PEMBENTUKAN DAFTAR KATA KUNCI
UNTUK PENGKLASIFIKASIAN OPINI PADA MEDIA SOSIAL
DENGAN PENDEKATAN KORPUS DAN KAMUS

Laporan Tugas Akhir I

Disusun sebagai syarat kelulusan mata kuliah
IF4091/Tugas Akhir I

Oleh
WILSON FONDA
NIM : 13510015

PROGRAM STUDI TEKNIK INFORMATIKA
SEKOLAH TEKNIK ELEKTRO & INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG
Januari 2014

PEMBENTUKAN DAFTAR KATA KUNCI
UNTUK PENGKLASIFIKASIAN OPINI PADA MEDIA SOSIAL
DENGAN PENDEKATAN KORPUS DAN KAMUS

Laporan Tugas Akhir I

Oleh

WILSON FONDA

NIM : 13510015

Program Studi Teknik Informatika

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Telah disetujui untuk dimajukan dalam Seminar Tugas Akhir I

di Bandung, pada tanggal 13 Februari 2014

Pembimbing,

Dr. Eng. Ayu Purwarianti, S.T., M.T.

NIP. 197701272008012011

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I.

PENDAHULUAN

I.1. Latar Belakang

Media sosial telah menjadi salah satu bagian dari kehidupan masyarakat saat ini. Masyarakat sering menyampaikan opininya terhadap seseorang ataupun sesuatu melalui media sosial seperti *Facebook*, *Twitter*, *Tumblr*, ataupun *Path*. Opini yang diberikan oleh masyarakat ini dapat berupa saran, solusi ataupun kritik terhadap pemerintah. Tapi kebanyakan dari opini ini tidak dapat ditangkap oleh pemerintah, sehingga pemerintah menjadi kurang tanggap terhadap masalah-masalah yang sedang terjadi di masyarakat. Masyarakat juga terkadang bingung untuk mencari tempat atau departemen yang cocok untuk melaporkan keluhannya. Apabila opini-opini ini dikumpulkan dan diberikan kepada departemen pemerintah yang berhubungan maka penanganan terhadap masalah yang terjadi di masyarakat dapat lebih cepat dan tepat sasaran.

Salah satu metode dalam penangkapan opini masyarakat dari media sosial ini adalah pembuatan daftar kata kunci untuk mengklasifikasikan opini yang didapat dari *Twitter* atau media sosial lainnya. Dengan adanya daftar kata kunci maka pengklasifikasian opini dapat dilakukan dengan menganalisa kata kunci yang muncul pada opini tersebut. Namun, untuk saat ini belum ada daftar kata kunci dalam Bahasa Indonesia yang dapat menangani pengklasifikasian opini tersebut. Sehingga pembuatan daftar kata kunci ini menjadi hal yang cukup penting untuk dapat meningkatkan kualitas dari hasil pengklasifikasian opini. Terdapat metode lain seperti N-gram atau *clustering* untuk melakukan klasifikasi secara langsung terhadap opini-opini yang ditangkap dari *Twitter*, namun metode pembuatan daftar kunci mempunyai beberapa keuntungan dalam implementasinya bila dibandingkan dengan kedua metode tersebut. Keuntungan dari pembuatan daftar kata kunci ini adalah pengklasifikasian opini yang lebih cepat dan terkendali. Pengklasifikasian yang lebih cepat dikarenakan mesin hanya perlu mencari setiap kata dalam pesan *Twitter* dan mencocokkannya dengan daftar kata kunci yang telah ada.

Pengklasifikasian yang lebih terkendali dikarenakan daftar kata kunci yang telah dihasilkan secara otomatis dapat dievaluasi untuk menghasilkan akurasi klasifikasi yang lebih baik.

Pembuatan daftar kata kunci secara manual memang memungkinkan, namun hal ini memerlukan sumber daya yang sangat besar. Oleh karena itu, banyak berkembang penelitian untuk membuat daftar kata kunci secara otomatis. Diharapkan bahwa dengan bertambahnya kualitas hasil, sumber daya yang dikeluarkan untuk melakukan evaluasi terhadap daftar kata kunci tersebut akan berkurang. Salah satu penelitian daftar kata kunci ini adalah *SentiWordNet* (Esuli & Sebastiani, 2006.) yang merupakan daftar nilai sentimen dari setiap kata pada *WordNet*. Pada *SentiWordNet*, tidak terdapat kategori klasifikasi setiap kata, melainkan terdapat nilai riil yang menunjukkan tingkat positif dan negatif dari kata tersebut. Hal ini berbeda dengan daftar kunci yang dibuat pada tugas akhir ini, karena untuk setiap kata akan diberi label kategori dan label yang diberikan bukan berupa nilai riil.

Dua pendekatan yang dapat dilakukan untuk pembuatan daftar kata kunci secara otomatis adalah dengan melakukan penelusuran kamus ataupun dengan melakukan penelusuran pada dokumen (pendekatan korpus). Pendekatan dengan penelusuran kamus merupakan metode untuk mencari kata-kata yang dapat dibentuk menjadi model klasifikasi berdasarkan kedekatan kategori terhadap kata di dalam kamus elektronik. Metode ini telah banyak dilakukan dalam beberapa kasus pemrosesan teks, salah satunya adalah untuk menentukan nilai kedekatan kata pada kamus *WordNet* (Mihalcea, Corley, & Strapparava, 2006). Dengan menggunakan perhitungan kedekatan kata pada penelitian yang telah ada (Mihalcea, Corley, & Strapparava, 2006), akan ditentukan kata-kata yang termasuk ke dalam model klasifikasi yang akan dibuat. Namun, pendekatan kamus memerlukan kata-kata awal untuk dicari kata-kata lainnya, sehingga pendekatan ini diimplementasikan pada keluaran dari pendekatan korpus.

Pendekatan dengan penelusuran dokumen (pendekatan korpus) merupakan pendekatan untuk mencari kata-kata yang berpengaruh pada beberapa dokumen untuk dibentuk menjadi daftar kata yang digunakan sebagai model klasifikasi

dokumen lain. Pendekatan ini dapat dilakukan dengan beberapa teknik, seperti metode *clustering / cluster analysis* dan *Association Rule Learning*. Berbagai macam algoritma telah dikembangkan untuk mendapatkan hasil *clustering* yang lebih baik, namun hasil dari algoritma ini relatif terhadap studi kasus yang dihadapi. Oleh karena itu perlu dilakukan analisa atau evaluasi untuk dapat menentukan algoritma yang memberikan hasil *clustering* terbaik. Metode *clustering* ini digunakan untuk mengumpulkan kata-kata yang berpengaruh pada kategori dokumen tertentu. *Association Rule Learning* (Antonie & Zai'ane, 2002) merupakan algoritma yang umumnya digunakan untuk mencari hubungan antar data pada basis data. Kedua metode ini dilakukan pada dokumen yang telah diberi label dan diterapkan pada dokumen-dokumen dengan kategori yang sama untuk mendapatkan kategori dari setiap kelompok ataupun aturan (Antonie & Zai'ane, 2002). Agar dokumen tersebut dapat diproses oleh metode pada pendekatan korpus, maka perlu dilakukan juga pemrosesan awal dan ekstraksi fitur.

Terdapat beberapa hal yang menjadi masalah untuk dapat melakukan penangkapan opini dalam bahasa Indonesia. Salah satunya adalah belum adanya kamus Bahasa Indonesia elektronik yang telah memetakan seluruh kata dalam Bahasa Indonesia secara lengkap beserta arti/ keterangan dari kata tersebut. Untuk saat ini, hanya terdapat *WordNet* Bahasa Indonesia (Noor, Sapuan, & Bond, 2011) yang berisi terjemahan kata dari kamus *WordNet* dan belum memiliki arti / keterangan katanya dalam Bahasa Indonesia. *WordNet* merupakan Kamus Bahasa Inggris secara elektronik yang telah banyak digunakan sebagai basis dari berbagai penelitian yang berhubungan pemrosesan teks dan bahasa alami. Dengan demikian, pendekatan-pendekatan yang digunakan sebaiknya tidak memerlukan kamus Bahasa Indonesia yang lengkap (memiliki kata dan keterangan dari kata tersebut).

I.2. Rumusan Masalah

Belum adanya penelitian yang melakukan pembangunan daftar kata kunci dalam Bahasa Indonesia untuk masalah pengklasifikasian opini menjadi latar belakang dari pelaksanaan tugas akhir ini. Sehingga masalah yang akan diselesaikan pada tugas akhir ini adalah :

1. Mengevaluasi pendekatan korpus dan pendekatan korpus yang dilengkapi dengan pendekatan kamus (pendekatan *hybrid*) yang digunakan untuk membuat daftar kata kunci
2. Menganalisis sumber daya yang dapat digunakan dalam pembuatan daftar kata kunci

I.3. Tujuan

Tujuan akhir dari tugas akhir ini adalah :

1. Menghasilkan daftar kata kunci dengan pendekatan dari hasil eksperimen pengklasifikasian opini *Twitter* yang memberikan akurasi tertinggi
2. Mengevaluasi pendekatan untuk pembuatan daftar kata kunci yang cocok untuk penanganan masalah pengklasifikasian opini dalam Bahasa Indonesia

I.4. Batasan Masalah

Batasan masalah pada tugas akhir ini adalah :

1. Daftar hanya berisi kata-kata yang berhubungan dengan kategori dan tidak menunjukkan arti dari kata tersebut
2. Pengklasifikasian opini hanya pada pesan atau akun *Twitter* yang berhubungan dengan pemerintahan pada kota Bandung
3. Daftar kata kunci dapat berisi kata dengan 2 atau lebih label yang berbeda

I.5. Metodologi

Berikut metodologi yang digunakan untuk mencapai tujuan yang telah ditentukan :

1. Eksplorasi

   Pada tahap ini, dilakukan pembelajaran mengenai pendekatan korpus dalam pembuatan daftar kata kunci (*clustering* dan *association rule*), fitur-fitur yang digunakan pada pendekatan korpus, serta kamus *WordNet y*ang digunakan pada pendekatan kamus dan metode pencarian *word similarity* pada kamus. Selain itu juga, dilakukan pembelajaran mengenai penelitian terkait yang menggunakan pendekatan yang sama, seperti hasil penggunaan

pendekatan korpus dan kamus pada pencarian *Word Similarity* dari kata-kata tertentu ,ataupun penelitian yang menghasilkan daftar kata kunci sentimen dengan penelusuran kata-kata di *WordNet*, yaitu *SentiWordNet* 3.0. Pembelajaran tersebut dilakukan pada karya tulis yang mempunyai topik sesuai pembelajaran yang bersangkutan. Hasil Eksplorasi ini digunakan sebagai dasar tahap-tahap selanjutnya.

2. Analisis Metode

   Analisis yang dilakukan adalah perbandingan pendekatan pada pembuatan daftar kata, masalah-masalah yang mungkin muncul pada tahap implementasi, penjelasan secara rinci untuk pendekatan yang digunakan, dan penjelasan mengenai evaluasi yang dilakukan.

3. Pengumpulan data

   Tahap pengumpulan data dilakukan untuk mengumpulkan data dari *Twitter*, Lapor.ukp.go.id, dan artikel berita yang akan digunakan sebagai data latih dan data uji. Selain pengumpulan data, pada tahap ini juga akan dilakukan pelabelan data sesuai dengan kategori hasil klasifikasi opini yang diinginkan.

4. Implementasi

   Tahap implementasi merupakan tahap pembentukan daftar kata kunci dari pendekatan dan fitur yang telah ditetapkan. Sumber data dari daftar kata kunci didapatkan dari hasil tahap pengumpulan data yang dijadikan data latih. Daftar kata kunci yang dihasilkan dari setiap metode akan disimpan untuk dibandingkan pada tahap evaluasi.

5. Evaluasi

   Terdapat eksperimen yang dilakukan untuk menguji pembuatan daftar ini :

   a. Pemeriksaan hasil pelabelan kata-kata yang berhubungan pada daftar kata
   b. Pengujian masing-masing metode yang digunakan

I.6. Jadwal Pelaksanaan Tugas Akhir

Berikut adalah jadwal pengerjaan tugas akhir saat ini (Table I-1). Diharapkan jadwal dapat lebih dipercepat bila memungkinkan dan disesuaikan dengan keadaan pengerjaan tugas akhir.

Untuk bimbingan akan dilakukan secara berkala 1 minggu 1 kali untuk melakukan pelaporan kemajuan pengerjaan tugas akhir atau disesuaikan dengan kebutuhan bimbingan untuk tugas akhir. Diharapkan bahwa pelaksanaan sidang dapat dilakukan pada Bulan Mei.

Tabel I-1. Tabel jadwal pengerjaan Tugas Akhir

No.	Metodologi	Sep	Okt	Nov	Des	Jan	Feb	Mar	Apr	Mei
1	Eksplorasi	√	√	√	√	√				
2	Analisis Metode				√	√	√			
3	Pengumpulan Data					√	√			
4	Implementasi						√	√		
5	Evaluasi						√	√	√	
6	Dokumentasi	√	√	√	√	√	√	√	√	√

BAB II.

STUDI LITERATUR

Pembuatan daftar kata kunci ini akan menggunakan pendekatan korpus sebagai pendekatan utama. Pendekatan korpus merupakan pendekatan dengan mencari kata-kata yang berpengaruh untuk menentukan kategori dari beberapa dokumen yang dijadikan data latih. Untuk meningkatkan kualitas dari daftar kunci hasil, akan digunakan juga pendekatan secara kamus. Pembuatan daftar kata kunci dengan pendekatan kamus merupakan pendekatan dengan menelusuri kamus *WordNet* Bahasa Indonesia. Studi Literatur ini akan menjelaskan lebih lanjut tentang hal-hal yang mendukung pelaksanaan eksperimen dan penelitian-penelitian lain yang terkait.

II.1. Pendekatan Korpus

Pendekatan Korpus merupakan pendekatan untuk membuat daftar kunci dengan menelusuri dokumen-dokumen yang berhubungan dengan salah satu kategori dalam daftar kunci. Dari dokumen-dokumen tersebut, akan dicari kata-kata yang mungkin berpengaruh untuk melakukan klasifikasi dokumen. Pengaruh dari setiap kata yang ditelusuri ditunjukkan dalam beberapa fitur yang diambil dari kata tersebut. Dari fitur-fitur yang ada, akan diterapkan algoritma-algoritma untuk menghasilkan daftar kata kunci yang diinginkan.

II.1.1. Fitur

Fitur merupakan nilai-nilai yang digunakan untuk menerapkan algoritma-algoritma pada pendekatan korpus. Fitur-fitur tersebut menentukan posisi dari kata terhadap dokumen atau kategori tertentu. Beberapa fitur yang dapat diambil dari teks atau dokumen adalah TFxIDF, *Information Gain* (IG), *Mutual Information* (MI), dan X^2 *Statistic* (CHI).

II.1.1.1. TFxIDF

TFxIDF (*Term Frequency x Inverse Document Frequency*) adalah statistik numerik yang menunjukkan pentingnya kata pada dokumen dalam korpus (Rajaraman, Leskovec, & Ullman, 2011). Umumnya TFxIDF digunakan sebagai faktor untuk menghitung bobot pada pengambilan informasi. Nilai TFxIDF meningkat setiap banyaknya kata muncul pada dokumen, namun turun apabila frekuensi kata sering muncul pada korpus, hal ini untuk menangani kata-kata yang sering muncul. Oleh karena itu, bobot yang dihasilkan dari TFxIDF dapat dijadikan salah satu fitur untuk melakukan *Clustering*/ pengelompokan dari kata. TF dapat ditunjukkan dengan 3 cara (Salton & Buckley, 1988), yaitu :

1. Menggunakan nilai biner, yaitu diberi nilai 1 untuk kata-kata yang terdapat pada dokumen dan nilai 0 terhadap kata-kata yang tidak muncul pada dokumen. Pada konsep ini, frekuensi kemunculan kata tidak dimasukkan ke dalam perhitungan.
2. Menggunakan nilai frekuensi kemunculan kata secara langsung untuk menjadi TF
3. Menggunakan nilai pecahan term yang telah dilakukan normalisasi, rumus tersebut dapat dilihat pada rumus (II.1) (Salton & Buckley, 1988):

$$\mathrm{TF(t,d)} = 0{,}5 + 0{,}5\,\frac{\mathrm{f(t,d)}}{\max\{\mathrm{f(w,d)}:\mathrm{w} \in \mathrm{d}\}} \qquad (\mathrm{II.1})$$

Pada rumus (II.1) (t,d) merupakan frekuensi kata t muncul pada dokumen d dan $\max\{f(w,d): w \in d\}$ merupakan frekuensi maksimum dari term lain pada dokumen d.

Sedangkan untuk menghitung IDF digunakan rumus sebagai berikut (Salton & Buckley, 1988):

$$IDF(t,D) = \log\frac{N}{Df(t,D)} \qquad (\mathrm{II.2})$$

Pada rumus (II.2) N merupakan jumlah dokumen pada korpus dan $Df(t,D)$ sebagai banyak dokumen dalam kumpulan dokumen D yang mengandung *term* t. Namun

bila term tidak muncul pada korpus, maka akan terdapat nilai 0 pada pembagian, sehingga perlu penanganan untuk menggantinya menjadi 1+$Df(t, D)$

Untuk menghitung TFxIDF dari kata, digunakan rumus :

$$TFxIDF(t, d, D) = tf(t, d)\ x\ idf(t, D) \qquad (II.3)$$

Dari rumus (II.3) akan didapatkan nilai yang dapat dijadikan sebagai pembobotan kata pada saat dilakukan pengelompokan kata.

Tabel II-1. Contoh dokumen data latih

Angkutan Sampah di Kab. Bandung Baru Capai 30 Persen

SOREANG, (PRLM).- Pengangkutan sampah di Kab. Bandung baru mencapai 30 persen dari produksi sampah sekitar 7.000 meter kubik per harinya. Sisa sampah yang tidak terangkut terbuang ke aliran sungai, pinggir jalan, maupun dibuang dalam lubang-lubang yang dibuat masyarakat.

"Seharusnya tak ada istillah sampah tidak terangkut karena sampah merupakan pelayanan pemerintah," kata Wabup Bandung Deden Rukman Rumaji, dalam rakor pengelolaan sampah di Hotel Aston, Senin (20/1).

Rakor dihadiri Bupati Bandung H. Dadang M. Naser, para kepala organisasi perangkat daerah (OPD), para camat, dan para kepala UPTD pasar se-Kab. Bandung. Rakor diisi pemaparan pengelolaan sampah dari Direktur PT Bumi Resik, Jaka Winarso.

Lebih jauh Deden Rumaji mengatakan, pengelolaan sampah di Kab. Bandung terhambat dengan minimnya anggaran. "Pemkab Bandung memprioritaskan perbaikan jalan-jalan menjadi jalan mantap dengan target 2015. Akhirnya sampah harus kita kelola dengan cara gotong royong atau sabilulungan," katanya.

Menurut Deden, persoalan sampah di Kab. Bandung bukan kepada tempat pembuangan akhir karena TPA Babakan masih mencukupi sampai tahun 2020. "Persoalan sampah pada pemilahan dari rumah tangga sampai pengangkutan dari TPS ke TPA," katanya.(A-71/A-108)***

Contoh aplikasi TFxIDF dengan nilai TF yang dinormalisasi: Bila terdapat kumpulan dokumen D sebanyak 10.000 dokumen dan terdapat kata “jalan” yang muncul pada 100 dokumen dari kumpulan dokumen tersebut. Maka :

$$IDF(jalan, D) = \log\frac{10.000}{100} = 2$$

Pada salah satu dokumen d dari Tabel II-1,dapat diketahui bahwa f(jalan) adalah sebesar 4 kata dan jumlah maksimum frekuensi kata yang muncul adalah kata sampah, yaitu sebanyak 12 kata. Maka :

$$TF(jalan, d) = 0{,}5 + 0{,}5\frac{4}{12} = \frac{4}{6}$$

$$TFxIDF(jalan, d, D) = \frac{4}{6} \times 2 = \frac{4}{3}$$

Dan bila kata sampah muncul pada 10.000 dokumen dan sebanyak 12 kata muncul pada Tabel II-1, maka :

$$IDF(sampah, D) = \log\frac{10.000}{10.000} = 1$$

$$TF(sampah, d) = 0{,}5 + 0{,}5\frac{12}{12}$$

$$TFxIDF(sampah, d, D) = 1 \times 1 = 1$$

Dari kedua contoh di atas, dapat dilihat bahwa kata yang mempunyai pengaruh lebih tinggi adalah kata yang paling sedikit muncul pada kumpulan dokumen dan jumlah frekuensi kata yang paling tinggi pada salah satu dokumen. Hal ini dikarenakan kata yang lebih berpengaruh merupakan kata yang dapat membedakan beberapa dokumen dari seluruh kumpulan dokumen tersebut.

II.1.1.2. Information Gain (IG)

Information gain (IG) merupakan nilai yang menyatakan kesesuaian kata terhadap kriteria yang ditetapkan pada pembelajaran mesin (Mitchell, 1997). Nilai tersebut mengukur banyak informasi yang didapat berdasarkan kategori tertentu dengan melihat banyak kata muncul dan tidak muncul pada dokumen. Nilai IG dari suatu kata dapat didefinisikan dengan rumus berikut (Yang & Pedersen, 1997) :

$$IG(t) = -\sum_{i=1}^{m} P(Ci)\, log\, P(Ci) + \sum_{i=1}^{m} P(Ci|t)\, log\, P(Ci|t) + \sum_{i=1}^{m} P(Ci|\bar{t})\, log\, P(Ci|\bar{t}) \quad (II.4)$$

Pada rumus (II.4), P(Ci|t) merupakan probabilitas kategori pada dokumen dimana kata t ditemukan, $\sum_{i=1}^{m} P(Ci)$ merupakan probabilitas seluruh kategori yang mungkin untuk dokumen, sedangkan $P(Ci|\bar{t})$ merupakan probabilitas seluruh kategori dimana kata-kata pada dokumen selain kata yang ingin dicari nilai informasinya ditemukan. Untuk memproses dokumen yang mempunyai kategori yang sama, maka rumus (II.4) dapat diubah menjadi rumus (II.5):

$$\mathrm{IG(t)} = -\mathrm{P(C)}\log\mathrm{P}(C) + \mathrm{P(C|t)}\log\mathrm{P}(C|t) + \mathrm{P(C|\bar{t})}\log\mathrm{P}(C|\bar{t})$$

$$= \mathrm{P(C|t)} \log \mathrm{P}(C|t) + \mathrm{P(C|\bar{t})} \log \mathrm{P}(C|\bar{t}) \quad \text{(II.5)}$$

Rumus (II.5) dihasilkan karena P(C) akan menghasilkan 1 untuk dokumen dengan kategori yang sama dan karena log(1) = 0, maka [–P(C) log P(C)] dapat dihilangkan.

Contoh aplikasi IG: Bila terdapat kumpulan dokumen D sebanyak 10.000 dokumen dan terdapat kata "jalan" yang muncul pada 100 dokumen dari kumpulan dokumen tersebut. Maka :

$$\text{IG(jalan)} = \frac{100}{10.000} \log \frac{100}{10.000} + \frac{9.900}{10.000} \log \frac{9.900}{10.000} = -0{,}024$$

II.1.1.3. Mutual Information (MI)

Mutual Information (MI) merupakan fitur yang sering digunakan untuk memodelkan hubungan antar kata (Yang & Pedersen, 1997). Bila terdapat kata t dan kategori c, dengan A sebagai banyak dokumen dengan t dan c muncul secara bersamaan, B sebagai banyak dokumen dengan t muncul tanpa c, C sebagai banyak dokumen dengan c muncul tanpa t, N sebagai jumlah seluruh dokumen, P(t∩c) sebagai probabilitas kata t dan muncul pada kategori c, P(t) sebagai probabilitas kata t muncul dan P(c) sebagai probabilitas kategori c dipilih. Maka MI dapat didefinisikan sebagai berikut (Church & Hanks, 1990):

$$\mathrm{MI(t,c)} = \log \frac{P(t \cap c)}{P(t) \times P(c)} \quad \text{(II.6)}$$

Dan dapat juga dihitung dengan (Yang & Pedersen, 1997) :

$$\mathrm{MI(t,c)} \approx \log \frac{A \times N}{(A+C) \times (A+B)} \quad \text{(II.7)}$$

II.1.1.4. X^2 *Statistic* (CHI)

X^2 *Statistic* (CHI) merupakan perhitungan tingkat keterkaitan antara kata(t) dan kategori(c). Perbedaan terbesar antar CHI dan MI adalah bahwa CHI merupakan nilai hasil normalisasi, sehingga nilainya dapat dibandingkan secara langsung antara *term* dengan kategori yang sama. Kelemahan dari CHI adalah kinerjanya yang buruk untuk pemrosesan dokumen dengan frekuensi *term* yang kecil. Bila terdapat kata t dan kategori c, dengan A sebagai banyak dokumen dengan t dan c

muncul secara bersamaan, B sebagai banyak dokumen dengan t muncul tanpa c, C sebagai banyak dokumen dengan c muncul tanpa t, D sebagai banyak dokumen dengan c dan t tidak muncul, dan N sebagai jumlah seluruh dokumen, maka nilai keterkaitan kata dengan kategori dapat didefinisikan sebagai berikut (Yang & Pedersen, 1997) :

$$X^2(t,c) = \frac{N \times (AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \quad (\text{II.8})$$

II.1.2. Algoritma pendekatan korpus

Pendekatan secara korpus menggunakan teknik *Association Rule* (Agrawal, Imielinski, & Swami, 1993) dengan algoritma Apriori dan *clustering* / *cluster analysis* (Mitchell, 1997) dengan 1 atau lebih fitur yang dapat digunakan untuk melakukan pengelompokan terhadap kata-kata yang termasuk dan tidak termasuk ke dalam daftar kata kunci. Agar hasil yang didapat mempunyai akurasi yang lebih baik, perlu dilakukan pemrosesan awal pada dokumen-dokumen yang menjadi data latih dan dilakukan ekstraksi fitur-fitur dari setiap kata untuk menentukan kelompok dari setiap kata.

II.1.2.1. Association Rule Learning

Association Rule Learning merupakan pembelajaran yang mencari hubungan menarik antar variabel dalam basis data yang besar. Salah satu pengaplikasian teknik ini adalah untuk mencari hubungan antar produk (produk-produk yang dibeli dalam 1 transaksi yang sama) pada basis data transaksi supermarket (Agrawal, Imielinski, & Swami, 1993), contohnya bila ditemukan aturan {bawang,kentang}=>{susu} maka dapat dinyatakan bahwa pelanggan yang membeli bawang dan kentang dalam 1 transaksi memiliki kemungkinan yang cukup besar untuk membeli susu juga.

Dengan menggunakan teknik ini sebagai dasar, terdapat penelitian yang membentuk *classifier* term (Antonie & Zai'ane, 2002) dari dokumen latih untuk melakukan klasifikasi pada dokumen-dokumen selanjutnya. Penentuan aturan term dari dokumen latih pada penelitian tersebut menggunakan algoritma Apriori.

Algoritma ini hanya akan memeriksa kombinasi term yang mempunyai jumlah kemunculan di atas *threshold* tertentu dan mengeliminasi term yang berada di bawah *threshold*, dengan demikian biaya komputasi dari aturan akan berkurang. Pencarian aturan akan selesai pada saat semua kombinasi term telah selesai diperiksa dan setiap aturan yang terbentuk diberi label sesuai kategori dokumennya. Saat model klasifikasi telah selesai dibuat, maka dilakukan evaluasi dengan melakukan klasifikasi pada dokumen-dokumen baru. Pengklasifikasian dokumen baru dilakukan dengan melihat aturan yang paling memenuhi untuk dokumen tersebut. Secara garis besar, *pseudo code* algoritma apriori dari *Association Rule Learning* dapat dilihat pada Tabel II-2 (Antonie & Zai'ane, 2002). Eksperimen dilakukan pada data yang merupakan bagian dari data Reuters, yaitu sebanyak 12202 dokumen dengan 9603 dokumen latih dan 3299 dokumen uji. Terdapat 135 topik yang seharusnya menjadi hasil klasifikasi dokumen, namun pada eksperimen tersebut dipilih 10 kategori dengan jumlah dokumen terbanyak. Hasil eksperimen ini dapat dilihat pada Tabel II-3.

Tabel II-2. Pseudo code Association Rule (Antonie & Zai'ane, 2002)

Algoritma *Association Rule*, untuk menemukan aturan asosiasi dari data latih yang telah dibagi berdasarkan kelompok kategorinya

Input kumpulan dokumen D_i dengan C_i sebagai kategori dari kelompok dokumen dan t_j sebagai *term* yang sedang diproses pada dokumen; *threshold* minimum

Output kumpulan aturan dalam bentuk $t_1 \wedge t_2 \wedge t_3 \wedge ..... \wedge t_n \Rightarrow C_i$ dimana C_i sebagai kategori dan t_j sebagai *term*

Metode :

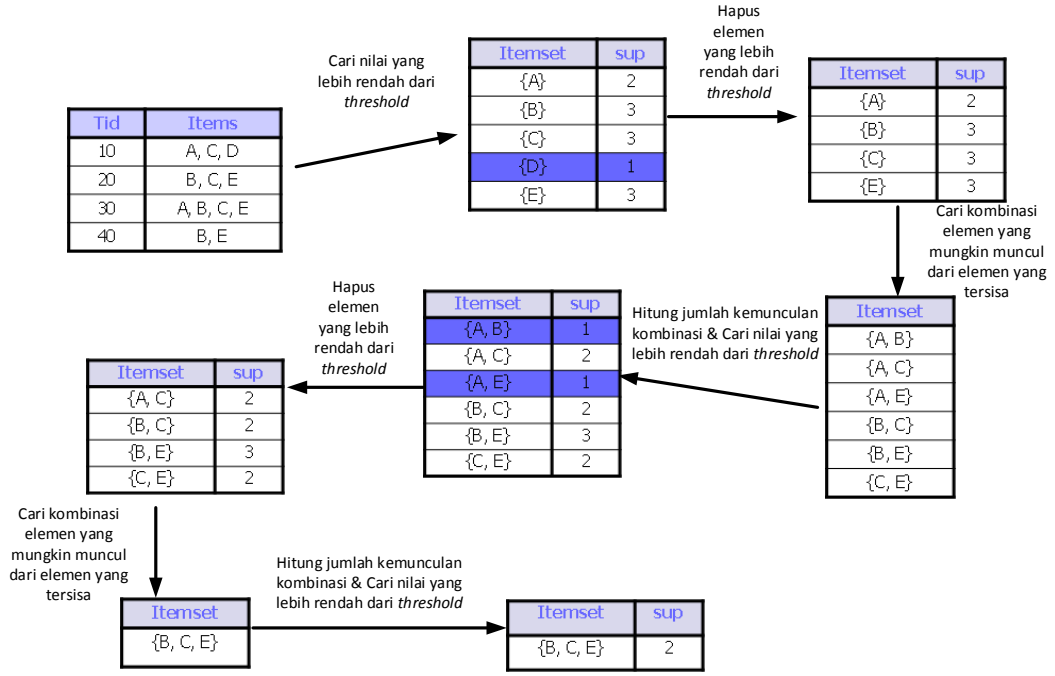
(1). $C_i \leftarrow$ {kandidat 1 dari kelompok *term*}
(2). $F_i \leftarrow$ {*Frequent* 1 dari kelompok *term*}
(3). For ($i \leftarrow 2$; $F_{i-1} \neq 0$; $i \leftarrow i+1$) do{
(4). $C_i \leftarrow (F_{i-1} \bowtie F_{i-1})$
(5). $C_i \leftarrow C_1 - \{c | (i-1)$ item-set of $c \notin F_{i-1}\}$
(6). $D_i \leftarrow$ FilterTable(D_{i-1}, F_{i-1})
(7). foreach document d in D_i do{
(8). foreach c in C_i do{
(9). c.support←c.support+Count(c,d)
(10). }
(11). }
(12). $F_i \leftarrow \{c \in C_i \mid c.support > \sigma\}$
(13). }
(14). Sets $\leftarrow U_i\{c \in F_i \mid i > 1\}$
(15). R=0

```
(16).  foreach itemset I in Sets do {
(17).       R←R+{I => Cat}
(18).  }
```

Tabel II-3. Hasil klasifikasi dokumen reuter ke dalam 10 kategori dengan Association Rule (Antonie & Zai'ane, 2002)

BEP	ARC-BC with δ=50 10%	15%	20%	Bayes	Rocchio	C4.5	k-NN	bigrams	SVM (poly)	SVM (rbf)
acq	**90.9**	89.9	87.8	91.5	92.1	85.3	92.0	73.2	94.5	95.2
corn	69.6	**82.3**	70.9	47.3	62.2	87.7	77.9	60.1	85.4	85.2
crude	77.9	77.0	**80.7**	81.0	81.5	75.5	85.7	79.6	87.7	88.7
earn	**92.8**	89.2	86.6	95.9	96.1	96.1	97.3	83.7	98.3	98.4
grain	68.8	72.1	**73.1**	72.5	79.5	89.1	82.2	78.2	91.6	91.8
interest	70.5	70.1	**75.3**	58.0	72.5	49.1	74.0	69.6	70.0	75.4
money-fx	70.5	**72.4**	70.5	62.9	67.6	69.4	78.2	64.2	73.1	75.4
ship	**73.6**	73.2	63.0	78.7	83.1	80.9	79.2	69.2	85.1	86.6
trade	68.0	69.7	**69.8**	50.0	77.4	59.2	77.4	51.9	75.1	77.3
wheat	84.8	**86.5**	85.3	60.6	79.4	85.5	76.6	69.9	84.5	85.7
micro-avg	**82.1**	81.8	81.1	72.0	79.9	79.4	82.3	73.3	85.4	86.3
macro-avg	76.74	**78.24**	76.32	65.21	79.14	77.78	82.05	67.07	84.58	86.01

Bila telah ditentukan *threshold* dari jumlah kemunculan minimal adalah sebesar 2 dan terdapat 4 dokumen dengan masing-masing mempunyai *term* : {(A, C, D), (B, C, E), (A,B,C,E), (B, E)}, maka contoh dari penerapan ARL terhadap kasus tersebut dapat dilihat pada Gambar II-1.



Gambar II-1. Contoh penerapan ARL

II.1.2.2. K-*Means* Clustering / Cluster Analysis

Cluster Analysis merupakan metode pengelompokan beberapa objek yang mempunyai kedekatan tertentu dibandingkan dengan objek pada kelompok lain. *Clustering* memiliki beberapa algoritma yang telah dikembangkan untuk menyelesaikan masalah-masalah pada pengelompokan objek, diantarnya adalah :

1. *Hierarchical Clustering*

   Hierarchical Clustering merupakan algoritma yang mengadopsi pemikiran bahwa objek lebih berhubungan dengan objek yang berdekatan dibandingkan dengan objek yang jauh. Algoritma ini menghubungkan objek-objek untuk membentuk kelompok berdasarkan jarak antar objek. Suatu kelompok dapat dinyatakan dengan menentukan jarak maksimum untuk mengelompokkan objek.

2. *Centroid-based clustering*

   Algoritma ini merepresentasikan kelompok sebagai vektor utama yang mungkin bukan bagian dari kumpulan data. Algoritma centroid ini disebut sebagai *K-Means*. Namun algoritma ini memerlukan jumlah pengelompokan yang diinginkan agar dapat menentukan kelompok dari objek.

3. *Distribution-based clustering*

   Metode ini menggunakan pemodelan statistik untuk mendapatkan kelompok. Namun metode ini mengalami masalah *overfitting* (terlalu spesifik untuk permasalahan tertentu), kecuali diberikan batasan pada kompleksitas dari model. Salah satu algoritma pada metode ini adalah model *Gaussian mixture*.

4. *Density-based clustering*

   Metode ini mengelompokkan objek dengan mendefinisikan area dengan kerumunan data yang lebih tinggi bila dibandingkan dengan kumpulan data lainnya. Algoritma yang digunakan salah satunya adalah DBSCAN.

Pengelompokan kata-kata pada dokumen membutuhkan bobot yang dapat memperhitungkan jarak atau tingkat keterhubungan antar kata agar dapat

dikelompokkan ke dalam kelompok yang tepat. Bobot untuk melakukan pengelompokan ini umumnya disebut sebagai fitur.

Banyak penelitian dalam pemrosesan teks menggunakan algoritma K-*Means* (dapat dilihat pada referensi : {(Dash, Choi, Scheuermann, & Liu, 2002), (Liu, Liu, Chen, & Ma, 2003), (Abbas, 2008)}). K-*Means* akan membagi data ke dalam k kelompok yang ditentukan secara manual terlebih dahulu. Algoritma ini sering digunakan karena mudah untuk diimplementasikan dan mempunyai kompleksitas waktu sebesar O(n), dengan n merupakan jumlah pola/ kelompok (Jain, Murty, & Flynn, 1999).

Langkah-langkah dari implementasi algoritma K-*Means* adalah sebagai berikut (Jain, Murty, & Flynn, 1999):

1. Pilih k pusat kelompok sesuai dengan banyak kelompok yang diinginkan atau pilih secara acak dari kumpulan data
2. Tempatkan data lain pada pusat kelompok dengan jarak yang terdekat
3. Hitung ulang titik pusat dari kelompok dengan menggunakan seluruh anggota kelompok
4. Bila belum konvergen, ulangi langkah 2. Kriteria dari konvergen adalah : tidak terjadi (atau hanya terjadi sedikit) pengubahan titik pusat dari kelompok yang baru, atau perubahan minimal pada rumus :

$$e^2(K,L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left\| {X_i}^j - c_j \right\|^2 \qquad (\text{II.9})$$

Pada rumus (II.9) $e^2(K,L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left\| {X_i}^j - c_j \right\|^2$ (II.9), e^2 menunjukkan tingkat kesalahan, K menunjukkan banyak kelompok, L merupakan data yang dikelompokkan, ${X_i}^j$ merupakan pole ke i pada kelompok ke j, dan c_j merupakan pusat dari kelompok ke j.

Dari langkah-langkah tersebut, pemilihan pusat dari masing-masing kelompok merupakan hal yang penting dalam menentukan kualitas kelompok yang dihasilkan. Dari penelitian yang dilakukan Penã (Penã, Lozano, & Larrañaga, 1999), ditemukan bahwa algoritma yang diajukan oleh Kaufman dan Rousseeuw (Kaufman & Rousseeuw, 1990), yaitu *Kaufman Approach* (KA), menghasilkan

kelompok yang terbaik. *Pseudocode* dari penentuan KA dapat dilihat pada Gambar II-2.

```
Langkah 1. Pilih pusat dari data yang paling terkumpul sebagai
           titik yang pertama
Langkah 2. Untuk setiap data yang tidak terpilih wi lakukan
     Langkah 2.1. Untuk setiap data yang tidak terpilih wj lakukan
                  Hitung Cji=max ( Dj - dji, 0) dimana dji = || wi -
                  wj|| dan Dj= minsdsj dengan s sebagai salah satu
                  data yang telah terpilih
     Langkah 2.2. Hitung gain dari pemilihan wi dengan Σj Cji
Langkah 3. Pilih data wi yang belum terpilih dan dapat
           memaksimalkan Σj Cji
Langkah 4. Jika terdapat K titik yang telah terpilih, maka berhenti
           Jika tidak, ulangi langkah 2
Langkah 5. Implementasikan algoritma clustering yang dapat
           mengelompokkan data yang belum terpilih ke titik
           terdekat
```

Gambar II-2. Pseudocode metode inisialisasi KA (Penã, Lozano, & Larrañaga, 1999)

II.2. Pendekatan Kamus

Pendekatan Kamus merupakan pembuatan daftar kata kunci melalui penelusuran kamus elektronik yang telah ada, dalam hal ini kamus *WordNet*. Penelusuran kamus ini berupa pengambilan kata-kata yang berhubungan dengan kategori yang telah ditetapkan. Hubungan tersebut dapat berupa sinonim, antonim, hipernim ataupun hiponim dari kategori. Keterhubungan ini telah digambarkan dalam bentuk *synset* pada *WordNet*, namun hasil pencarian hubungan juga dapat dilakukan dengan melihat *Word Similarity* antar kata.

II.2.1. *WordNet*

WordNet (Esuli & Sebastiani, 2006.) merupakan basis data leksikal dari Bahasa Inggris. Pada basis data ini dilakukan juga pengelompokan kata ke dalam *synset* (*synonim set*) sesuai konsep dari kata tersebut. setiap *synset* dikelompokkan berdasarkan kata benda, kata sifat, kata kerja, dan kata keterangan dari kata tersebut. Berikut adalah contoh hasil pencarian kata "*word*"pada *WordNet* (Tabel II-4) :

Tabel II-4. Contoh hasil pencarian pada WordNet

```
The noun word has 10 senses (first 10 from tagged texts)
```

```
1. (958) word -- (a unit of language that native speakers can identify;
"words are the blocks from which sentences are made"; "he hardly said
ten words all morning")

2. (101) word -- (a brief statement; "he didn't say a word about it")

3. (31) news, intelligence, tidings, word -- (new information about
specific and timely events; "they awaited news of the outcome")

4. (21) Son, Word, Logos -- (the divine word of God; the second person
in the Trinity (incarnate in Jesus))

5. (11) parole, word, word of honor -- (a promise; "he gave his word")

6. (10) password, watchword, word, parole, countersign -- (a secret word
or phrase known only to a restricted group; "he forgot the password")

7. (8) discussion, give-and-take, word -- (an exchange of views on some
topic; "we had a good discussion"; "we had a word or two about it")

8. (7) Bible, Christian Bible, Book, Good Book, Holy Scripture, Holy
Writ, Scripture, Word of God, Word -- (the sacred writings of the
Christian religions; "he went to carry the Word to the heathen")

9. (5) word -- (a verbal command for action; "when I give the word,
charge!")

10. (3) word -- (a word is a string of bits stored in computer memory;
"large computers use words up to 64 bits long")


The verb word has 1 sense (first 1 from tagged texts)


1. (2) give voice, formulate, word, phrase, articulate -- (put into words
or an expression; "He formulated his concerns to the board of trustees")
```

Pada contoh di atas dapat dilihat bahwa konsep (*sense*) dari setiap jenis kata akan ditunjukkan terlebih dahulu (bila terdapat jenis kata untuk kata tersebut). Pada penjelasan masing-masing konsep dapat diketahui bahwa :

Kolom 1 : nomor dari konsep pada jenis kata tersebut

Kolom 2 : ID *WordNet*

Kolom 3 : kata-kata yang termasuk ke dalam *synset*

Kolom 4 : penjelasan dari konsep *synset* dan contoh penggunaan kata pada kalimat.

WordNet saat ini terdiri dari 147278 kata yang unik, namun terdapat kata yang mempunyai lebih dari 1 cara/aturan penggunaan pada kalimat. Statistik lengkapnya dapat dilihat pada Tabel II-5.

Tabel II-5. Statistik distribusi kata pada WordNet

POS Tag / Jenis Kata	Banyak Kata Unik	*Synset*	Total pasangan kata-sense
Kata Benda	117798	82115	146312
Kata Kerja	11529	13767	25047
Kata Sifat	21479	18156	30002
Kata Keterangan	4481	3621	5580
Total	155287	117659	206941

Saat ini telah ada *WordNet* Bahasa Indonesia (Noor, Sapuan, & Bond, 2011) yang kata-katanya telah diterjemahkan dalam bahasa Indonesia, namun untuk arti dari konsep katanya belum dapat diterjemahkan ke dalam Bahasa Indonesia. Berikut adalah contoh tampilan *WordNet* Bahasa Indonesia :

```
00015388-n    B      X      fauna
00015388-n    M      Y      haiwan
00015388-n    I      Y      hewan
```

Keterangan dari contoh tampilan *WordNet* Bahasa Indonesia :

Kolom 1 : ID *WordNet* - jenis kata

Kolom 2 : bahasa hasil terjemahan :

- I (Indonesian = ind)
- M (Malay = zsm)

- B (Bahasa = msa) -> gabungan dari ind dan zsm

Kolom 3 : kualitas hasil terjemahan (untuk versi yang umumnya di rilis hanya hasil terjemahan dengan kualitas Y dan O) :

- Y = sudah dievaluasi secara manual dan berkualitas baik
- O = hasil otomatis dalam kualitas yang baik
- M = hasil otomatis dalam kualitas yang cukup baik
- L = hasil otomatis dalam kualitas yang (mungkin) buruk
- X = sudah dievaluasi secara manual dan berkualitas buruk

II.2.2. Word Similarity dengan Pendekatan Kamus

Word Similarity atau dapat disebut sebagai *semantic similarity* merupakan keterhubungan antar kata pada sekumpulan dokumen ataupun data. Keterhubungan ini dapat berupa sinonim, antonim, hiponim ataupun hipernim dari kata yang dimaksud. Untuk mencari keterhubungan antar kata pada dokumen dapat digunakan 2 pendekatan (Mihalcea, Corley, & Strapparava, 2006), yaitu pendekatan korpus dan pendekatan kamus. Pendekatan korpus dari *word similarity* menggunakan fitur yang sama, yaitu TFxIDF. Oleh karena itu, untuk perhitungan *word similarity* hanya digunakan pendekatan dengan kamus.

Pendekatan Kamus dari *word similarity* merupakan metode pencarian kata pada kamus dan menentukan tingkat keterhubungan antar kata dari hasil pencarian kamus elektronik. Pada pendekatan kamus, untuk mendapatkan tingkat keterhubungan kata dapat didefinisikan sebagai beberapa rumus berikut :

- Leacock & Chodorow : *length* merupakan panjang dari jalur terpendek antara 2 konsep pada perhitungan simpul, dan D adalah kedalaman maksimum dari taksonomi. Rumus keseluruhan dapat dilihat pada rumus (II.10).

$$\text{Sim}_{\text{lch}} = -log\,\frac{length}{2*D} \qquad \text{(II.10)}$$

- Lesk : kesamaan dari 2 konsep yang didefinisikan sebagai fungsi dari irisan antara definisi yang berhubungan, sesuai yang tertulis pada kamus.

- Wu and Palmer : perhitungan *metric* kesamaan untuk kedalaman dari 2 konsep pada taksonomi *WordNet*, dan kedalaman dari LCS (*least common subsumer*), dan digabungkan menjadi persamaan pada rumus (II.11).

$$\text{Sim}_{\text{wup}} = \frac{2 * depth(LCS)}{depth(concept1) + depth(concept2)} \quad (\text{II.11})$$

- Resnik : mengembalikan *information content* (IC) dari LCS 2 konsep :

$$\text{Sim}_{\text{res}} = \text{IC}(\text{LCS}) \quad (\text{II.12})$$

$$\text{IC}(c) = -\log P(c) \quad (\text{II.13})$$

  Pada rumus (II.13), P(c) adalah probabilitas menemukan instansiasi konsep c dari korpus yang besar.

- Jiang & Conrath :

$$\text{Sim}_{\text{jnc}} = \frac{1}{IC(concept1) + IC(concept2) - 2 * IC(LCS)} \quad (\text{II.14})$$

  IC pada rumus (II.14) menyatakan *information content* dari konsep. Perhitungan *information content* dapat menggunakan rumus (II.13).

Semua nilai yang dihasilkan dari rumus similaritas berada pada interval antara 0 – 1, dengan 0 menyatakan kata sangat tidak berhubungan dan 1 menyatakan kata sangat berhubungan.

II.3. Penelitian Terkait

Penelitian mengenai pembuatan daftar kata kunci telah banyak dikembangkan, penelitian ini dapat disebut juga sebagai *latent sematic indexing* (LSI). LSI merupakan metode untuk membuat daftar kata untuk mengidentifikasi pola antara kata dengan dokumennya pada kumpulan dokumen yang tidak terstruktur. Salah satu penerapannya adalah untuk melakukan kategorisasi terhadap dokumen-dokumen yang mempunyai term dengan bobot yang menyerupai (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990). Pembobotan yang digunakan dapat berupa nilai *boolean* (1 untuk kata-kata yang terdapat pada dokumen dan 0 untuk kata-kata yang tidak terdapat pada dokumen) dan banyak kemunculan kata dalam dokumen. Pada Gambar II-3, dapat dilihat bahwa setiap kata diberi nilai banyak

kemunculannya pada judul dokumen. Dari pola kemunculan tersebut dicari hubungan antara judul. Dengan menggunakan pola yang didapatkan dari data latih, dapat dilakukan klasifikasi terhadap data-data baru yang akan masuk.

Titles:

c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

Terms	**Documents**								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Gambar II-3. contoh penggunaan LSI (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990)

Selain untuk membuat kamus, pendekatan penelusuran kamus dan dokumen ini telah banyak diaplikasikan dalam kasus pemrosesan teks lainnya. Salah satunya untuk mengukur nilai similaritas antar kata pada dokumen (Mihalcea, Corley, & Strapparava, 2006). Pada penelitian ini, algoritma yang digunakan pada pendekatan korpus adalah *Pointwise Mutual Information* dengan data dari hasil *Information Retrieval* (PMI-IR) dan *Latent Semantic Analysis* (LSA). Sedangkan untuk pendekatan kamus, evaluasi dilakukan dengan menghitung kedekatan kata pada *WordNet*. Pada pendekatan kamus, digunakan beberapa ilmu perhitungan kedekatan kata yang telah dikembangkan oleh beberapa orang untuk mencari keterhubungan kata pada kamus elektronik *WordNet*. Dengan menggunakan pendekatan korpus dan pendekatan kamus didapatkan hasil yang cukup baik dari masing-masing algoritma yang digunakan.

Penelitian terkait lain yang menggunakan pendekatan kamus dan membuat daftar kata adalah *SentiWordNet*. *SentiWordNet* (Esuli & Sebastiani, 2006.)merupakan kamus kata yang berisi nilai sentimen dari kata. Kata-kata dari *SentiWordNet* merupakan kata-kata yang diambil dari basis data leksikal *WordNet* yang dibuat dalam Bahasa Inggris. *SentiWordNet* umumnya digunakan untuk melakukan klasifikasi sentimen dan *opinion mining* dari kalimat, dokumen ataupun media lainnya. Berikut contoh leksikal dari *SentiWordNet* (Tabel II-6):

Tabel II-6. Contoh leksikal SentiWordNet

a	01657056	0	0	basined#1	enclosed in a basin
n	15246775	0	0	hot_spell#1	a spell of hot weather

Keterangan contoh leksikal dari *SentiWordNet* :

Kolom 1 : Jenis Kata

Kolom 2 : ID *WordNet*

Kolom 3 : Nilai polaritas positif

Kolom 4 : Nilai polaritas negatif

Kolom 5 : Kata

Kolom 6 : Deskripsi kata

Pemberian anotasi secara otomatis pada kamus *SentiWordNet* 3.0 melalui 2 proses (Baccianella, Esuli, & Sebastiani, 2010), yaitu pembelajaran *semi-supervised* dan *random walk*.

BAB III.

DESKRIPSI SOLUSI

Pada bab III ini, akan dilakukan analisis dari perbandingan pendekatan yang digunakan pada tahap eksperimen, analisis masalah dan analisis setiap pendekatan. Analisis ini dilakukan untuk memastikan algoritma yang digunakan relevan terhadap pembuatan daftar kata kunci dan dapat menghasilkan akurasi yang cukup baik. Evaluasi untuk menentukan pendekatan dan pembobotan yang terbaik akan dilakukan melalui akurasi hasil klasifikasi beberapa teks opini masyarakat dengan menggunakan model klasifikasi yang dihasilkan dari setiap pendekatan.

III.1. Analisis Perbandingan Pendekatan

Pendekatan korpus dan *hybrid* (pendekatan korpus dan pendekatan kamus) akan menghasilkan daftar kata yang berbeda. Perbedaan daftar kata disebabkan oleh penambahan sumber penelusuran kata. Pendekatan kamus mempunyai kemungkinan untuk menghasilkan akurasi yang tinggi, namun diperlukan kata-kata sumber untuk dapat menghasilkan daftar yang lebih baik. Dengan pertimbangan tersebut, pendekatan kamus dilakukan untuk meningkatkan kualitas dari pendekatan korpus dan disebut sebagai pendekatan *hybrid*. Pendekatan *hybrid* yang menggunakan pendekatan korpus dan kamus diharapkan dapat menangani masalah data yang tidak dapat diklasifikasikan ataupun akurasi hasil yang kecil dari pendekatan korpus.

III.2. Analisis masalah

Pesan dari *Twitter* memiliki beberapa masalah yang dapat mengurangi akurasi klasifikasi, diantaranya adalah :

1. Struktur kalimat yang kadang tidak sesuai struktur kalimat yang benar
2. Penggunaan singkatan, slank atau kata-kata yang sesuai dengan EYD Bahasa Indonesia
3. Istilah-istilah yang sering muncul pada *Twitter*, seperti RT.

Permasalahan ini dapat dilihat pada contoh pesan *Twitter* berikut (Tabel III-1):

Tabel III-1. Contoh Pesan Twitter

Pak, keren deh Pak liat anak2 SD Ujungberung pake kebaya n ikat kepala batik.. Top @ridwankamil
2 pompa sdg difungsikan. Mudah2an secepatnya RT @rambutkulimis: Pa ruas jalan gedebage cibiru banjir mudah2an dpt segera ter atasi ya pak

Sehingga untuk dapat melakukan klasifikasi pada pesan tersebut perlu dilakukan pemrosesan awal. Pesan dari *Twitter* tersebut kurang lengkap untuk dijadikan data latih, sehingga akan digunakan juga pesan dari lapor.ukp.go.id dan artikel berita untuk menjadi data latih pembentukan daftar kata kunci. Contoh Pesan yang diambil dari Lapor.ukp dapat dilihat sebagai berikut (Tabel III-2):

Tabel III-2. Contoh Pesan Lapor.ukp.go.id

Pesan	Kategori
Penjelasan tentang Kartu Jakarta Pintar **Kepada Yth. Pemprov DKI Jakarta. Saya ingin bertanya bagaimana prosedur pembuatan Kartu Jakarta Pintar?. Mohon Penjelasannya. Terimakasih karna sudah menyempatkan membaca pesan dari saya.**	Pendidikan
Penanganan Pohon Rindang di Gunung Sahari **Kepada Yth. Pemerintah Provinsi DKI Jakarta, saya ingin melaporkan bahwa pohon di depan Bank BCA Gunung Sahari, No. 45 sudah lebat sekali. Dimohon tindakannya untuk dirapihkan karena ditakutkan dapat membahayakan pengguna jalan lagi bila cuaca buruk. Terima kasih.**	Lingkungan hidup dan Penanganan Bencana

Sedangkan contoh artikel berita yang diambil dari infobandung.co.id adalah sebagai berikut (Tabel III-3):

Tabel III-3. Contoh Artikel Berita

Masyarakat Masih Banyak Yang mengeluh Soal Pelayanan SJSN Dan BPJS.

Bandung,Infobandung.co.id – Aksi demo penolakan UU SJSN dan BPJS mengundang komentar Atmadja sekretaris Komisi D DPRD Kota Bandung, dia mengatakan Demo mahasiswa tersebut cukup bagus namun anggapan bahwa sistem yang diberlakukan untuk SJSN dan BPJSN merupakan pemalakan bagi rakyat itu kurang baik.

"Demo mahasiswa ini cukup bagus terus terang saja banyak masyarakat yang mengeluh masalah pelayanan BPJS tersebut karena terlalu banyak masyarakat yang mendaftar ke BPJS namun kalau oleh pihak mahasiswa yang berdemo tadi menyatakan, bahwa masalah ini dikatakan pemerasan atau pemalakan,saya pikir terlalu kasar bila istilah itu digunakan" ungkap Atmadja kepada wartawan di Kantor DPRD Kota Bandung, Jalan Aceh. Rabu (8/1/2013).

Dirinya juga mengatakan secara garis besar tujuan pemerintah sudah cukup baik yaitu memberikan pelayanan kesehatan bagi seluruh rakyat Indonesia.

"Tujuan diberlakukannya sistem jaminan sosial nasional (SJSN) oleh pemerintah sudah cukup baik yaitu memberikan pelayanan bagi seluruh masyarakat, tinggal praktek nya saja seperti apa", paparnya. (Ods)

Pesan dari Lapor.ukp.go.id dan artikel berita dari infobandung.co.id mempunyai struktur kalimat dan penggunaan kata-kata yang lebih baik bila dibandingkan dengan pesan *Twitter*, selain itu jumlah kata yang dapat ditangkap lebih banyak bila dibandingkan dengan pesan dari *Twitter*. Dengan demikian, sumber data latih yang digunakan untuk membuat daftar kata berasal dari Lapor.ukp.go.id dan artikel berita yang ditulis secara elektronik. Sumber data uji yang digunakan berasal dari pesan *Twitter* untuk menyesuaikan dengan lingkungan tujuan pembuatan daftar kata kunci. Pengambilan data dari *Twitter* dilakukan dengan bantuan oleh perusahaan NoLimit, data dari perusahaan NoLimit mengandung nama pengirim tweet, konten, dan tanggal dari pesan *Twitter*. Untuk data dari lapor.ukp.go.id, didapatkan langsung beberapa pesan yang sudah diklasifikasikan, sehingga tidak perlu melakukan pengambilan data secara manual. Sedangkan untuk artikel berita,

data didapatkan secara manual dengan melakukan pengambilan masing-masing berita yang relevan dengan pemerintahan kota Bandung.

Terdapat 21 kategori pelabelan untuk data latih, data uji dan setiap kata yang masuk ke dalam daftar kata kunci. Kategori-kategori tersebut ditentukan dari organisasi yang mengurus kota Bandung yang telah didefinisikan pada (Laksana, 2013). Kategori yang termasuk dapat dilihat pada Gambar III-1.

1. Dinas Bina Marga dan Pengairan	8. Dinas Pemuda dan Olahraga	15. Dinas Tata Ruang dan Cipta Karya
2. Dinas Kebakaran	9. Dinas Pengelolaan Keuangan dan Aset Daerah	16. PD Air Minum Tirtawening
3. Dinas Kebudayaan dan Pariwisata	10. Dinas Koperasi, UKM, dan Perindustrian Perdagangan	17. PD Bank Perkreditan Rakyat Kota Bandung
4. Dinas Kependudukan dan Pencatatan Sipil	11. Dinas Perhubungan	18. PD Kebersihan
5. Dinas Kesehatan	12. Dinas Pertanian dan Ketahanan Pangan	19. PD Pasar Bermartabat
6. Dinas Komunikasi dan Informatika	13. Dinas Sosial	20. Dinas Pemakaman dan Pertamanan
7. Dinas Pendidikan	14. Dinas Tenaga Kerja	21. Dinas Pelayanan Pajak Kota Bandung

Gambar III-1. Kategori data (Laksana, 2013)

III.3. Pendekatan korpus

Pendekatan korpus merupakan pendekatan dengan menghitung keterhubungan antar kata ataupun besar pengaruh kata pada data latih. Secara umum, pendekatan korpus dibagi menjadi 2 metode, yaitu metode *clustering* dan metode *Association Rule Learning* (ARL). Kedua metode tersebut akan melalui 2 tahap utama, yaitu tahap persiapan data dan tahap yang dilakukan sistem. Tahap persiapan data dari kedua metode mencakup:

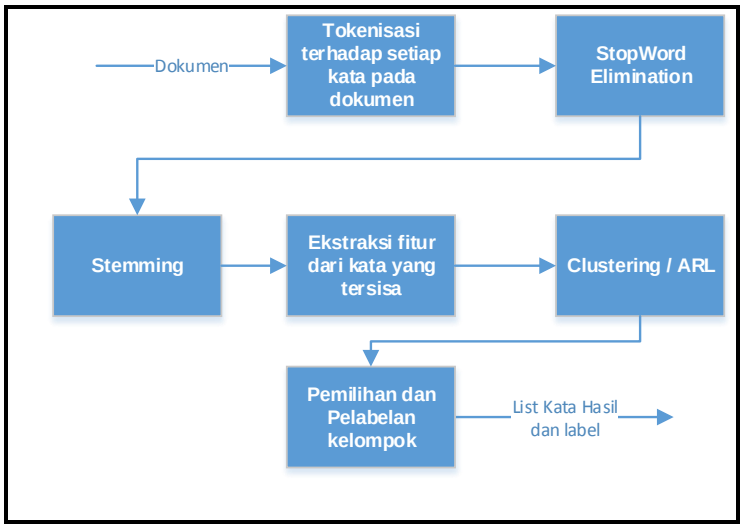
1. pengumpulan data latih yang diambil dari tulisan-tulisan pada lapor.ukp.go.id dan artikel berita dengan jumlah latih $\pm$ 100 tulisan dan artikel
2. Pelabelan dokumen secara manual terhadap data latih sesuai dengan kategori dari dokumen yang telah dikumpulkan
3. Pemrosesan dokumen dengan kategori dokumen yang sama

Tahap yang dilakukan sistem pada pendekatan korpus dengan metode *clustering* mencakup:

1. Penghapusan kata-kata yang sangat umum digunakan pada kalimat agar tidak masuk ke dalam daftar kata yang mungkin berhubungan (*StopWord*)
2. Pengubahan menjadi kata dasar dari setiap kata yang diproses
3. Ekstraksi fitur TFxIDF, *Information gain, Mutual Information,*dan X^2 *Statistic* dari setiap kata yang tersisa. Untuk fitur TFxIDF, dilakukan penjumlahan nilai TFxIDF keseluruhan dari dokumen-dokumen dengan kategori yang sama.
4. Penerapan metode *clustering* atau ARL
5. Pemilihan kelompok kata yang sesuai dengan kriteria
6. Setiap kata dari kelompok yang diambil akan disimpan dan diberikan label kategori dari dokumen tersebut. Proses pada tahap otomatis ini akan diulang hingga hasil daftar kata dari semua kategori telah tersimpan

Tahap sistem dari pendekatan korpus dengan metode ARL berbeda pada langkah ekstraksi fitur, dimana fitur yang diambil hanya TFxIDF. Perbedaan lain antara ARL dengan *clustering* adalah perlunya ditetapkan nilai batas kata dianggap sering muncul pada ARL. Kumpulan kata terakhir yang masih berada di atas nilai batas akan diambil untuk menjadi daftar kata. Dengan demikian daftar kata yang didapatkan merupakan kumpulan kata yang paling banyak muncul pada dokumen-dokumen dengan kategori yang sama. Jumlah kata yang dihasilkan berkemungkinan lebih sedikit bila dibandingkan dengan metode *clustering*, namun akurasi klasifikasi mempunyai kemungkinan lebih akurat.

Ilustrasi dari tahap yang dilakukan sistem dapat dilihat pada Gambar III-2.



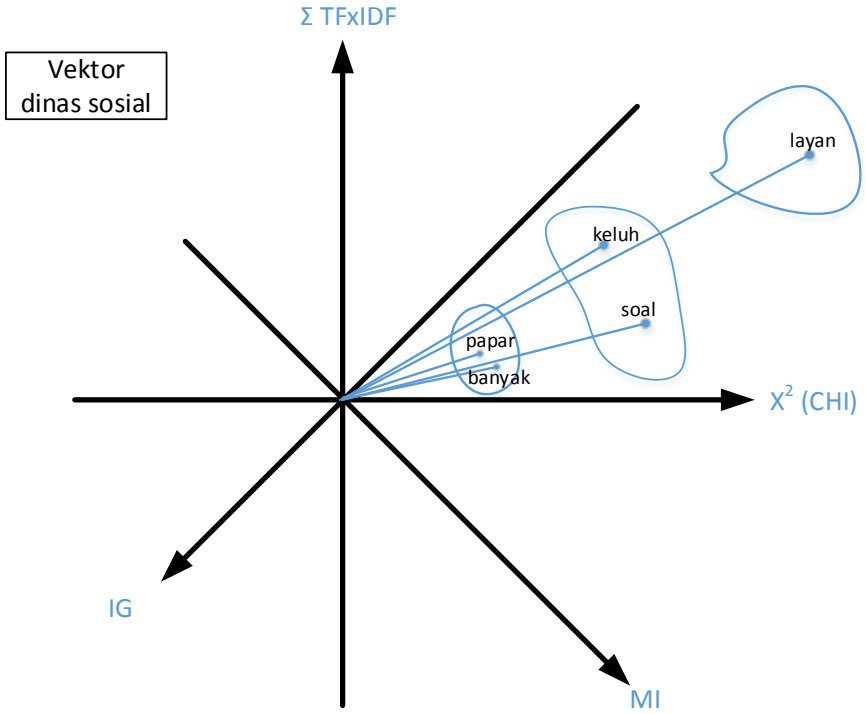
Gambar III-2. tahap yang dilakukan sistem pada pendekatan korpus

Contoh penerapan tahap sistem pada pemrosesan dokumen dari Tabel III-3 dapat dilihat pada tabel berikut (Tabel III-4):

Tabel III-4. Contoh penerapan tahap sistem pendekatan korpus

Tahap	**Input**	**Output**
Tokenisasi	Dokumen Tabel III-3	”masyarakat” ”banyak” ”yang” “mengeluh” “soal” “pelayanan” .... “paparnya”
StopWord Elimination	”masyarakat” ”banyak” ”yang” “mengeluh” “soal” “pelayanan”	”masyarakat” ”banyak” “mengeluh” “soal” “pelayanan” .... “paparnya”

Tahap	Input	Output
	“paparnya”	
Stemming	”masyarakat” ”banyak” “mengeluh” “soal” “pelayanan” .... “paparnya”	”masyarakat” ”banyak” “keluh” “soal” “layan” .... “papar”
Ekstraksi Fitur	”masyarakat” ”banyak” “keluh” “soal” “layan” .... “papar”	$\sum TFxIDF$(“layan”,dinas sosial)=1,3 IG(“layan”,dinas sosial)=1 .... MI(“papar”,dinas sosial)=0,2 X^2(“papar”,dinas kebersihan)=0,4
Clustering / ARL	$\sum TFxIDF$(“layan”,dinas sosial)=1,3 IG(“layan”,dinas sosial)=1 .... MI(“papar”,dinas sosial)=0,2 X^2(“papar”,dinas kebersihan)=0,4	Kelompok 1 : {layan,....} Kelompok 2 : {masyarakat, keluh ,soal,...} Kelompok 3 : {papar, banyak,...}
Pemilihan dan pelabelan kelompok	Kelompok 1 : {layan,....} Kelompok 2 : {masyarakat, keluh ,soal,...} Kelompok 3 : {papar, banyak,...}	{layan,....} =>Dinas sosial



Gambar III-3. Ilustrasi proses clustering untuk kategori dinas sosial

Pada proses *clustering*, kata-kata tersebut akan dikelompokkan sesuai kedekatan nilai fiturnya. Ilustrasi dari proses *clustering* dapat dilihat pada Gambar III-3. Proses *clustering* dapat direpresentasikan dalam bentuk vektor 4 dimensi dimana semua fitur menjadi dimensi dari setiap kata. Namun, untuk kategori yang berbeda, maka ruang vektor dari fitur-fitur tersebut juga akan berbeda. Dengan demikian, jumlah ruang vektor yang perlu dibuat adalah sebanyak kategori yang diinginkan.

III.4. Pendekatan *Hybrid*

Pendekatan *hybrid* merupakan pendekatan gabungan antara pendekatan korpus dan pendekatan kamus. Pendekatan kamus menggunakan metode penelusuran kamus untuk mendapatkan hubungan antar kata. Oleh karena itu, Pendekatan *hybrid* akan mengambil kata-kata yang mungkin berhubungan dengan menggunakan kata-kata yang didapatkan dari pendekatan korpus. Pendekatan ini menggunakan kamus elektronik *WordNet* untuk mendapatkan kata-kata lain. Kata-kata yang didefinisikan berupa kata-kata dengan *sense* / konsep yang sama dengan kata dari kategori departemen yang bersangkutan. Kata-kata yang telah didefinisikan pada

WordNet ini diharapkan dapat menambah kosakata pada kamus yang telah terbentuk dari pendekatan secara korpus. Pendekatan secara kamus ini akan menggunakan kamus *WordNet* dalam bahasa Indonesia (Noor, Sapuan, & Bond, 2011) / kamus *SentiWordNet* yang telah diterjemahkan ke dalam bahasa Indonesia (Lunando, 2013). Selain menggunakan kata-kata yang berada dalam *sense* yang sama, akan digunakan juga perhitungan *word similarity* dengan pendekatan kamus yang telah ditemukan dari referensi.

Langkah dari pendekatan *hybrid* yang dilakukan secara garis besar adalah sebagai berikut:

1. Pemrosesan masing-masing kata hasil pendekatan korpus
2. Dilakukan pencarian kata yang mempunyai nilai *word similarity* yang tinggi (melebihi *threshold* yang telah ditetapkan secara manual)
3. Mengumpulkan kata-kata hasil dan memberi label kategori sesuai kategori kata sumber
4. Melakukan evaluasi kata-kata hasil (dilakukan penghapusan kata-kata yang tidak berhubungan)
5. Membuat daftar kata kunci dari kata-kata hasil.

Contoh penerapan langkah di atas dapat dilihat pada tabel berikut (Tabel III-5):

Tabel III-5. Contoh penerapan langkah pendekatan hybrid

Langkah	**Input**	**Output**
Pemrosesan masing-masing kata	{layan,...}=>Dinas sosial	Layan
Pencarian kata	layan	Bantu -> 0,9 Jamu -> 0,8
Mengumpulkan kata dan pemberian label	Bantu -> 0,9 Jamu -> 0,8	Bantu => Dinas sosial Jamu => Dinas sosial
Evaluasi kata	Bantu => Dinas sosial Jamu => Dinas sosial	Bantu => Dinas sosial

Langkah	Input	Output
Membuat daftar kata kunci	Bantu => Dinas sosial	{layan, bantu,... } =>Dinas sosial ..... {sampah,kotoran,...} =>Dinas kebersihan

III.5. Evaluasi

Untuk melakukan evaluasi, dilakukan klasifikasi otoritas dari data yang diambil dari pesan *Twitter*. Evaluasi dilakukan dengan mengimplementasikan daftar kata kunci dan menganalisis akurasi hasil implementasi terhadap data uji yang telah disiapkan. Hasil implementasi dibandingkan dan diambil pendekatan dan metode yang menghasilkan akurasi klasifikasi otoritas terbaik dan tingkat yang rendah untuk data yang tidak dapat diklasifikasi. Daftar kata kunci hasil akan dievaluasi lebih lanjut untuk meningkatkan nilai akurasinya dan menghasilkan daftar kata kunci yang siap digunakan.

DAFTAR REFERENSI

Abbas, O. (2008). Comparisons Between Data Clustering Algorithms. *The International Arab Journal of Information Technology, 5*, 320--325.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (hal. 207-216). ACM.

Antonie, M.-L., & Zai'ane, O. R. (2002). Text Document Categorization by Term Association. *IEEE International Conference* (hal. 19-26). IEEE.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Language Resources and Evaluation* (hal. 2200-2204). Genova: IT.

Church, K. W., & Hanks, P. (1990, mar). Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist., 16*(March 1990), 22-29.

Dash, M., Choi, K., Scheuermann, P., & Liu, H. (2002). Feature selection for clustering-a filter solution. *IEEE International Conference* (hal. 115-122). IEEE.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS, 41*, 391--407.

Esuli, A., & Sebastiani, F. (2006.). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *Language Resources and Evaluation* (hal. 417–422). Genova: IT.

Jain, A., Murty, M., & Flynn, P. (1999, sep). Data Clustering: A Review. *ACM Comput. Surv., 31*, 264--323. doi:10.1145/331499.331504

Kaufman, L., & Rousseeuw, P. (1990). Finding Groups in Data. Dalam *An Introduction to Cluster Analysis.* Canada: John Wiley & Sons,Inc.

Laksana, J. (2013). *KLASIFIKASI OTORITAS TEKS PENDEK JEJARING SOSIAL TWITTER UNTUK PEMERINTAH KOTA BANDUNG.* Bandung.

Liu, T., Liu, S., Chen, Z., & Ma, W.-Y. (2003). An evaluation on feature selection for text clustering. *ICML*, (hal. 488--495).

Lunando, E. (2013). *Analisis Sentimen Mengandung Ironi dan Menggunakan Sentiwordnet.* Bandung.

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. *AAAI*, (hal. 775-780).

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill Science/Engineering/Math;.

Noor, N. H., Sapuan, S., & Bond, F. (2011). Creating the Open Wordnet Bahasa. *25th Pacific Asia Conference on Language, Information*, (hal. 258–267). Singapore.

Penã, J., Lozano, J., & Larrañaga, P. (1999). An empirical comparison of four initialization methods for the K-Means algorithm. *Pattern Recognition Lett. 20*, 1027-1040.

Rajaraman, A., Leskovec, J., & Ullman, J. D. (2011). Data Mining. Dalam A. Rajaraman, J. Leskovec, & J. D. Ullman, *Mining of Massive Datasets* (hal. 1–17).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*, 513 - 523. doi:http://dx.doi.org/10.1016/0306-4573(88)90021-0

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *ICML*, (hal. 412--420).