Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.201861226
Vol. 6, No. 1, Februari 2019, hlm. 41-48 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 30/E/KPT/2018 e-ISSN: 2528-6579

EVALUASI DAFTAR STOPWORD BAHASA INDONESIA

Faisal Rahutomo[1], Ariadi Retno Tri Hayati Ririd[2]

[1,2]Politeknik Negeri Malang
Email: [1]faisal@polinema.ac.id, [2]ariadi.retno@polinema.ac.id

(Naskah masuk: 11 November 2018, diterima untuk diterbitkan: 18 Desember 2018)

Abstrak

Pada sistem temu kembali informasi berbentuk teks maupun *text mining*, terdapat proses pengindeksan. Teks diproses dengan tujuan mengintisarikan informasi berbentuk teks tersebut. Salah satu proses yang dilakukan adalah *stopword filtering*, beberapa kata yang tidak layak diindeks diabaikan berdasar sebuah daftar. Di dalam sistem berbahasa Indonesia, terdapat beberapa versi daftar *stopword* yang tersedia bebas. Penelitian ini bertujuan mengevaluasi daftar yang telah tersedia tersebut. Tujuan akhir dari penelitian ini adalah telaah daftar yang tersedia berdasarkan tata bahasa Indonesia, cara penyusunan, dan kebiasaan perambah internet. Dari hasil telaah diperoleh fakta bahwa daftar yang tersedia dibangun dengan analisis frekuensi kemunculan kata pada sebuah korpus (*corpus*) teks, tanpa memperhatikan jenis kata ataupun kebiasaan pengguna internet. Hasil lain penelitian ini adalah beberapa rekomendasi lebih lanjut bagi para peneliti di bidang ini ketika membutuhkan daftar *stopword* bahasa Indonesia, yaitu daftar yang memperhatikan jenis kata dan kebiasaan pengguna internet melalui mesin perambah yang tersedia.

Kata kunci: *daftar stopword, bahasa Indonesia, temu kembali informasi, text mining, evaluasi*

INDONESIAN STOPWORD LIST EVALUATION

Abstract

Most of text-based information retrieval system uses indexing process. The system processes the texts in order to obtain the information essence. One of the process is stopword filtering, several words are being ignored based on a stopword list. Several Indonesian stopword list are available openly. Therefore, this paper evaluates the available lists based on Indonesian formal grammar, its preparation technique, and internet surfer habit. The results show all of the list are developed by term frequency analysis based on a text corpus. This paper also provides several recommendations for researcher both in text mining and text-based information retrieval field, developing stoplist by the word type and internet surfer habit.

Keywords: *stopword list, Indonesian, information retrieval, text mining, evaluation*

1. PENDAHULUAN

Di dalam dokumen teks terdapat banyak jenis kata seperti kata depan, kata sambung, kata ganti, kata sifat dan lain sebagainya. Sebagian kata tersebut tidak berpotensi dijadikan indeks dokumen karena kemunculannya tidak unik untuk sebuah dokumen tertentu. Untuk itu dilakukan proses penyaringan kata-kata itu (G Salton, Wong, & Yang, 1975)(Gerard Salton & Buckley, 1988)(Baeza-Yates & Ribeiro-Neto, 2008)(Manning, Raghavan, & Schütze, 2008). Langkah ini adalah pembersihan teks dari kata-kata yang tidak relevan dijadikan indeks, disebut sebagai langkah *stopword filtering*.

Daftar kata tersebut diberi istilah daftar *stopword* atau *stoplist* (Luhn, 1959)(Flood, 1999). Daftar kata ini bersifat unik, tiap-tiap bahasa memiliki daftar katanya tersendiri. Di dalam bahasa Indonesia terdapat beberapa versi daftar kata ini. Keragaman versi daftar kata yang dimaksud menjadi masalah tersendiri bagi peneliti di bidang sistem temu kembali informasi teks. Daftar kata mana yang lebih tepat digunakan sebagai *stoplist*. Idealnya, kata-kata yang sering digunakan di dalam perambahan internet tidak ikut masuk ke dalam daftar tersebut, dengan kata lain masih tercantum di dalam indeks dalam proses *stopword filtering*. Bila kata-kata tersebut dihapus dari indeks, maka dokumen yang mengandung kata terkait tidak lagi bisa ditemukan kembali oleh pencari.

Untuk itu di dalam penelitian ini dilakukan telaah daftar *stopword* bahasa Indonesia yang telah tersedia. Daftar yang dievaluasi di dalam penelitian ini adalah daftar yang disusun Fadillah Z. Tala (Z Tala, 2003), Damian Doyle (Doyle, n.d.), dan Wibisono (Wibisono, 2008). Validasi dilakukan

dengan telaah pendalaman struktur *stoplist* yang telah tersedia termasuk proses penyusunannya. Kemudian daftar *stopword* yang ada diperbandingkan dengan beberapa sumber yang lain, yaitu: hasil kuesioner, kata kunci pencarian dari Google Trends, dan istilah-istilah baku di dalam bahasa Indonesia.

Hasil penelitian ini berupa komentar proses penyusunan *stopword* yang telah ada. Dengan demikian diharapkan dapat dipilih daftar kata mana yang paling tepat digunakan oleh peneliti dan pengembang aplikasi di bidang ini. Penelitian ini juga memberikan rekomendasi daftar *stopword* yang sesuai dengan kebutuhan pengembang aplikasi ataupun peneliti di bidang ini berdasarkan kebutuhannya masing-masing.

2. STOPWORD FILTERING

Salah satu langkah pemrosesan teks di bidang sistem temu kembali informasi teks atau *text mining* adalah pembersihan teks dari kata-kata yang tidak relevan dijadikan indeks. Di dalam sebuah dokumen teks bisa jadi terdapat banyak jenis kata seperti kata depan, kata sambung, kata ganti, kata sifat, dan lain sebagainya. Sebagian kata tersebut bisa jadi tidak berpotensi dijadikan indeks dokumen karena kemunculannya tidak unik atau tidak pernah digunakan di dalam *query* pencarian. Untuk itu dilakukan proses penyaringan kata-kata tersebut (Luhn, 1959)(Flood, 1999). Penyaringan dilakukan dengan menyediakan sebuah daftar kata-kata yang tidak penting diindeks (*stopword list*). Hukum Zipf terkadang digunakan sebagai landasan pembentukan *stoplist*, utamanya pada analisis kemunculan kata (Zipf, 1949). Daftar *stopword* yang tersedia untuk Bahasa Indonesia diterangkan secara berturut-turut di Bagian 2.1, 2.2, dan 2.3.

2.1. Daftar *Stopword* Fadillah Z. Tala

Stoplist yang disusun Fadillah Z. Tala disusun bersamaan dengan penyusunan tesis master yang bersangkutan (Z Tala, 2003). Di appendix D, terdapat dua daftar yang diberikan: daftar yang disarankan (Tabel D.1.), dan daftar kata yang umum muncul di dalam *korpus* teks yang diteliti (Tabel D.2.). Daftar tersebut diturunkan dari analisis kemunculan kata yang dilakukan dengan menjalankan eksperimen *korpus* bahasa Indonesia. Eksperimen ini menggunakan koran daring Indonesia sebagai sumber teks. Satu tahun edisi dikumpulkan dari Kompas daring, http://www.kompas.com, sebagai salah satu koran yang banyak dibaca di Indonesia. Edisi ini diambil berurutan tiap hari selama setahun, dimulai dari Januari 2001 hingga Desember 2001 dengan total 3160 dokumen. Dokumen-dokumen tersebut hanyalah berita utama harian koran. *Korpus* untuk analisis ini berisi 50.000 kota-kata yang unik, setelah membuang nama-nama orang, kota, organisasi, negara, dll. Hasilnya adalah sebuah daftar *stopword*. Buku tesis Tala dapat diakses secara terbuka melalui tautan sebagai berikut: http://www.illc.uva.nl/Research/Reports/MoL-2003-02.text.pdf.

2.2. Daftar *Stopword* Damian Doyle

Daftar *Stopword* yang disediakan oleh Damian Doyle (Doyle, n.d.) dapat diakses terbuka di https://www.ranks.nl/stopwords. Selain *stoplist* Bahasa Indonesia, Doyle juga menyediakan *stoplist* untuk bahasa-bahasa lainnya. Ranks NL sendiri adalah sebuah perangkat pengembangan laman web yang ramah terhadap mesin perambah. Tidak terdapat informasi rinci di dalam website tersebut tentang bagaimana daftar stopword ini dibentuk. Begitu pula tidak ada naskah akademik yang memberikan informasi tahapan yang mendasarinya.

2.3. Daftar *Stopword* Yudi Wibisono

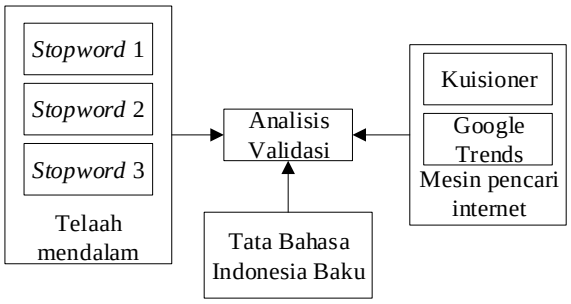
Yudi Wibisono membuat daftar *stopword* bahasa Indonesia untuk tugas salah satu matakuliah (Wibisono, 2008). Tujuannya waktu itu bukan untuk sistem temu kembali informasi, tetapi untuk klasifikasi *text mining*. Ia menggunakan langkah *stopword filtering* untuk mengurangi jumlah kata yang harus diproses algoritma klasifikasi.

Langkah yang dilakukan untuk membuat daftar *stopword* dengan cara mengumpulkan kata paling banyak muncul pada korpus (c*orpus).* Ia menggunakan beberapa ratus artikel berita Kompas sebagai sumber data korpus. Setelah diurutkan kemudian diperiksa secara manual satu persatu. Karena, daftar itu dibuat secara manual dan untuk keperluan klasifikasi, ada beberapa kata yang mungkin dapat diperdebatkan apakah layak masuk ke dalam daftar *stopword* atau bukan, misalnya "utara", "senin", "gedung" dan sebagainya. Jadi, ia mempersilahkan daftar itu untuk diedit sesuai kebutuhan. Daftar tersebut sudah diurutkan dari kata yang frekuensinya paling tinggi. *Stoplist* yang disusun Yudi ini dapat diakses secara di: https://yudiwbs.wordpress.com/2008/07/23/stop-words-untuk-bahasa-indonesia/.

3. METODE EVALUASI

Metode penelitian yang digunakan di dalam makalah ini ditunjukkan pada Gambar 1. Beberapa sumber data disandingkan satu dengan lainnya untuk dapat dianalisis lebih lanjut. Data pertama adalah tiga daftar *stopword* sebagaimana telah dibahas sebelumnya. Ketiga daftar *stopword* tersebut diperbandingkan dengan kebiasaan pengguna mesin pencari di internet. Penelitian ini akan menggunakan dua pendekatan untuk mengetahui kebiasaan pengguna mesin pencari. Pendekatan pertama menggunakan kuesioner. Pendekatan kedua melakukan telaah tren kata pencarian di dalam mesin pencari komersial Google. Langkah ini dilakukan untuk melihat, apakah daftar *stopword* yang ada

melibatkan jenis kata yang sebenarnya digunakan oleh pengguna atau tidak. Bila daftar *stopword* melibatkan jenis kata tersebut, akan terjadi kesenjangan antara indeks yang ada dengan keinginan pencari. Dengan demikian pengaruh jenis kata tersebut dinafikan ketika dilakukan penelusuran dengan mesin pencari.



Gambar 1. Diagram penelitian

Pelengkap dalam langkah validasi yang dilakukan adalah tinjauan terhadap tata bahasa baku Bahasa Indonesia yang disusun di dalam aturan ejaan bahasa Indonesia (Alwi, Dardjowidjojo, Lapoliwa, & Moeliono, 2010)(Hamizan, 2015). Istilah-istilah yang digunakan di dalam pedoman ini dijadikan timbangan atas istilah yang digunakan di dalam kuesioner, analisis tren Google, dan istilah di *stoplist* yang ada. Timbangan yang digunakan adalah jenis kata di dalam bahasa Indonesia: kata benda, kata bilangan, kata depan, kata ganti, kata keadaan, kata kerja, kata keterangan, kata sandang, kata sambung, kata seru, kata sifat (adjektiva), partikel, dan pronomina (Bahasa, 2008).

3.1. Telaah *Stoplist*

Telaah mendalam *stoplist* yang dilakukan di dalam penelitian ini meliputi:

- Menghitung jumlah kemunculan kata berdasar jenis katanya. Langkah ini dilakukan untuk mengetahui sebaran kata berdasar jenisnya di *stoplist*.
- Membandingkan satu *stoplist* dengan *stoplist* lainnya dari segi daftar katanya. Langkah ini untuk mengetahui di mana perbedaan jumlah dan kata yang ada.
- Melacak cara penyusunan *stoplist*.
- Mengurutkan tipe kata berdasar kemunculannya di *stoplist*. Langkah ini untuk mengetahui persentase kemunculan kata berdasar jenis kata.
- Menghitung persentase kata yang muncul di *stoplist* dengan kata setara di Kamus Tesaurus Bahasa Indonesia untuk tiap jenis katanya.

3.2. Validasi dengan Kuesioner

Tujuan langkah ini untuk mengetahui kebiasaan pengguna ketika merambah internet dengan mesin pencari. Kuesioner dibagikan ke responden mahasiswa D4 Teknik Informatika tingkat 4 kelas 4A, 4B, 4C, 4D, dan 4E Jurusan Teknologi Informasi, Politeknik Negeri Malang. Setelah kuesioner terkumpul, langkah- langkah selanjutnya adalah sebagai berikut:

- Input data dari kertas kuesioner ke dalam komputer menggunakan aplikasi Excel. Untuk pertanyaan tertutup, skor untuk setiap jawaban dari pertanyaan.
- Hasil pengolahan data kuesioner ditampilkan dalam bentuk deskriptif.
- Langkah selanjutnya adalah pengujian hipotesis berdasarkan data kuesioner yang ada.

3.3. Validasi dengan Tren Pencarian Google

Tujuan langkah ini sama dengan tujuan bagian sebelumnya, untuk mengetahui kebiasaan pengguna ketika merambah di internet dengan mesin pencari. Langkah yang dilakukan di dalam tahap ini meliputi:

- Dilakukan pencarian tren kata pencarian di Google dalam waktu satu bulan.
- Mengategorikan kata pencarian yang tren tersebut dengan jenis kata.
- Menghitung persentase penggunaan jenis kata dalam lingkup watu tertentu.

3.4. Hipotesis

Hipotesis penelitian ini adalah:

- *Stoplist* yang ada dibangun dengan memilih jenis-jenis kata tertentu yang dianggap tidak penting.
- *Stoplist* mengandung kata-kata yang tidak pernah digunakan oleh perambah internet.
- *Stoplist* yang dibangun Fadillah Tala adalah *stoplist* yang paling bisa dipertanggung jawabkan secara ilmiah.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Telaah *Stoplist*

Tabel 1. Padanan istilah

buku teks	**kamus tesaurus**	***stopword***
kata benda	nomina, tertulis n	*noun*
kata kerja	verba, tertulis v	*verb*
kata sambung	partikel, tertulis p	*particle*
kata depan	partikel, tertulis p	*particle*
kata keadaan	adjectiva, tertulis a	*adjective*
kata keterangan	adverbia, tertulis adv	*adverb*
kata bilangan	numeralia, tertulis num	*numeralia*
kata ganti	pronomina, tertulis pron	*pronomina*
kata sandang	partikel, tertulis p	-
kata seru	partikel, tertulis p	*particle*

Telaah dilakukan setelah mengetahui padanan penggunaan istilah antara buku teks tata bahasa baku bahasa Indonesia, buku Kamus Tesaurus Bahasa Indonesia, dan istilah yang digunakan di dalam *stoplist* yang ada. Padanan tersebut ditunjukkan di Tabel 1. Tampak dari tabel tersebut, baik kata sambung, kata depan, kata sandang, dan kata seru disebut sebagai partikel di kamus tesaurus dan di

Tabel 2. Kemunculan kata berdasar jenisnya

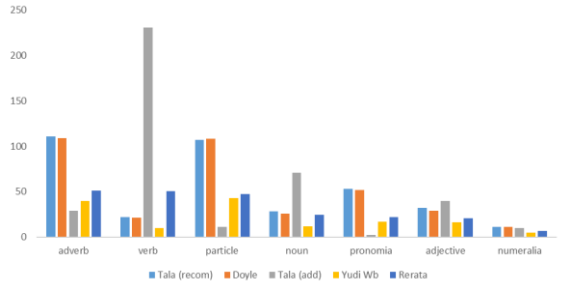
tipe	tala (recom)	%	doyle	%	tala (add)	%	yudi wb	%	rerata	%
adverbia	111	30,49	109,00	30,62	29,00	7,36	40,00	27,97	51,07	24,11
verba	22	6,04	21,00	5,90	231,00	58,63	10,00	6,99	50,65	19,39
partikel	107	29,40	108,00	30,34	11,00	2,79	43,00	30,07	47,36	23,15
nomina	28	7,69	26,00	7,30	71,00	18,02	12,00	8,39	24,29	10,35
pronomina	53	14,56	52,00	14,61	2,00	0,51	17,00	11,89	21,95	10,39
Adjektiva	32	8,79	29,00	8,15	40,00	10,15	16,00	11,19	20,58	9,57
numeralia	11	3,02	11,00	3,09	10,00	2,54	5,00	3,50	6,52	3,04
jumlah	**364**		**356**		**394**		**143**			

stoplist. Setelah dilacak mendetail, *stoplist* tidak melibatkan kata sandang, baik itu yang disusun Tala, Doyle, dan Wibisono.

4.1.1. Jumlah Kemunculan Kata Berdasar Jenis Katanya

Untuk mengetahui sebaran kata berdasar jenisnya di *stoplist*, dilakukan langkah ini. Hasil telaah disarikan di dalam Tabel 2. Secara umum, kata keterangan adalah jenis kata yang paling banyak terlibat di *stoplist*. Disusul kata kerja , partikel, kata benda, kata ganti, kata keadaan, dan kata bilangan. Perlu ditinjau lebih detail bahwa sebenarnya partikel itu terdiri atas kata depan, kata sambung, dan kata seru. Hasil tersebut banyak berubah akibat data *stoplist* Tala tambahan yang jauh berbeda polanya dibandingkan tiga *stoplist* lainnya.

Tiga *stoplist* lainnya memiliki pola urutan pelibatan jenis kata yang sama. Berturut-turut dari yang paling banyak ke sedikit: kata keterangan, partikel, kata ganti, kata sifat, kata benda, kata kerja, dan kata bilangan. Pola data ini ditampilkan lebih jelas secara visual di dalam Gambar 2. Gambar 2 menunjukkan dengan jelas, penggunaan kata kerja yang amat banyak di *stoplist* tambahan Tala besar pengaruhnya ke dalam nilai rerata yang ada.



Gambar 2. Pola kemunculan kata berdasar *stoplist*

4.1.2. Perbandingan Daftar Kata

Perbedaan kata yang ada tidak terlalu banyak antara *stoplist* Tala yang disarankan dan *stoplist* Doyle. Malah bisa dikatakan kedua *stoplist* ini hampir identik, perbedaannya sangat tipis. Beberapa kata yang dihapus *stoplist* Doyle dari Tala yang disarankan: berkali-kali, bermacam-macam, bersama-sama, masing-masing, sama-sama, sekali-kali, selama-lamanya, seolah-olah, dan setidak-tidaknya. Selain menghapus beberapa kata ulang tersebut, Doyle menambahkan satu kata yang tidak ada di *stoplist* Tala yang disarankan: selagi.

Stoplist Wibisono sama sekali berbeda dengan *stoplist* Tala yang disarankan. Meskipun, sama-sama menggunakan pendekatan analisis frekuensi kemunculan kata, Tala dan Wibisono menggunakan korpus yang berbeda.

4.1.3. Cara Penyusunan *Stoplist*

Tala (Z Tala, 2003) menyusun *stoplist*-nya dengan menggunakan analisis frekuensi kemunculan kata (Fox, 1992) yang diterapkan ke dalam Bahasa Indonesia. Hasilnya diperbandingakan dengan hasil *stoplist* di dalam bahasa lainnya. Frekuensi kemunculan kata dilakukan pada korpus Bahasa Indonesia. Tala menggunakan koran Indonesia daring sebagai sumber teksnya. Edisi satu tahun dikoleksi dari Kompas daring, http://www.kompas.com. Edisi ini diambil secara berurutan setiap harinya selama satu tahun (dimulai bulan January 2001 hingga December 2001) dengan total 3160 dokumen. Dokumen-dokumen tersebut adalah tajuk utama koran tersebut. Hasilnya berupa *korpus* yang memiliki 50.000 kata-kata yang unik, sesudah membuang nama orang, kota, organisasi, negara, dll. Dari hasil analisis ini, diperoleh *stoplist*, yang berasal dari kata-kata yang paling banyak muncul di dalam *korpus* tersebut (Z Tala, 2003).

Doyle (Doyle, n.d.) menyusun *stoplist*-nya dengan memodifikasi *stoplist* Tala. Ia menghilangkan 9 kata dari *stoplist* Tala yang berbentuk kata ulang, dan menambah satu kata sambung (partikel) di dalam *stoplist*-nya: selagi.

Wibisono (Wibisono, 2008) mengambil jalan yang sama dengan Tala, menggunakan analisis frekuensi kemunculan kata Bahasa Indonesia. Setelah diurutkan yang paling banyak muncul, diperiksa secara manual satu-persatu. Perbedaannya, Wibisono haya menggunakan beberapa ratus berita Kompas. Dari proses yang mirip, Wibisono hanya berisi 143 kata. Hanya sekitar 40% daftar Tala. Meskipun demikian, terdapat beberapa kata yang bisa diperdebatkan kelayakannya masuk *stoplist*: utara, senin, gedung, dlsb. Tala memproses penyusunan *stoplist* untuk

Tabel 3. Persentase kemunculan jenis kata di *stoplist*

tipe	total di kamus tesaurus	tala (suggested)	%	doyle	%	tala (add)	%	yudi wb	%
adverbia	796	111	13,94	109	13,69	29	3,64	40	5,03
verba	15.465	22	0,14	21	0,14	231	1,49	10	0,06
partikel	388	107	27,58	108	27,84	11	2,84	43	11,08
nomina	18.702	28	0,15	26	0,14	71	0,38	12	0,06
pronomina	106	53	50,00	52	49,06	2	1,89	17	16,04
adjektiva	8.018	32	0,40	29	0,36	40	0,50	16	0,20
numeralia	150	11	7,33	11	7,33	10	6,67	5	3,33
jumlah	**43.625**	**364**		**356**		**394**		**143**	
rerata			**14,22**		**14,08**		**2,49**		**5,12**

meraih gelar master, sedangkan Wibisono sebagai tugas kuliahnya.

4.1.4. Perbandingan Kata *Stoplist* dengan Kamus Tesaurus Bahasa Indonesia

Tabel 3 menunjukkan persentase kemunculan jenis kata di *stoplist*. Kata-kata yang dijadikan acuan adalah kata-kata yang muncul di Kamus Tesaurus Bahasa Indonesia. Dari tabel tersebut tampak bahwa *stoplist* yang ada tidak melibatkan banyak kata yang tersedia di kamus. Hal tersebut sangat dimungkinkan karena sumber data yang digunakan dalam penyusunannya berupa artikel berita daring. Besar kemungkinan kosakata yang digunakan di dalamnya adalah kosakata bahasa sehari-hari yang tidak banyak menggunakan kosakata di dalam kamus. Rata-rata tidak lebih 15% kata-kata di dalam kamus untuk jenis yang ada yang digunakan. Jumlah terbanyak pelibatan kosakata di dalam *stoplist* ada pada daftar *stoplist* Tala yang melibatkan 50% jenis kata ganti.

4.2. Hasil Kuesioner

Tabel 4. Hasil Kuesioner

jenis kata	jumlah	% penggunaan
kata benda	120	89,55
kata kerja	104	77,61
kata sambung	96	71,64
kata depan	86	64,18
kata keadaan	74	55,22
kata keterangan	72	53,73
kata bilangan	64	47,76
kata ganti	40	29,85
kata sandang	27	20,15
kata seru	18	13,43

Tujuan langkah ini untuk mengetahui kebiasaan pengguna ketika merambah internet dengan mesin pencari. Kuesioner dibagikan ke responden mahasiswa tingkat 4 kelas 4A, 4B, 4C, 4D, dan 4E Jurusan Teknologi Informasi, Politeknik Negeri Malang. Total responden 134 orang. Hasil kuesioner ditampilkan pada Tabel 4. Tabel tersebut menunjukkan pengalaman natural perambah internet ketika menggunakan Bahas Indonesia. Setelah diurutkan dari frekuensi penggunaan kata yang paling besar ke kecil, didapatkan hasil urutan berturut-turut: kata benda, kata kerja, kata sambung, kata depan, kata keterangan, kata bilangan, kata ganti, kata sandang, dan kata seru.

Nilai kemunculan minimum 18, median 73, dan maksimal 120. Tabel 4.4 menunjukkan kata seru adalah kata yang paling jarang digunakan dalam pencarian ketika merambah di internet. Dengan logika yang berkebalikan, jenis kata ini paling potensial masuk ke dalam *stoplist*. Demikian selanjutnya berturut-turut.

Hasil yang agak mengejutkan adalah seringnya responden menggunakan kata sambung ketika merambah internet. Pendalaman kuesioner dengan teknik dialog singkat menjelaskan responden memerlukan kata sambung ketika membandingkan sesuatu atau menyandingkan, sebagaimana fungsi kata sambung. Makna ini akan hilang bila jenis kata ini tidak ikut diindeks.

Kata sandang yang tidak masuk ke dalam *stoplist*, perlu dipertimbangkan masuk. Berdasarkan data responden, kata sandang ini jarang mereka gunakan di mesin perambah internet.

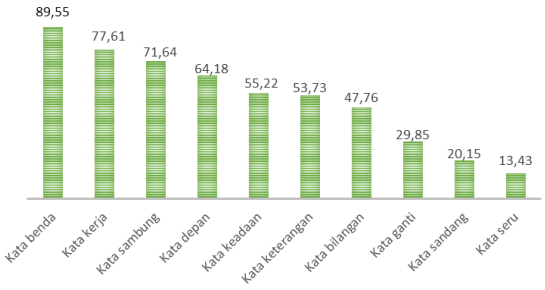
Menariknya, kata depan yang banyak dilibatkan di dalam *stoplist* ternyata sering digunakan perambah internet. Pendalaman kuesioner mengungkapkan bahwa responden menggunakannya ketika mencari hotel atau makanan, semacam kata pencarian "makanan enak di sawojajar". Terdapat perbedaan makna yang besar antara kata pencarian tersebut dengan "makanan enak sawojajar". Atau perambah mencari cara "bagaimana cara ke Surabaya". Sangat jauh berbeda dengan makna "bagaimana cara Surabaya".

Ringkasan hasil kuesioner di dalam Tabel 4 ditampilkan secara visual di dalam Gambar 3. Dari data ini didapatkan ide baru, penyusunan *stoplist* yang bisa diatur berdasarkan kebutuhan unjuk kerja sistem yang dibangun. Persentase kemunculan kata bisa diterapkan sebagai nilai ambang batas, kemudian daftar katanya diambil dari kamus. Hipotesisnya, semakin banyak jenis kata yang dilibatkan, semakin sederhana indeks sistem yang dibangun, semakin sedikit media penyimpan yang diperlukan, dan semakin cepat unjuk kerjanya. Sebaliknya ia semakin tidak sensitif dan kehilangan banyak makna penting. Pengguna dapat mencari

Tabel 5. Kemunculan Jenis Kata di Google Trends

jenis kata	topik		bulan		konsistensi
	ada/ tidak	contoh	ada/ tidak	contoh	
kata benda	ada	salon	ada	kukang	√
kata kerja	ada	penyergapan	ada	sarapan	√
kata sambung	ada	yang	ada	dan	√
kata depan	tidak		tidak		√
kata keadaan	ada	gerhana	ada	dingin	√
kata keterangan	ada	dekat	tidak		x
kata bilangan	ada	2016	ada	7	√
kata ganti	tidak		tidak		√
kata sandang	tidak		tidak		√
kata seru	tidak		tidak		√

sesuatu konten yang ternyata tidak tersedia di indeks karena kata yang mewakili konten tersebut telah dipangkas dengan sistem *stopword filtering*.



Gambar 3. Persentase penggunaan jenis kata

4.3. Hasil Tren Pencarian Google

Hasil penelusuran kata pencarian populer di Google Trends dapat dilihat berdasarkan topik maupun berdasarkan Bulan. Penelitian ini menggunakan kata pencarian populer di tahun 2016. Dari daftar kata tersebut, penelitian ini membentuk matriks yang menyandingkan jenis kata di dalam Bahasa Indonesia dengan kata-kata yang muncul secara populer di dalam Google Trends. Hasilnya ditampilkan di Tabel 5. Tampak dari tabel tersebut, beberapa jenis kata secara konsisten tidak pernah muncul di dalam pencarian populer Google: kata depan, kata ganti, kata sandang, dan kata seru. Kata sambung, bagian dari kata partikel muncul dengan frekuensi yang kecil. Hanya satu jenis kata yang tidak tentu munculnya, yaitu kata keterangan. Kata ini di daftar kata populer per topik muncul, tetapi per bulan tidak muncul. Kata-kata yang secara konsisten tidak muncul sangat layak untuk dilibatkan di dalam daftar *stoplist*.

4.4. Pembahasan Hipotesis

Di Bagian 3 disampaikan beberapa hipotesis penelitian. Fakta-fakta yang ditemukan berdasarkan penelitian dapat dipaparkan sebagai berikut:

- Hipotesis: *Stoplist* yang ada dibangun dengan memilih jenis-jenis kata tertentu yang dianggap tidak penting.
  Fakta: *Stoplist* dibangun dengan analisis frekuensi kemunculan kata. Hampir seluruh jenis kata masuk ke dalam *stoplist*, tetapi dengan persentase kecil. Hanya satu jenis kata yang tidak terlibat: kata sandang.
- Hipotesis: *Stoplist* mengandung kata-kata yang tidak pernah digunakan oleh perambah internet.
  Fakta: *Stoplist* mengandung kata-kata yang ternyata juga digunakan perambah internet: kata sambung dan kata depan. Bahkan *stoplist* juga mengandung kata benda dan kata kerja yang banyak digunakan.
- Hipotesis: *Stoplist* yang dibangun Fadilah Tala adalah *stoplist* yang paling bisa dipertanggung jawabkan secara ilmiah.
  Fakta: Proses pengembangan *stoplist* Tala dilakukan saat yang bersangkutan menyelesaikan studi masternya. Buku tesis yang bersangkutan dapat diunduh dengan mudah oleh orang lain, sehingga dapat dirujuk pijakan ilmiahnya.

5. KESIMPULAN

Kesimpulan dari penelitian ini bisa disarikan sebagai berikut. *Stoplist* yang ada: Tala, Doyle, dan Wibisono dibangun dengan analisis frekuensi kemunculan kata. Hampir seluruh jenis kata masuk ke dalam *stoplist* tersebut, tetapi dengan persentase kecil dengan rata-rata di bawah 15% dari kata yang ada di dalam kamus. Kata sandang tidak digunakan di dalam *stoplist*, padahal ia hanya digunakan 20,15% responden penelitian ini. Kata tersebut juga tidak muncul di dalam kata populer pencarian Google Trends. *Stoplist* banyak mengandung kata-kata yang juga digunakan perambah internet: kata sambung (71,6% responden) dan kata depan (64,18% responden). Dengan persentase yang cukup besar bergabung di dalam istilah partikel sebanyak rata-rata 23,15%. Bahkan *stoplist* juga mengandung sedikit kata benda (rata-rata 10,35%) dan kata kerja (19,39%) yang banyak digunakan reponden, berturut-turut: 89,55% dan 77,61% responden. Proses pengembangan *stoplist* Tala dilakukan

dengan pijakan ilmiah yang dapat dipertanggung-jawabkan dan mudah diakses.

Sedangkan saran lanjutan penelitian ini adalah menggagas sebuah mekanisme *stoplist* dinamik yang dapat menghasilkan *stoplist* yang tepat sesuai unjuk kerja sistem yang diperlukan. Data hasil kuesioner dan data Kamus Tesaurus Bahasa Indonesia dapat digunakan untuk tujuan ini.

6. DAFTAR PUSTAKA

ALWI, H., DARDJOWIDJOJO, S., LAPOLIWA, H., & MOELIONO, A. M., 2010. *Tata Bahasa Baku Bahasa Indonesia* (3rd ed.). Jakarta: Pusat Bahasa dan Balai Pustaka.

BAEZA-YATES, R., & RIBEIRO-NETO, B., 2008. *Modern Information Retrieval: The Concepts and Technology Behind Search* (2nd ed.). USA: Addison-Wesley Publishing Company.

BAHASA, P., 2008. *Kamus Tesaurus Bahasa Indonesia*. Departemen Pendidikan Nasional.

DOYLE, D., tanpa tahun. Indonesian Stopword. https://www.ranks.nl/stopwords/indonesian

FLOOD, B. J., 1999. Historical Note: The Start of a Stop List at Biological Abstracts. *JASIS*, *50*(12), 1066.

FOX, C., 1992. Information Retrieval. In W. B. Frakes & R. Baeza-Yates (Eds.) (pp. 102–130). Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

HAMIZAN, Y., 2015. *Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan dan Intisari Kata Bahasa Indonesia* (1st ed.). Seruni Multi Aksara.

LUHN, H. P., 1959. Key word-in-context index for technical literature (kwic index). *American Documentation*, *11*(4), 288–295.

MANNING, C. D., RAGHAVAN, P., & SCHÜTZE, H., 2008. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

SALTON, G., & BUCKLEY, C., 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, *24*(5), 513–523.

SALTON, G., WONG, A., & YANG, C. S., 1975. A Vector Space Model for Automatic Indexing. *Commun. ACM*, *18*(11), 613–620.

WIBISONO, Y., 2008. Indonesian Stopword. https://yudiwbs.wordpress.com/2008/07/23/stop-words-untuk-bahasa-indonesia/

Z TALA, F., 2003. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.

ZIPF, H., 1949. *Human Behaviours and the Principle of Least Effort*. Cambridge, MA: Addison- Wesley.

Halaman ini sengaja dikosongkan