kata kunci diambil dari kosa kata terbatas atau tesaurus (Gazendam, 2010).

Pusat Dokumentasi dan Informasi Ilmiah - Lembaga Ilmu Pengetahuan Indonesia yang lebih dikenal dengan PDII LIPI merupakan sebuah lembaga yang melakukan pengolahan dokumen, termasuk pengindeksan dengan menggunakan kosakata terkontrol *(controlled vocabulary)*. Artikel yang akan diolah dibaca satu persatu oleh para pengindeks atau analis informasi. Kemudian, tiap artikel diringkas dan ditentukan kata kuncinya. Untuk dokumen yang jumlahnya tidak terlalu banyak, jenis pekerjaan seperti ini tidak akan terlalu menyulitkan. Namun, untuk dokumen yang jumlahnya ribuan, pekerjaan seperti ini menjadi satu hal yang sangat membosankan. Untuk itu, diperlukan adanya sebuah solusi yang bisa membantu menangani permasalahan ini. Dalam hal ini, ekstraksi kata kunci secara otomatis

untuk artikel berbahasa Indonesia merupakan satu kebutuhan yang dapat direalisasikan.

Terdapat banyak metode yang dapat digunakan untuk membangun sistem ekstraksi kata kunci otomatis. Kegiatan ini dilakukan dengan tujuan untuk mencari metode terbaik untuk sistem ekstraksi kata kunci otomatis. Data yang digunakan salam kajian ini berupa 33 artikel jurnal koleksi PDII LIPI.

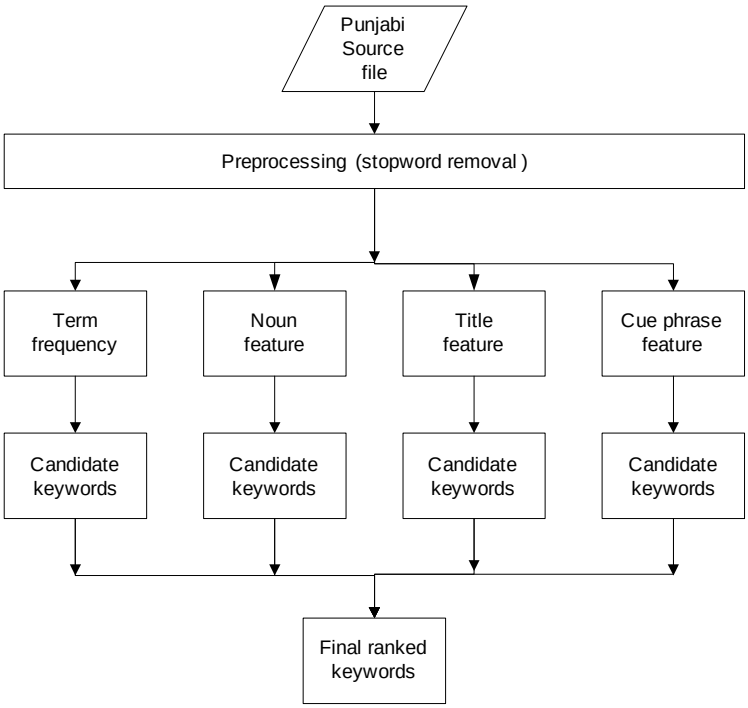
## 2. TINJAUAN PUSTAKA

Pemrosesan bahasa alami merupakan interdisipliner bidang dari ilmu komputer dan linguistik yang membahas interaksi antara bahasa (alami) manusia dengan komputer. Ekstraksi kata kunci merupakan bagian dari cakupan konsep pemrosesan bahasa alami (Kaur, 2011).

Lebih lanjut Kaur (2011) menyebutkan ada empat metode yang digunakan untuk melakukan ekstraksi kata kunci otomatis, yaitu sebagai berikut:

1) Pendekatan statistik sederhana
   Pendekatan ini sederhana dan tidak memerlukan *data training*. Informasi statistik dari kata dapat digunakan untuk mengidentifikasi kata kunci pada dokumen. Metode ini mencakup frekuensi kata, TFxIDF, *word co-occurance*, dll.
2) Pendekatan linguistik
   Pendekatan ini menggunakan fitur linguistik dari kata dan dokumen. Pendekatan ini mencakup analisis leksikal, analisis siktaksis, dll.
3) Pendekatan *machine learning*
   Pendekatan ini dapat dilihat sebagai *supervised learning*. Pendekatan *machine learning* menggunakan kata kunci yang terekstrak dari dokumen *training* untuk mempelajari sebuah model dan menerapkan model tersebut untuk menemukan kata kunci dari dokumen baru. Pendekatan ini mencakup Naïve Bayes, Support Vector Machine, dll
4) Pendekatan lain
   Pendekatan lain dari ekstraksi kata kunci dengan mengombinasikan metode yang disebutkan di atas atau menggunakan pengetahuan heuristik, seperti posisi, panjang, *layout* fitur dari kata-kata, tag html di sekitar kata, dll.

Beberapa pendekatan dilakukan untuk mengekstrak kata kunci dokumen dalam berbagai ragambahasa. Pendekatan yang dilakukan oleh Kamaldeep Kaur untuk mengekstrak kata kunci bahasa Punjabi meliputi: a) *noun chunking*; b) jumlah kemunculan *term*; c) keberadaan *term* dalam judul; d) frase seperti kata-kata dalam kalimat yang mengandung “frase ringkasan” atau “frase transisi” (Kaur, 2011). Gambar 1 berikut ini memperlihatkan arsitektur sistem untuk mengekstrak kata kunci bahasa Punjabi.




Gambar 1. Arsitektur sistem ekstraksi kata kunci untuk bahasa Punjabi (Kaur, 2011)

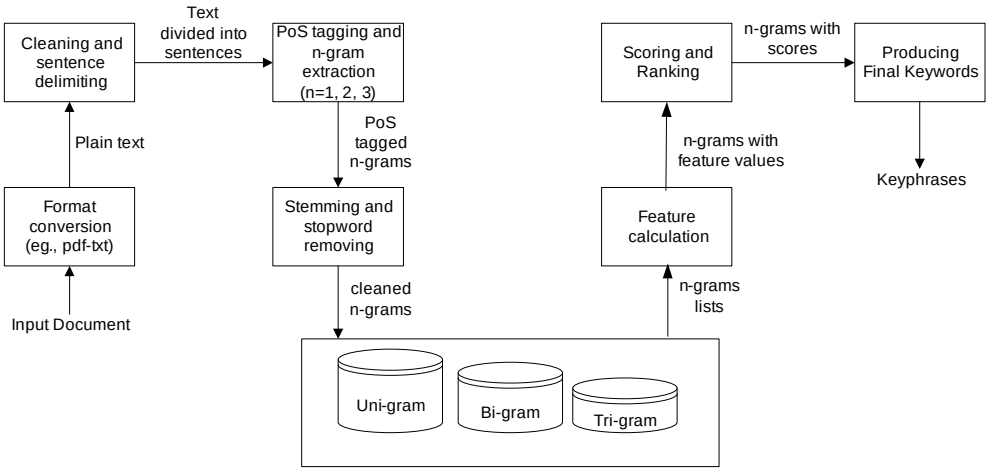
Pendekatan lain untuk mengekstrak kata kunci berbahasa Inggris adalah seperti yang digunakan oleh Hulth (2003), yaitu sebagai berikut:

1) N-gram, yaitu mengekstraksi seluruh unigram, bigram, dan trigram.
2) NP chunks, yaitu mengekstraksi seluruh *noun phrase chunk* dalam dokumen.
3) Pola POSTAG, dilakukan untuk mengekstraksi kata yang tidak berhasil di-*chunk*. Proses ini mengekstrak kata atau urutan kata berdasarkan pola POSTAG yang sudah didefinisikan sebelumnya.

Selain pendekatan yang dikemukakan Hulth di atas, Pudota (2010) juga mengatakan bahwa terdapat pendekatan lain yang digunakan untuk mengekstraksi kata kunci, yaitu:

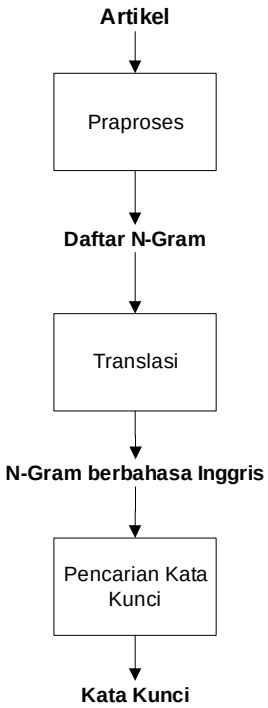
1) Ekstraksi kandidat kata kunci, dengan tahapan: a) konversi; b) pembersihan dan pembatasan kalimat; c) POSTAG dan ekstraksi N-Gram; d) *stemming* dan *stopword removal*; serta e) memisahkan daftar N-Gram.
2) Kalkulasi fitur. Fitur yang digunakan dalam tahapan ini, adalah frekuensi frase, nilai pos, kedalaman frase, frase terakhir muncul dan umur frase.
3) Pembobotan dan pengurutan. Dalam tahapan ini dilakukan pembobotan dan pengukuran yang nantinya akan menunjukkan kata/frase yang paling representatif dalam dokumen.

Arsitektur dari ektraksi kata kunci yang diajukan Pudota diatas, dapat dilihat pada Gambar 2 berikut ini.




Gambar 2. Arsitektur sistem ekstraksi kata kunci (Pudota, 2010)

Dokumen berupa artikel ilmiah yang masuk ke PDII LIPI tidak memiliki aturan struktur penulisan yang baku. Namun, secara umum, struktur penulisan dokumen pada data yang terkumpul terdiri dari judul, abstrak, pendahuluan, pembahasan, kesimpulan, dan daftar pustaka. Sesuai dengan definisi kata kunci yang telah diungkapkan oleh Hartinah (2002) bahwa bagian dari struktur penulisan artikel yang diambil sebagai poin penting untuk menentukan kata kunci adalah judul, abstrak atau isi dari artikel. Namun, dalam perancangan sistem ekstraksi kata kunci otomatis untuk dokumen berbahasa Indonesia, ditambahkan lagi satu unsur, yaitu kata kunci yang ditentukan oleh penulis. Dengan demikian, unsur penting untuk menentukan kata kunci terdiri dari judul, abstrak, dan kata kunci yang ditentukan oleh penulis. Sementara itu sumber yang dijadikan sebagai acuan untuk menentukan kata kunci dengan cara *controlled vocabullary* adalah daftar kata kunci yang dibangun oleh para analis informasi PDII LIPI. Tahapan yang dilalui dalam proses ekstraksi kata kunci dapat dilihat pada Gambar 3 berikut ini.




Gambar 3. Tahapan ekstraksi kata kunci otomatis

## 3. METODE

Kajian ini dilaksanakan untuk kajian sistem. Dalam pelaksanaannya, metode yang dilaksanakan meliputi tiga tahapan, yaitu: a) praproses; b) translasi; c) pencocokan kandidat kata kunci dengan daftar kata kunci. Tahapan praproses terdiri dari proses pemecahan kalimat menjadi kata (tokenisasi), pelabelan kata (POSTAG), dan pencarian kata/frase kandidat kata kunci (pencarian N-gram). Setelah didapatkan frase kandidat kata kunci, selanjutnya dilakukan proses penyaringan kata/frase untuk meyakinkan bahwa kata/frase yang diujikan merupakan bagian dari judul atau kata kunci yang ditentukan oleh penulis artikel. Kata/frase ini kemudian melalui proses pembobotan dengan menggunakan tiga teknik pembobotan, yaitu TF, TFxIDF, dan WIDF. Setiap kata yang telah diberi bobot diurutkan berdasarkan nilai tertinggi. Setelah itu, dilakukan proses translasi dan proses pemeriksaan kandidat kata kunci dalam daftar kata kunci. Kata/frase yang memiliki pasangan yang cocok dalam daftar kata kunci merupakan kata kunci yang dihasilkan oleh sistem.

### 3.1 Praproses

Tahapan praproses dokumen dalam sistem ini terdiri dari proses tokenisasi, POSTAG, dan ekstraksi frase. Dalam proses tokenisasi kalimat, dokumen dipecah menjadi unit terkecil (token), yaitu kata. Sistem akan memisahkan kata berdasarkan tanda baca dan spasi.

Kata-kata yang terkumpul hasil dari proses tokenisasi selanjutnya dikumpulkan dalam sebuah file untuk berikutnya dilakukan proses POSTAG. Dalam proses ini. POSTAG yang digunakan adalah POSTAG Bahasa Indonesia dengan metode HMM yang dibangun oleh Wicaksono dan Purwarianti (2010). Kelas kata yang diperlukan sebagai kandidat kata kunci adalah kata benda yang dalam POSTAG Alfan Farizki Wicaksono disebut dengan NN *(Common Noun)*, NNP *(Proper Noun)* dan NNG *(Genitive Noun)*, kata asing *(FW/Foreign Words*) dan kata sifat *(JJ/Adjective)*.

Kata-kata yang telah memiliki kelas kata selanjutnya akan melalui tahapan ekstraksi frase sebagai kandidat kata kunci. Dalam proses ini, dihasilkan kandidat kata kunci yang terdiri dari satu kata *(unigram)*, dua kata *(bigram)*, dan tiga kata *(trigram)*. Proses ini disebut juga dengan proses pencarian N-Gram. Kata/frase hasil dari pencarian N-Gram ini selanjutnya akan melalui proses penyortiran untuk meyakinkan bahwa kata/frase yang akan masuk ke proses berikutnya adalah kata/frase yang termasuk ke dalam unsur penting untuk menentukan kata kunci (judul, abstrak dan kata kunci yang ditentukan oleh penulis).

### 3.2 Translasi (Penerjemahan)

Proses ini merupakan tahapan yang menghasilkan keluaran kandidat kata kunci berbahasa Inggris. Tujuan dilakukan proses ini adalah mencocokkan istilah dalam artikel yang berbahasa Indonesia dengan daftar kata kunci yang istilah-istilahnya berbahasa Inggris. Sebelum dilakukan proses translasi, kata/frase yang telah melalui tahapan praproses diberi bobot untuk mengetahui kata/frase yang paling berkualitas untuk dijadikan kandidat kata kunci. Proses pembobotan menggunakan tiga metode, yaitu TF, TFxIDF dan WIDF. Ketiga metode pembobotan ini digunakan untuk mencari metode pembobotan terbaik yang dapat digunakan untuk membangun sistem ekstraksi kata kunci otomatis.

Manning (1999) menyatakan bahwa pendekatan yang paling sederhana untuk memberikan bobot pada sebuah *term* adalah jumlah kemunculan *term* t pada dokumen d. Selain itu, terdapat

juga istilah *document frequency* dft, yang didefinisikan sebagai jumlah dokumen dalam koleksi yang mengandung term t. Kemudian, untuk menghitung pembobotan *term* dengan menggunakan *document frequency* dalam sebuah koleksi dokumen N, digunakan *inverse document frequency.* Dengan demikian, perhitungan TFxIDF diperoleh dengan cara mengalikan *term frequency* dengan *inverse* dari *document frequency*.

WIDF *(Weighted Inverse Document Frequency)* merupakan lanjutan dari IDF untuk dengan mempertimbangkan TF dalam koleksi dokumen. WIDF dikembangkan untuk menutupi kekurangan IDF, dimana kata yang muncul pada tiap dokumen diperlakukan sama. IDF tidak membedakan kemunculan sebuah kata pada satu dokumen dan banyak dokumen (Tokunaga, 1994). Pembobotan kata dengan menggunakan WIDF membedakan bobot kata pada dokumen $d_1$ sampai dengan dokumen $d_1$.

Setelah melalui tahapan pembobotan, dalam proses ini kata/frase yang termasuk dalam judul, abstrak dan kata kunci yang ditentukan oleh penulis diberi bobot ganda (bobot normal dikali dua). Hal ini dilakukan untuk menandai bahwa kata/frase tersebut merupakan kata/frase yang dianggap penting untuk dijadikan sebagai kandidat kata kunci. Kata/frase yang telah diberi bobot kemudian diurutkan berdasarkan nilai tertinggi untuk masing-masing N-Gram. Kata/frase yang diambil untuk masuk ke proses berikutnya adalah 5 kata/frase yang memiliki nilai tertinggi untuk masing-masing pembobotan.

Setelah melalui proses pembobotan, selanjutnya dilakukan proses translasi. Metode yang digunakan dalam proses ini yaitu menarik informasi hasil translasi dari situs http://translate.google.com. Data yang diambil berupa hasil translasi dan alternatif/ sinonimnya.

### 3.3 Pencarian Kata Kunci

Hasil translasi yang didapatkan selanjutnya dicocokkan dengan daftar kata kunci yang telah tersedia. Jika tidak ditemukan istilah yang cocok dalam daftar kata kunci maka proses alternatif dieksekusi, yaitu menjalankan proses pencarian istilah yang paling dekat dengan kata kunci menggunakan algoritma levenshtein. Keluaran dari proses ini merupakan hasil akhir dari sistem.

## 4. HASIL DAN PEMBAHASAN

Data yang digunakan dalam pengujian sistem ini adalah sebanyak 33 dokumen yang diambil dari pangkalan data jurnal ilmiah Indonesia (ISJD). Data ini tersedia dalam bentuk format PDF yang terbentuk dari hasil scan dokumen cetak. Untuk menyesuaikan dengan sistem yang memerlukan masukan dalam bentuk teks biasa, data ini dikonversi disesuaikan dengan kebutuhan sistem. Pada dokumen yang hasil konversinya tidak sempurna, isi dokumen diperbaiki sedemikian rupa sehingga dokumen siap menjadi masukan tanpa kesalahan untuk sistem.

Pengujian dilakukan untuk membandingkan nilai akurasi dari penggunaan tiga metode pembobotan (TF, TFxIDF, dan WIDF). Tabel 1 memperlihatkan hasil pengujian tiga metode pembobotan terhadap satu dokumen.

Tabel 1. Pengurutan Kata/Frase Berdasarkan Nilai TF, TFxIDF dan WIDF

| TF | | TFxIDF | | WIDF | |
|---|---|---|---|---|---|
| **Kata/Frase** | **Nilai TF** | **Kata/Frase** | **Nilai TFxIDF** | **Kata/Frase** | **Nilai WIDF** |
| scada | 6 | scada | 18.2222 | *mixing* | 1 |
| *manager* | 6 | wonderware | 15.1851 | *controller* | 1 |
| wonderware | 5 | *manager* | 12.4967 | *supervisory* | 1 |
| program | 4 | intouch | 12.1481 | *and* | 1 |
| recipe | 4 | recipe | 12.1481 | *acquisiition* | 1 |
| wonderware intouch | 4 | wonderware intouch | 12.1481 | baik scada | 1 |
| sql access | 3 | sql access | 9.11108 | wonderware in | 1 |
| program scada | 3 | recipe manager | 9.11108 | *software* wonderware | 1 |
| *recipe manager* | 3 | *access manager* | 9.11108 | ornron c200hg | 1 |
| *access manager* | 3 | program scada | 4.55554 | plc ornron | 1 |
| sql *access manager* | 3 | sql access manager | 9.11108 | perancangan scada *software* | 1 |
| proses pembuatan kertas | 2 | proses pembuatan kertas | 3.03703 | komposisi campuran bahan | 1 |
| perancangan scada *software* | 1 | perancangan scada software | 1 | wonderware intouch | 1 |
| komposisi campuran bahan | 1 | komposisi campuran bahan | 1 | *software* wonderware in | 1 |
| wonderware in touch | 1 | wonderware in touch | 1 | plc ornron c200hg | 1 |

Berdasarkan Tabel 1 di atas, dapat dilihat bahwa nilai tertinggi diperoleh dari hasil pembobotan dengan menggunakan TFxIDF. Selanjutnya, urutan kata/frase berdasarkan bobot tertinggi ini dihitung nilai ketepatannya jika dibandingkan dengan kata kunci hasil manual yang dikerjakan oleh analis informasi PDII-LIPI. Hasil yang diperoleh dapat dilihat pada Tabel 2 berikut ini.

Tabel 2. Perbandingan Nilai Akurasi Sistem dengan Tiga Pembobotan

| | **Jumlah Kata Kunci Cocok dengan Manual** | | | **Jumlah Kata Kunci yang Seharusnya Ditemukan** |
|---|---|---|---|---|
| | **TF** | **TFxIDF** | **WIDF** | |
| Jumlah | 36 | 38 | 21 | 124 |
| % akurasi | 29.03% | 30.65% | 16.94% | |

Dari tabel 2 di atas, dapat dilihat bahwa nilai akurasi tertinggi diperoleh dari hasil pembobotan dengan menggunakan TFxIDF dengan nilai ketepatan/akurasi sebesar 30,65%.

Melihat nilai akurasi yang terhitung masih rendah selanjutnya dilakukan analisis terhadap faktor-faktor yang mempengaruhi kerendahan nilai akurasi yang diperoleh. Hasil yang diperoleh menunjukkan bahwa terdapat beberapa faktor yang berpengaruh. Faktor-faktor tersebut antara lain terdapat pada proses: a) POSTAG, yaitu pemberian kelas kata yang tidak tepat; serta b) pencocokan dengan daftar

kata kunci, ketidaksesuaian antara istilah yang diperoleh dari artikel dengan istilah dalam daftar kata kunci.

Untuk mengatasi hal tersebut, maka dilakukan proses *update* kelas kata pada Lexicon POSTAG, *update* istilah dalam daftar kata kunci, dan penggunaan algoritma kedekatan kata levenshtein untuk meningkatkan nilai akurasi. Hasil yang diperoleh setelah dilakukan perbaikan dapat dilihat pada Tabel 3.

Tabel 3. Perbandingan Nilai Akurasi Sistem dengan Tiga Pembobotan Setelah Dilakukan Perbaikan terhadap Unsur-Unsur yang Mempengaruhi Nilai Akurasi

| | **Jumlah Kata Kunci Cocok dengan Manual** | | | **Jumlah Kata Kunci yang Seharusnya Ditemukan** |
|---|---|---|---|---|
| | **TF** | **TFxIDF** | **WIDF** | |
| Jumlah | 36 | 38 | 21 | 124 |
| % akurasi | 45,16% | 45,97% | 32,26% | |

Pada Tabel 3 di atas, memperlihatkan bahwa setelah dilakukan perbaikan, nilai akurasi meningkat lebih tinggi dan tetap diperoleh dari hasil pembobotan TFxIDF.

## 5. KESIMPULAN

Kesimpulan yang dapat diambil dari hasil eksperimen adalah bahwa tahapan yang dilakukan untuk mengekstrak kata kunci dokumen berbahasa Indonesia secara otomatis, secara umum adalah tahapan praproses, tahapan translasi, dan tahapan pencocokan kandidat kata kunci dengan daftar kata kunci. Nilai akurasi terbaik yang diperoleh dari hasil keluaran sistem didapat dengan menggunakan teknik pembobotan TFxIDF. Dengan melakukan *update* pada lexicon POSTAG dan daftar kata kunci serta penggunaan metode kedekatan string levensthein, nilai akurasi yang diperoleh meningkat lebih tinggi dari hasil sebelumnya.

Saran yang dapat diberikan untuk penelitian selanjutnya adalah penggunaan teknik pembobotan lain selain tiga metode yang digunakan dalam penelitian ini; penggunaan daftar kata kunci yang lebih *up to date* dan penggunaan teknik ekstraksi kata kunci yang lain, seperti penggunaan *machine learning* dsb.

## DAFTAR PUSTAKA


Gazendam, L., Waterna, C., Brussee, R. 2010. Thesaurus based term ranking for keyword extraction. 7th International Workshop on Text-based Information Retrieval.

Hartinah, S. 2002. Penggunaan Dalil Zipf pada pengindeksan otomatis. Kumpulan Makalah Kursus Bibliometrika. Depok: Masyarakat Informetrika Indonesia.

Hulth, A. 2003. Improved automatic keyword extraction given more linguistic knowledge. *Proceedings of the 2003 Conference on Emprical Methods in Natural Language Processing. 216-223*.

Kaur, K., Vishal, G. 2011. Keyword extraction for punjabi language. *Indian Journal of Computer Science and Engineering (IJSCE)*. 2 (3), 364-370.

Manning, C. D., Raghavan, P., Scütze, H. 1999. An Introduction to Information retrieval. England: Cambridge University Press.

Pudota, N., Dattolo, A., Baruzzo, A., Tasso, C. 2010. A New Domain independent keyphrase extraction system. Italian Research Conference on Digital Library Management Systems - IRCDL.

Tokunaga, T., Iwayama, M. 1994. Text Categorization based on weighted inverse document frequency. Technical Report.

Wicaksono, A. F. and Purwarianti, A. 2010. HMM Based part-of-speech tagger for bahasa indonesia. Proceedings of the 4th- International MALINDO Workshop (MALINDO). Jakarta, Indonesia.