Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses *Text Mining*

Aris Tri Jaka H.
Program Studi Informatika, Fakultas TEKNIK, Universitas PGRI Semarang
Gedung B Lantai 3, Kampus 1 Jl. Sidodadi Timur 24, Semarang
E-mail : aristrijaka@upgris.ac.id

Abstract*—The growing world of information technology course, the growing impact of data outstanding and continues to grow significantly, and initial data processing or preprocessing text in text mining process is expected to reduce by eliminating the word - the word or text that are not necessary or do not have the meaning of text database or document. By decreasing the amount of text was expected to ease further processing in order to mine the information contained within the document - document or text - text in a miraculous process by applying existing methods to produce useful information from the text without reducing the sense or meaning and information contained in the document.*

Keyword : *data, text mining, information, preprocessing*

***Abstrak*—**Berkembangnya dunia teknologi informasi tentu saja membawa dampak semakin besarnya data yang beredar dan terus bertambah besar secara signifikan, dan pengolahan data awal atau preprocessing text dalam proses *text mining* di harapkan dapat mengurangi dengan menghilangkan kata – kata atau teks yang tidak perlu atau tidak mempunyai arti dari database teks atau dokumen. Dengan berkurangnya jumlah teks diharapkan dapat meringankan proses selanjutnya dalam rangka menambang informasi yang berada dalam dokumen – dokumen ataupun teks- teks yang di proses dengan menerapkan bebrapa metode yang ada untuk dapat menghasilkan informasi yang berguna dari teks tersebut tanpa mengurangi arti ataupun makna serta informasi yang dikandung dalam dokumen tersebut.

***Kata Kunci* :** data, text mining, informasi, preprocessing

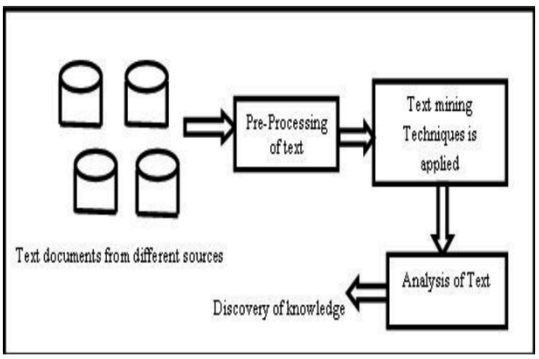
PENDAHULUAN

Dengan perkembangan teknologi yang semakin besar maka kebutuhan akan penyajian informasi yang cepat dan akurat menjadi salah satu focus utama dalam penelitaian dan pengembangan guna memenuhi kebutuhan informasi yang semakin cepat dan akurat. Data Mining merupakan kompleks teknologi yang berakar pada berbagai disiplin ilmu: matematika, statistik, ilmu komputer, fisika, teknik, biologi, dll, dan dengan beragam aplikasi dalam berbagai macam domain yang berbeda: bisnis, kesehatan, sains dan teknik , dll Pada dasarnya, data mining dapat dilihat sebagai ilmu menjelajahi dataset besar untuk mengekstraksi informasi tersirat, yang sebelumnya tidak diketahui dan berpotensi berguna [1].

Sedangkan *Text mining* adalah salah satu penambangan informasi yang berguna dari data – data yang berupa tulisan,

dokumen atau text dalam bentuk klasifikasi maupun clustering. Text mining masih merupakan bagian dari data mining dimana akan memproses data – data atau text – text serta dokumen – dokumen yang bisa jadi dalam jumlah sangat besar. Untuk memproses data yang sangat besar tentulah akan memakan sumber daya yang tidak sedikit kaitanya dengan pengolahan data tersebut. Disinilah diperukanya sebuah pemrosesan awal atau preprocessing data text tersebut sebelum data tersebut di lakukan proses text mining sesuai algoritma yang akan diterapkan.

Dengan *text mining* maka kita akan melakukan proses mencari atau penggalian informasi yang berguna dari data tekstual[2]. Ini juga merupakan salah satu kajian penelitian yang sangat menarik dan juga sangat berguna di kemudian hari dimana seperti mencoba untuk menemukan pengetahuan dari dokumen–dokumen atau teks - teks yang tidak terstruktur. *Text mining* sekarang juga memiliki peran yang semakin penting dalam negara berkembangaplikasi, seperti mengetahui isi dari teks secara langsung dari proses *text mining* tanpa perlu membaca satu persatu teks atau tulisan yang ada. Proses Text mining adalah sama dengan data mining, kecuali, beberapa metode dan data yang di kelola nya seperti data teks yang tidak terstruktur, terstruktur sebagian maupun terstruktur seperti teks email, teks HTML, maupun teks komentar serta dari berbagai sumber[3].

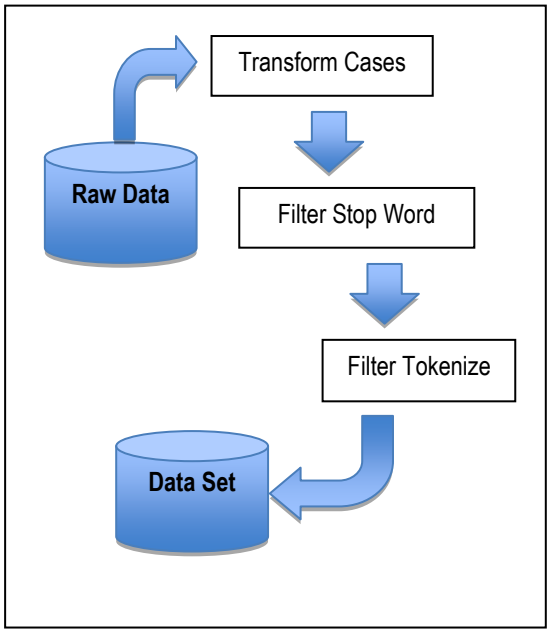


Gambar. 1. Proses Teks Mining

Untuk dapat melakukan penambangan informasi atau text mining maka perlu dilakukan beberapa tahapan yang harus dilakukan untuk mengolah sumber data baik yang terstruktur, terstuktur sebagian dan yang tidak terstruktur dari beberapa sumber maka data-data tersebut perlu dilakukan proses awal atau di sebut sebagai preprocessing text yang bermaksud mengolah data awal yang masih bermacam – macam untuk dijadikan sebuah data teratur yang dapat dikenai atau diterapkan beberapa metode text mining yang ada.

PREPROCESSING TEXT

Dalam penelitain ini di terapkan text preprocessing untuk data yang akan di gunakan dalam proses analisa sentimen, dimana data yang kita proses akan kita ambil informasi yang terkandung didalmnya dalam hal sentimen penulisnya yaitu negaitf atu positif. Guna memudahkan dalam mengelola data maka data perlu kita berikan analisa sentimen secara manual dengan membaca maksud dari kalimat yang ada dalam sentimen tersebut, sehingga dapat diberikan penilaian bahwa sentimen tersebut merupakan setimen negatif atau positif.



Gambar 2. Alur preprocessing text

Transform Cases

Dengan fitur *transform cases* kita dapat secara otomatis mengubah semua huruf pada teks menjadi huruf kecil semua atau menjadi huruf kapital semua, pada penelitian ini semua huruf dirubah kedalam huruf kecil karena mayoritas teks berupa tulisan opini yang sebagian besar merupakan huruf kecil semua[4].

Filter Stop Word

Dengan fitur ini maka teks sebelum di klasifikasikan di hilangkan dulu teks yang tidak berhubungan dengan analisa sentimen sehingga dimensi teks akan berkurang tanpa mengurangi isi sentimen dari teks tersebut[5].

Fiter stopword bahasa indonesia ini penulis ambil dari internet yang dibuat oleh Wang Pidong seorang Ph.D dari National University Singapore dengan penulis menambahkan beberapa kata yang memiliki arti sama dengan kata – kata yang sudah ada dalam daftar stopword tersebut.

Filter Tokenize

Tahapan ini juga menghilangkan karakter-karakter tertentu seperti tanda baca serta memfilter berdasarkan panjang teks[6]. Untuk metode pengujian atau evaluasi dilakukan pengujian terhadap model-model yang diteliti untuk mendapatkan informasi model diusulkan.

EXPERIMEN DAN PENGUJIAN

Tools yang digunakan dalam experimen ini adalah Rapidminer [7]yang di update dengan penambahan *plugin text processing* yang telah memiliki fitur pemrosesan teks diataranya:

Transform Cases

Dengan fitur *transform cases* kita dapat secara otomatis mengubah semua huruf pada teks menjadi huruf kecil semua atau menjadi huruf kapital semua, pada penelitian ini semua huruf dirubah kedalam huruf kecil karena mayoritas teks berupa tulisan opini yang sebagian besar merupakan huruf kecil semua.

Filter Stop Word (Indonesia)

Dengan fitur ini maka teks sebelum di klasifikasikan di hilangkan dulu teks yang tidak berhubungan dengan analisa sentimen sehingga dimensi teks akan berkurang tanpa mengurangi isi sentimen dari teks tersebut.

Fiter stopword bahasa indonesia ini penulis ambil dari internet yang dibuat oleh Wang Pidong seorang Ph.D dari National University Singapore dengan penulis menambahkan beberapa kata yang memiliki arti sama dengan kata – kata yang sudah ada dalam daftar stopword tersebut.

Filter Tokenize

Tahapan ini juga menghilangkan karakter-karakter tertentu seperti tanda baca serta memfilter berdasarkan panjang teks.

Untuk metode pengujian atau

evaluasi dilakukan pengujian terhadap model-model yang diteliti untuk mendapatkan informasi model diusulkan. Evaluasi dan validasi menggunakan metode sebagai berikut:

Mengubah Teks Menjadi Matrix

Untuk dapat diolah atau di proses maka data awal yang berupa kalimat setelah dilakukan pemrosesan awal data akan menjadi suatu atribut berupa teks, dan utnuk dapat diterpakan kedalam algoritma selanjutnya maka dibutuhkan adanya transformasi data dari teks kedalam sebuah matrix yang berisi numerik.

Pada penelitian ini penulis menggunakan proses pembentukan vector kalimat dengan menggunakan TF-IDF (*term frequency-inverse document frequency*) Matrix yang dirumuskan sebagai berikut:

TF-IDF = TF*IDF = TF*log(n/df)
Dimana: tf = frekuensi teks
df = frekuensi dokument
n = jumlah dokumen
contoh perhitungan matriks TF-IDF
jika ada tabel atribut sebagai berikut :

Tabel 1Tabel Atribut contoh Perhitungan TF-IDF

	Dok 1	Dok 2	Dok 3	df
Aplikasi	6	0	7	2
Bagus	9	2	3	3
Buruk	10	4	0	2

Maka perhitungna matrix TF-IDF nya

Tabel 2. Tabel Perhitungan TF-IDF

	Dok 1	Dok 2	Dok 3
Aplikasi	6*log(3/2)	0	7*log(3/2)
Bagus	9*log(3/3)	2*log(3/3)	3*log(3/3)
Buruk	10*log(3/2)	4*log(3/2)	0

Dan hasil matrix TF-IDF nya adalah sebagai berikut:

Tabel 3.Tabel Matrix TF-IDF

	Dok 1	Dok 2	Dok 3
Aplikasi	1.06	0.00	1.23
Bagus	0.00	0.00	0.00
Buruk	1.76	0.70	0.00

HASIL DAN PEMBAHASAN

Hasil Eksperimen Transform Cases

Pada tahapan pemrosesan awal data dengan menggunakan *transform case* ini mengubah semua huruf kedalam huruf kecil semua, namun jika teks sudah dalam huruf kecil maka tidak di ubah. Dari dataset refiew aplikasi android berbahas indonesia ini ada beberapa teks atau huruf yang yang dirubah dari huruf besar kedalam huruf kecil seperti Tabel 4 berikut.

Tabel 4.Trasform cases dari huruf besar ke huruf kecil

Text Awal	Text Akhir
Ini aplikasi hlr lookup no tsb dikeluarkan didaerah mana. Tp kl misal no tsb dbawa ke luar daerah ttp ngga bs update alias ttp ngebaca daerah asal..totally useless! , yg komen jg sok tau ttg telekomunikasi modal ngenet, bocah jaman skr..	ini aplikasi hlr lookup no tsb dikeluarkan didaerah mana. tp kl misal no tsb dbawa ke luar daerah ttp ngga bs update alias ttp ngebaca daerah asal..totally useless! , yg komen jg sok tau ttg telekomunikasi modal ngenet, bocah jaman skr..

Proses *tranform cases* ini dilakukan pada seluruh data sentimen yang ada di folder dataset, baik sentimen positif maupun negatif. Jadi dihasilkan fitur atau kata – kata dalam format teks huruf kecil semua.

Hasil Eksperimen Filter Tokens

Pada tahapan ini menyeleksi fitur atau kata kata yang bukan merupakan kata, dalam hal ini peneliti mengambil menghilangkan semua tanda baca dan segala sesuatu yang bukan huruf jadi teks menjadi bersih dari tanda baca dan angka ataupun apapun yang bukan huruf. Juga dilakukan limitasi minimal huruf dan maksimal huruf yang terdapat dalam satu kata. Karena dalam sentimen berbahasa Indonesia ini peneliti memasukkan minimal satu huruf sudah dapat di anggap sebagai kata karena banyak review berbahasa Indonesia tidak menggunakan bahasa baku dan menggunakan bahasa alay atau bahasa gaul yang beberapa hanya terdiri satu huruf saja dalam tiap kata.

Tabel 5. Proses *Filter tokens*

Teks sebelum di tokenize	Teks setelah di tokenize
Buat pengguna multi operator sangat berguna. Tarif normal serta paket nelpon tiap operator tidaklah sama. Ada yang murah ke sesama operator saja. Ada yang murah walau beda operator, tetapi hanya nomor-nomor lokal saja. Ada yang sedikit lebih mahal, tetapi pukul rata untuk semua operator.! Dan lain-lain. Dengan mengetahui tempat asal nomor dikeluarkan, kita jadi bisa menentukan sebaiknya pakai nomor yang mana buat menelpon..:D	buat pengguna multi operator sangat berguna tarif normal serta paket nelpon tiap operator tidaklah sama ada yang murah ke sesama operator saja ada yang murah walau beda operator tetapi hanya nomor nomor lokal saja ada yang sedikit lebih mahal tetapi pukul rata untuk semua operator dan lain lain dengan mengetahui tempat asal nomor dikeluarkan kita jadi bisa menentukan sebaiknya pakai nomor yang mana buat menelpon d

Hasil Ekspeimen Filter Stopword

Pada tahapan ini filter stopword berfungsi untuk mengurangi atau menghilangkan beberapa kata yang tidak memiliki hubungan terhadap sentimen, yaitu kata kata yang tidak berpengaruh terhadap hasil sentimen pada review tersebut.dari dataset awal yang berjumlah 2.000 file yang terdiri dari 1.000 sentimen positif dan 1.000 sentimen negatif di hasilkan atribut atau kata sebanyak 228 atribut, setelah di kurangi dengan stopword makan fitur yang perlu diperhitungkan maka tinggal 114 atribut. Hasil dari beberapa kata yang dihilangkan pada dataset ini adalah seperti pada Tabel 6 berikut:

Tabel 6 Daftar kata yang dihilangkan dengan *filter stopword*

No.	Kata	Nama Atribut	Jumlah Muncul	Jumlah Dokumen	Positif	Negatif
1	ada	ada	255	227	110	145
2	akan	akan	26	26	19	7
3	anak	anak	31	28	31	0
4	and	and	23	20	18	5

5	ane	ane	54	42	8	46
6	apa	apa	40	37	11	29
7	atau	atau	40	37	11	29
8	awal	awal	20	20	1	19
9	baik	baik	62	60	29	33
10	banget	banget	91	90	52	39
11	banyak	banyak	64	62	28	36
12	baru	baru	49	48	15	34
13	belum	belum	21	21	4	17
14	berita	berita	92	70	33	59
15	bintang	bintang	48	43	16	32
16	bisa	bisa	410	350	153	257
17	boleh	boleh	23	22	20	3
18	bs	bs	53	46	13	40
19	buat	buat	134	121	88	46
20	close	close	48	47	2	46
21	cukup	cukup	22	22	11	11
22	cuma	cuma	56	55	4	52
23	dalam	dalam	27	27	24	3
24	dan	dan	272	239	165	107
25	dari	dari	67	65	30	37
26	dengan	dengan	63	59	45	18
27	detik	detik	41	33	15	26
28	dgn	dgn	26	25	10	16
29	di	di	423	331	101	322
30	dibuka	dibuka	31	28	3	28
31	dong	dong	51	51	23	28
32	dr	dr	20	20	8	12
33	dulu	dulu	52	51	14	38
34	for	for	36	33	27	9
35	g	g	29	26	5	24
36	ga	ga	263	209	39	224
37	gak	gak	170	142	32	138
38	gk	gk	36	26	7	29
39	gw	gw	36	30	4	32
40	hanya	hanya	49	46	10	39
41	harus	harus	57	54	16	41
42	i	i	54	46	30	24
43	iklan	iklan	57	48	16	41
44	in	in	24	22	6	18
45	indonesia	indonesia	29	28	19	10
46	ini	ini	261	233	146	115

47	it	it	55	47	28	27
48	itu	itu	42	38	12	30
49	jadi	jadi	94	88	33	61
50	jalan	jalan	38	36	4	34
51	jangan	jangan	23	21	3	20
52	jd	jd	47	40	17	30
53	jelas	jelas	29	28	7	22
54	jg	jg	36	31	13	23
55	juga	juga	65	63	32	33
56	kalau	kalau	52	50	33	19
57	kalo	kalo	81	75	34	47
58	kan	kan	21	20	9	12
59	karena	karena	27	27	14	13
60	kata	kata	26	20	16	10
61	ke	ke	130	113	42	88
62	kecil	kecil	20	20	6	14
63	keluar	keluar	31	30	8	23
64	kenapa	kenapa	22	21	2	20
65	kita	kita	49	38	30	19
66	klo	klo	40	38	17	23
67	kok	kok	59	56	9	50
68	kompas	kompas	32	32	13	19
69	kurang	kurang	23	21	2	21
70	lagi	lagi	106	103	55	51
71	lah	lah	21	21	8	13
72	lain	lain	38	37	17	21
73	lama	lama	81	74	7	74
74	langsung	langsung	35	33	9	26
75	lbh	lbh	23	22	13	10
76	lebih	lebih	122	108	47	75
77	lg	lg	56	52	25	31
78	lokasi	lokasi	27	23	7	20
79	luar	luar	20	20	15	5
80	makin	makin	39	31	21	18
81	malah	malah	54	50	1	53
82	mana	mana	29	27	5	24
83	masa	masa	31	29	1	30
84	masih	masih	66	63	12	54
85	mau	mau	66	63	17	49
86	membantu	membantu	102	101	100	2
87	mudah	mudah	44	39	35	9
88	muncul	muncul	44	42	8	36

89	n	n	34	32	23	11
90	nggak	nggak	23	21	4	19
91	ni	ni	29	27	9	20
92	no	no	32	22	4	28
93	nomor	nomor	40	27	19	21
94	not	not	32	31	2	30
95	nya	nya	210	175	93	117
96	orang	orang	30	27	21	9
97	pada	pada	24	22	10	14
98	padahal	padahal	35	35	3	32
99	paling	paling	25	25	16	9
100	perlu	perlu	37	36	20	17
101	saat	saat	37	34	24	13
102	saja	saja	38	35	8	30
103	sama	sama	58	54	16	42
104	sangat	sangat	241	218	203	38
105	satu	satu	30	28	14	16
106	saya	saya	167	140	98	69
107	sebelumnya	sebelumnya	28	27	2	26
108	sekali	sekali	62	60	38	24
109	sekarang	sekarang	31	31	8	23
110	selalu	selalu	42	40	19	23
111	semua	semua	38	37	20	18
112	seperti	seperti	27	26	14	13
113	sering	sering	50	48	11	39
114	setelah	setelah	60	60	8	52
115	setiap	setiap	28	26	15	13
116	sudah	sudah	47	44	23	24
117	tambah	tambah	24	24	19	5
118	tapi	tapi	100	97	21	79
119	tau	tau	25	24	10	15
120	tdk	tdk	39	34	7	32
121	terlalu	terlalu	23	23	3	20
122	terus	terus	79	79	44	35
123	tetep	tetep	21	21	3	18
124	the	the	47	36	27	20
125	this	this	38	34	23	15
126	tidak	tidak	102	94	30	72
127	to	to	45	37	22	23
128	tp	tp	56	54	14	42
129	trus	trus	33	31	12	21
130	udah	udah	74	69	21	53

131	udh	udh	20	20	3	17
132	untuk	untuk	121	112	75	46
133	up	up	31	30	17	14
134	utk	utk	63	48	43	20
135	versi	versi	73	65	18	55
136	very	very	26	24	19	7
137	waktu	waktu	45	42	33	12
138	ya	ya	92	89	35	57
139	yang	yang	196	161	115	81
140	yg	yg	355	287	184	171
141	you	you	28	28	24	4

KESIMPULAN DAN SARAN

Dari prepossessing text maka banyak sekali di hasilkan beberapa pengurangan atau ringkasan terhadap berbagai kata yang tidak diperlukan untuk proses text mining selanjutnya, dalam penelitaian ini adalah untuk proses sentiment analisis. Dalam pemrosesan penghilangan atau peminimalisiran kata ini di perluakn beberapa tahapan diataranya penyesuaian jenis huruf (transform cases) penghilanagn tanda baca (filter tokenized) serta penghilangan stop word dalam bahasa Indonesia, dengan adanya proses preprosesing teks ini maka data yang banyak dan tidak terpakai akan tereliminasi terlebih dahulu sebelum dataset dikenakan metode penelusuran sentiment analisis yang ada.

Saran untuk selanjutnya mungkin dapat di gunakan berbagai kombinasi penggurangan kata, maupun stopword dengan bahasa yang lain atau campuran, karena banyak kata kata bahasa asing atau bahasa gaul yang di gunakan. Sehingga jika semakin kompleks stopword yang digunakan diharapkan dapat menambah pengurangan kata yang tidak berarti tanpa mengurangi sentimen yang ada dalam kalimat atau kata tersebut.

DAFTAR PUSTAKA

[1] F. Gorunescu, *Data Mining*, vol. 12. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[2] J. Han and M. Kamber, *Data mining: concepts and techniques*. 2006.

[3] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining - An Overview," vol. 5, no. 1, pp. 7–16.

[4] R. a Baeza-Yates, "Text retrieval: Theory and practice," *Proc. 12th {IFIP} World Comput. Congr.*, vol. I, no. JANUARY 1998, pp. 465–476, 1992.

[5] V. Srividhya and R. Anitha, "Evaluating preprocessing techniques in text categorization," *Int. J. Comput. Sci. Appl.*, no. 2010, pp. 49–51, 2010.

[6] S. Krishna and S. Bhavani, "An efficient approach for text clustering based on frequent itemsets," *Eur. J. Sci. ...*, vol. 42, no. 3, pp. 385–396, 2010.

[7] S. Land and S. Fischer, "RapidMiner 5," *docs.rapid-i.com*.