USULAN TUGAS AKHIR

IDENTITAS PENGUSUL

Nama : **Andreas Daniel Arifin**
NRP : **5108 100 132**
Dosen Wali : **Prof. Ir. Handayani Tjandrasa, M.Sc, Ph.D**

JUDUL TUGAS AKHIR

Implementasi Algoritma K-Nearest Neighbour berdasarkan Algoritma One Pass Clustering untuk Kategorisasi Teks

Implementation of K-Nearest Neighbour Algorithm based on One Pass Clustering Algorithm for Text Categorization

1 PENDAHULUAN

1.1 LATAR BELAKANG

Kategorisasi teks (atau klasifikasi teks) adalah suatu proses yang mengelompokkan suatu teks ke dalam suatu kategori tertentu. Secara teknis, tugas kategorisasi teks adalah memetakan fungsi tujuan: $D \times C \rightarrow \{T, F\}$, yang mana D adalah domain dokumen dan C adalah himpunan kategori yang telah ditentukan. Nilai T diberikan apabila suatu dokumen d_i termasuk ke dalam kategori c_j. Jika sebaliknya, diberikan nilai F.

Kategorisasi teks merupakan solusi yang tepat untuk mengelola informasi yang saat ini berkembang dengan sangat cepat dan melimpah. Kategorisasi teks membuat pengelolaan informasi tersebut menjadi efektif dan efisien. Dengan menggunakan kategorisasi teks, kita dapat melakukan penyaringan terhadap email *spam*, melakukan penggalian opini (*opinion mining*), dan analisis sentimen. Algoritma kategorisasi teks saat ini telah banyak berkembang, antara lain: support vector machines (SVM), naive bayessian (NB), pohon keputusan, k-nearest neighbour (KNN), dan lainnya. Dari berbagai macam algoritma yang telah dikembangkan tersebut, KNN dan SVM telah diakui lebih handal dibandingkan dengan algoritma yang lainnya.

Algoritma KNN sendiri adalah suatu algoritma yang sederhana, namun cukup efektif dalam melakukan kategorisasi teks. Tambahan pula, proses klasifikasi dari KNN mudah untuk direpresentasikan dibandingkan dengan algoritma klasifikasi lain, seperti: SVM dan Artificial Neural Networks (ANN). Namun dalam keunggulan tersebut, KNN mempunyai beberapa kekurangan. Permasalahan mendasar adalah masalah pemilihan nilai *k* yang mudah dipengaruhi oleh *noise*. Selain itu, algoritma KNN membutuhkan alokasi memori yang besar karena tidak membangun model klasifikasi dalam prosesnya. Dengan adanya kekurangan

tersebut, maka KNN bisa menjadi tidak sesuai diterapkan dengan kondisi data yang sangat melimpah dan terus berubah sekarang ini.

Untuk menyelesaikan masalah tersebut, Jiang, Pang, Wu, dan Kuan (2011) mengajukan suatu metode untuk mengatasi hal tersebut. Metode tersebut bekerja dengan mengelempokkan data teks yang akan dikategorisasi dengan KNN terlebih dahulu. Pengelompokan data teks tersebut dapat dilakukan dengan teknik *clustering*. Dengan mengelompokkan terlebih dahulu data teks, maka dapat dikatakan sudah melakukan pembangunan model klasifikasi. Oleh karena itu, dalam tugas akhir ini akan diimplementasikan sistem kategorisasi teks dengan algoritma KNN berdasarkan algoritma one pass clustering yang berguna untuk membangun model klasifikasi.

1.2 TUJUAN

Tugas akhir ini bertujuan untuk melakukan kategorisasi teks dengan menggunakan algoritma k-nearest neighbour berdasarkan algoritma one pass clustering.

1.3 MANFAAT

Dengan mengimplementasikan metode ini, diharapkan sistem kategorisasi teks, seperti: penyaringan email *spam*, pengelompokan artikel pada situs berita *online*, dsb, yang menggunakan algoritma KNN dapat berjalan lebih optimal dan lebih efisien.

1.4 RUMUSAN MASALAH

Untuk mengimplementasikan metode kategorisasi ini, maka dirumuskan permasalahan:

1. Bagaimanakah melakukan *preprocessing* untuk mengubah dokumen menjadi representasi vektor?
2. Bagaimanakah membuat model klasifikasi dengan algoritma one pass clustering?
3. Bagaimanakah melakukan kategorisasi teks menggunakan algoritma k-nearest neighbour berdasarkan model klasifikasi yang telah dibangun dengan one pass clustering?
4. Bagaimanakah mengevaluasi model yang telah dibangun?

1.5 BATASAN MASALAH

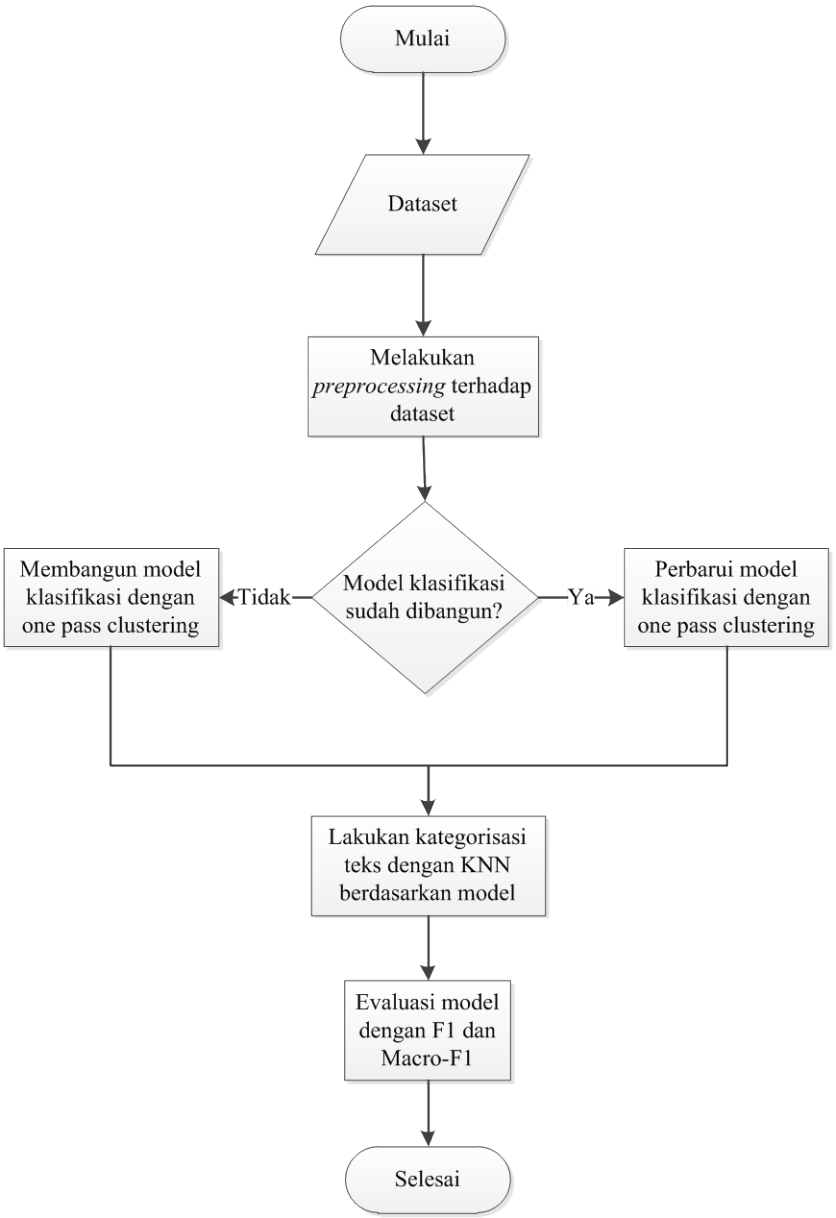
Asumsi dan ruang lingkup permasalahan yang dirumuskan dalam tugas akhir ini:

1. Dalam tahap *preprocessing* dokumen menjadi representasi vektor, *stemmer* yang digunakan adalah Porter stemmer.
2. *Dataset* yang digunakan dalam uji coba adalah Reuters-21578[1].
3. Sistem perangkat lunak dibangun dengan bahasa pemrograman Java SE.

2 RINGKASAN TUGAS AKHIR

Kategorisasi teks adalah salah satu solusi untuk mengatasi makin berkembangnya informasi. Salah satu algoritma kategorisasi teks yang cukup baik adalah k-nearest neighbour (KNN). Algoritma KNN tradisional memiliki beberapa kekurangan. Untuk mengatasi kekurangan tersebut, telah dirumuskan perbaikan dari algoritma KNN tradisional. Algoritma KNN yang diperbaiki tersebut ditambahkan kemampuan untuk membangun model klasifikasi lewat algoritma one pass clustering.

[1] http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection



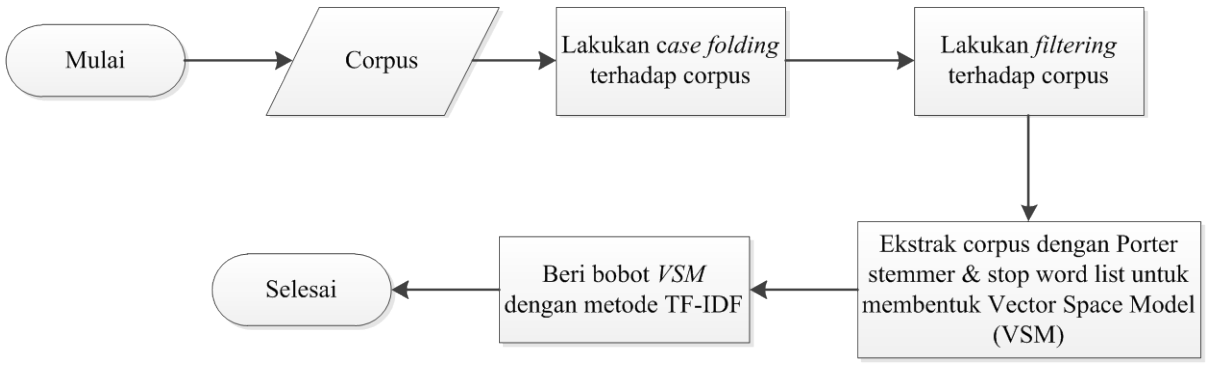
Gambar 2.1 Diagram alir sistem secara umum

Untuk mengimplementasikan kategorisasi teks dengan algoritma KNN berdasarkan one pass clustering, ada beberapa tahap yang harus dilalui. Gambar 2.1 menggambarkan aliran sistem secara umum. Tahap-tahap tersebut dijelaskan sebagai berikut:

1. Tahap *preprocessing*. Tahap ini diawali dengan mengubah semua karakter dalam satu dokumen menjadi huruf kecil. Kemudian, dokumen akan dilakukan *filtering* untuk mengeliminasi karakter-karakter yang tidak dibutuhkan. Setelah tahap *filtering*, dokumen akan diekstrak untuk mendapatkan kata-kata (*term*) dari setiap dokumen. Kata-kata yang didapat tersebut akan dilakukan proses *stemming* menggunakan Porter stemmer. Selanjutnya, kumpulan kata tersebut akan diberi bobot dengan metode tf-idf.
2. Tahap pembangunan model klasifikasi. Untuk mendapatkan model klasifikasi, digunakan algoritma one pass clustering. Dengan algoritma one pass clustering ini, model klasifikasi dapat dengan mudah diperbarui dengan penambahan data pelatihan.

3. Tahap kategorisasi/klasifikasi teks. Setelah model klasifikasi dibangun, model tersebut dapat digunakan untuk melakukan kategorisasi terhadap teks. Kategorisasi teks di sini menggunakan algoritma k-nearest neighbour (KNN).
4. Evaluasi model. Model yang telah dibangun, dapat dilakukan evaluasi untuk mengetahui sejauh mana kinerja model tersebut. Evaluasi model klasifikasi untuk kategorisasi teks menggunakan metode F_1 dan *Macro-*F_1.

Sistem ini diawali dengan mengolah terlebih dahulu *dataset* yang akan digunakan, baik untuk data pelatihan maupun data pengujian. *Dataset* yang digunakan adalah data Reuters-21578 yang sudah dianggap sebagai data standar untuk kategorisasi teks. *Dataset* ini merupakan suatu *corpus* yang terdiri dari 21578 dokumen dan 135 kategori. Kemudian, *dataset* ini akan diolah dalam tahap *preprocessing*.



Gambar 2.2 Diagram alir tahap *preprocessing*

Tahap *preprocessing* diawali dengan melakukan *case folding* terhadap teks dalam dokumen. *Case folding* ini akan mengubah semua huruf dalam teks menjadi huruf kecil. Setelah itu, dokumen akan dilakukan proses *filtering*, yaitu dengan membuang semua karakter yang tidak bersifat signifikan, seperti: tanda baca dan angka. Dokumen kemudian akan diekstrak untuk mendapatkan kata-kata (*term*) dari tiap dokumen. Kata-kata yang sudah didapat tersebut akan dilakukan proses *stemming* menggunakan Porter stemmer. Porter stemmer adalah *stemmer* standar dan dipakai secara luas dalam dokumen berbahasa inggris. *Stemmer* adalah suatu alat untuk mengubah morfologi suatu kata menjadi kata dasarnya, terutama akhiran infleksi dan bentuk morfologi yang umum. Selain dilakukan *stemming*, kata itu selanjutnya akan dieliminasi lebih lanjut berdasarkan *stop word list*. *Stop word list* merupakan kumpulan kata-kata yang sudah umum ditemukan dalam suatu dokumen sehingga memiliki nilai semantik yang kecil. Secara keseluruhan, tahap preprocessing dapat dilihat pada gambar 2.2.

Kumpulan kata-kata dasar yang sudah diekstrak akan direpresentasikan ke dalam bentuk Vector Space Model (VSM). VSM adalah suatu cara representasi dan *indexing* suatu dokumen dengan merepresentasikannya ke dalam suatu vektor yang berisi bobot dari kata-kata yang muncul di dalam dokumen. Bobot yang akan digunakan pada VSM tersebut adalah bobot tf-idf. Metode pembobotan tf-idf adalah suatu metode pembobotan yang merupakan hasil perkalian dari bobot tf (*term frequency*) dan bobot idf (*inverse document fequency*) dari suatu kata. Bobot tf adalah suatu bobot yang menyatakan frekuensi kemunculan kata dari suatu dokumen. Sedangkan, bobot idf adalah suatu bobot yang menyatakan inversi dari banyaknya dokumen yang mengandung suatu kata tertentu. Rumus tf-idf dinyatakan dalam persamaan berikut:

$$w_{t,d} = (1 + \log tf_{t,d}) \times \log_{10} \frac{N}{df_t} \quad (1)$$

Dalam persamaan (1), *t* merupakan term atau kata, dan *d* merupakan dokumen. Notasi *N* adalah total semua dokumen yang ada.

Data yang sudah direpresentasikan ke dalam bentuk VSM akan dijadikan acuan untuk pembangunan model klasifikasi. Pembangunan model klasifikasi ini menggunakan algoritma one pass clustering. Algoritma clustering ini memiliki keunggulan dari segi waktu daripada algoritma clustering tradisional, seperti k-mean, karena algoritma hanya membaca data satu kali selama proses *clustering*. Tahap-tahap proses one pass clustering dijelaskan sebagai berikut yang juga tergambar dalam gambar 2.3.

1. Buat suatu himpunan kosong untuk menampung *cluster*, m_0.
2. Baca teks *p* dari VSM. Bentuk *cluster* baru dengan anggota *p*, dan label dari *cluster* adalah label dari dari teks *p*.
3. Jika tidak ada teks yang bisa dibaca di VSM, maka menuju ke nomor 6. Sebaliknya, baca teks baru *p* dari VSM, hitung kemiripan teks *p* dengan semua *cluster C* di m_0 menggunakan fungsi *cosine*. Ambil *clutser* yang terdekat dengan *p* atau yang nilai kemiripannya paling besar, $sim(p,C^*) \geq sim(p,C)$.
4. Jika nilai $sim(p,C^*) < r$ atau label teks *p* berbeda dengan label *cluster C**, maka menuju ke 2. Variabel *r* adalah variabel *threshold* yang akan dijelaskan kemudian.
5. Gabungkan teks p ke dalam *cluster C** dan perbarui bobot dari tiap kata di *C**. Kemudian, menuju ke 3.
6. Hentikan proses *clustering*. Ambil hasil clustering, $m_0 = \{C_1, C_2, C_3, \ldots, C_4\}$, tiap *cluster* dalam m_0 berisi kumpulan kata berbobot dan label *cluster*. Himpunan m_0 adalah model klasifikasi.

Dalam proses *clustering* di atas, terdapat proses pembaruan bobot. Strategi untuk pembaruan bobot yang tercantum pada langkah 5 adalah sebagai berikut.

$$w_{C_i}^{i+1}(t) = \frac{w_{c_i}^{i}(t) \times |c_i| + w(t)_p}{|c_i| + 1}, \quad (2)$$

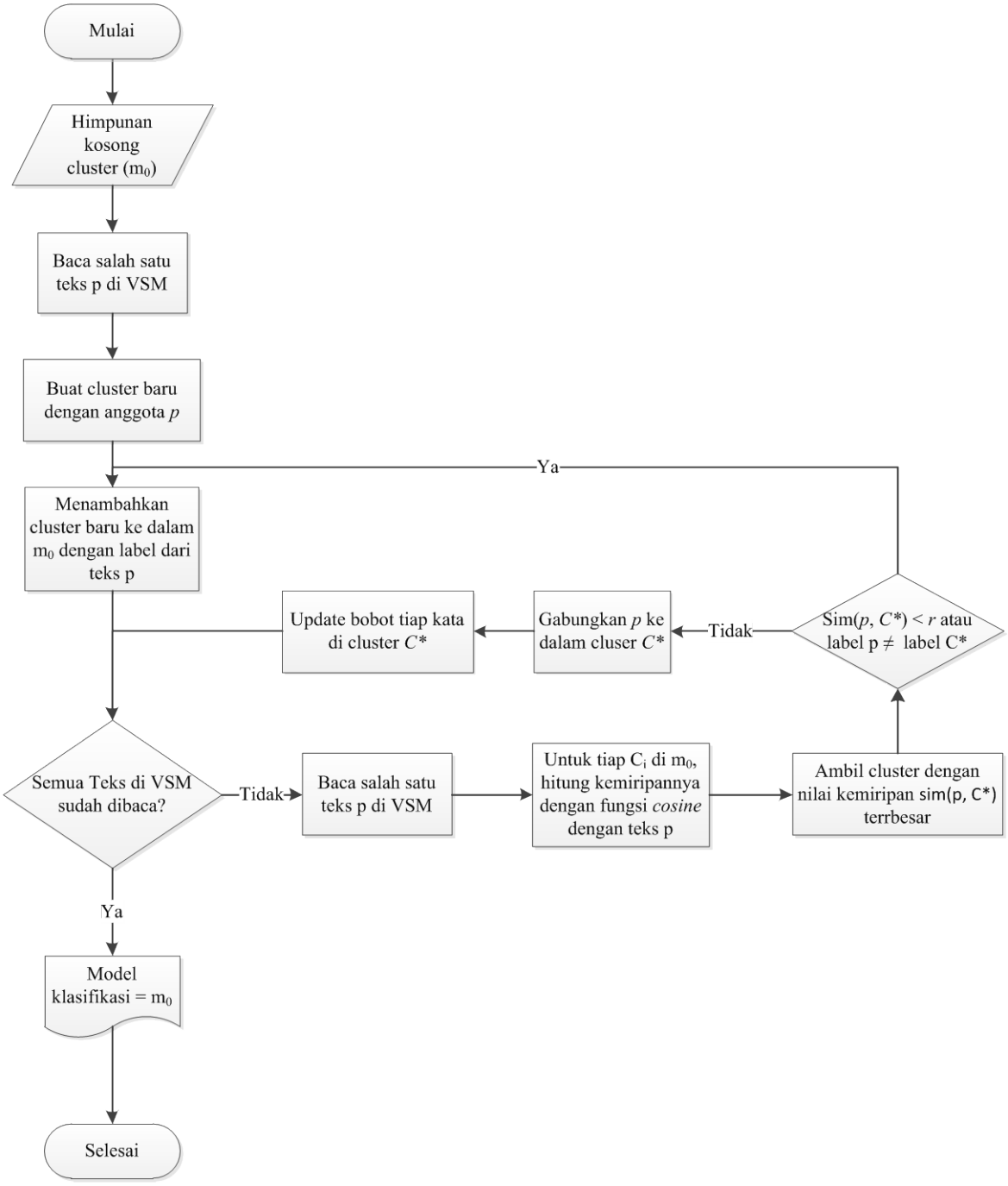
di mana $w_{C_i}^{i+1}(t)$ adalah bobot baru dari kata *t* dari *cluster* c_i. $w_{c_i}^{i}(t)$ adalah bobot lama dari kata *t* dari *cluster* c_i. $w(t)_p$ adalah bobot kata *t* pada teks *p*. Dan, $|c_i|$ adalah jumlah teks pada *cluster* c_i.

Selain pembaruan bobot, dalam proses *clustering* tersebut juga terdapat nilai *r*. Nilai r adalah nilai *threshold* yang dapat mempengaruhi efisiensi waktu dan kualitas dari *clustering*. Nilai r tersebut dapat diperoleh dengan menggunakan teknik sampling sebagai berikut.

1. Pilih secara acak N_0 pasang teks dari *corpus*.
2. Hitung kemiripan (*similarity*) tiap pasang teks.
3. Hitung nilai rata-rata kemiripan yang didapat dari tahap 2 (*ex*).
4. Tentukan nilai *r* dari persamaan ε *x ex*, di mana $\varepsilon \geq 1$.

Model klasifikasi tersebut dapat diperbarui (*update*) apabila terdapat data pelatihan baru yang ingin ditambahkan. Algoritma one pass clustering memungkinkan model klasifikasi dapat diperbarui dengan mudah karena sifatnya yang *incremental*. Proses pembaruan model dimulai dari tahap ke (3) dari algoritma one pass clustering.

Model klasifikasi yang telah didapat dari proses *clustering* sebelumnya akan digunakan sebagai acuan untuk melakukan kategorisasi teks. Proses kategorisasi teks ini akan menggunakan algoritma k-nearest neighbour (KNN). Secara sederhana, algoritma ini bekerja dengan membandingkan jarak data masukan dengan sejumlah *k* data pelatihan yang paling dekat. Namun, dengan adanya model klasifikasi dari one pass clustering, proses tersebut menjadi lebih sederhana dengan hanya membandingkan data masukan dengan *k cluster* yang telah terbentuk sebelumnya. Secara matematis, algoritma KNN berdasarkan one pass clustering dijabarkan dalam persamaan (3).



Gambar 2.3 Diagram alir proses pembangunan model dengan one pass clustering

$$f(x) = \text{argmax}_j \, ClusterScore(x, C_j) = \sum_{C_i \in KNN} sim(x, C_i) y(C_i, C_j), \quad (3)$$

di mana fungsi *f(x)* adalah fungsi yang mengembalikan label yang diberikan ke teks *x*. $ClusterScore(x, C_j)$ adalah skor kandidat perbandingan dokumen *x* dengan kumpulan kategori C_j. Fungsi $sim(x, C_i)$ mengembalikan nilai kemiripan (*similarity*) antara teks *x* dengan kategori C_i. Dan, $y(C_i, C_j)$ akan bernilai 1 apabila *cluster* C_i memiliki label C_j, dan bernilai 0 apabila sebaliknya. Secara sederhana, proses kategorisasi ini akan menghitung nilai kemiripan antara teks masukan dengan semua *cluster* dari model klasifikasi. Kemudian, akan dipilih *cluster* yang mempunyai nilai kemiripan paling besar dan label dari *cluster* tersebut akan menjadi label dari data teks masukan.

Terakhir, model klasifikasi tadi akan dievaluasi kinerjanya. Evaluasi perlu dilakukan untuk melihat sejauh mana kinerja dari model klasifikasi yang telah dibangun. Proses evaluasi ini akan menggunakan metode F_1 dan *Macro-*F_1. Metode evaluasi F_1 merupakan penerapan gabungan dari *recall* (*r*) dan *precision* (*p*) yang dinyatakan ke dalam persamaan berikut.

$$F_1 = \frac{2 \times recall \times precision}{recall + precision} \quad (4)$$

Sedangkan *Macro-*F_1 adalah nilai rata-rata dari nilai F_1 individu masing-masing kategori.

3 METODOLOGI PENGERJAAN TUGAS AKHIR

Metodologi yang akan dilakukan dalam tugas akhir ini memiliki beberapa tahapan, di antaranya sebagai berikut:

1. Studi Literatur

   Tahap awal ini dilakukan pencarian, pengumpulan, penyaringan, pembelajaran dan pemahaman literatur yang berhubungan dengan proses pengolahan dokumen teks, khususnya yang meliputi permasalahan mengenai *preprocessing*, *stemming*, dan representasi menggunakan Vector Space Model (VSM). Literatur tentang pembangunan model dengan one pass clustering, proses kategorisasi teks dengan KNN, dan evaluasi model dengan F_1 dan *Macro-*F_1 juga dibutuhkan dalam merancang sistem ini.

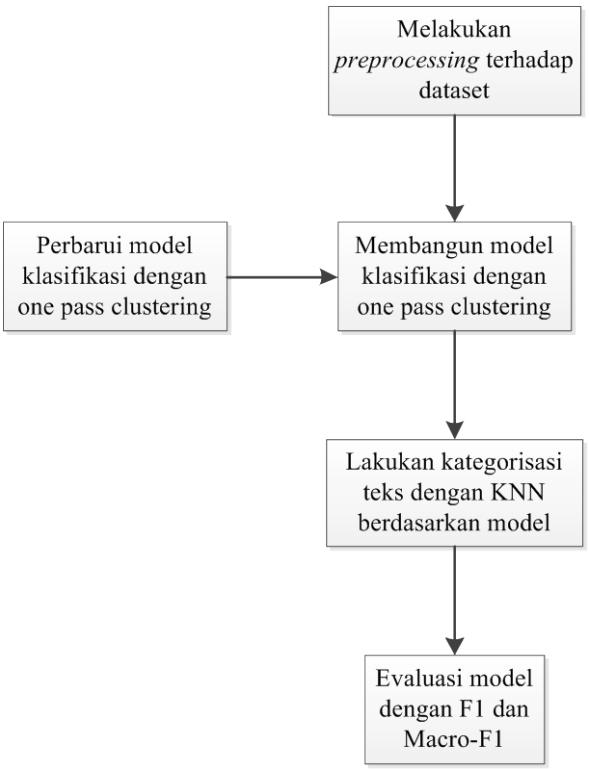
2. Implementasi

   Setelah melakukan studi literatur, tahap selanjutnya adalah melakukan implementasi dari bahan yang telah dipelajari sebelumnya. Pembangunan sistem ini akan dilakukan menggunakan bahasa pemrograman Java SE dan dibantu dengan DBMS MySQL. Tahap implementasi ini dibagi ke dalam 4 tahap yang dapat dilihat dalam gambar 3.1, yaitu: tahap implementasi *preprocessing*, tahap implementasi algoritma one pass clustering untuk pembangunan model klasifikasi dan pembaruan model klasifikasi, tahap implementasi kategorisasi teks dengan algoritma KNN, dan tahap implementasi evaluasi kinerja model klasifikasi.

   Pertama, diperlukan *dataset* untuk melakukan implementasi terhadap algoritma ini. *Dataset* yang digunakan adalah Reuters-21578. Selanjutnya, akan diimplementasikan proses *preprocessing* untuk mengolah *dataset*. Tahapan *Preprocessing* adalah sebagai berikut:

   1) Mengubah semua karakter alfabet pada dokumen menjadi huruf kecil (*case folding*).

2) Dilakukan proses *filtering* untuk membuang karakter yang tidak bersifat signifikan, seperti: tanda baca dan angka.
3) Mengekstrak dokumen untuk mendapatkan kata-kata (*term*) dengan Porter stemmer dan *stop word list* yang diambil dari http://download.csdn.net/source/1568518.
4) Kata-kata yang telah didapat tadi akan diberi bobot dengan metode tf-idf dan terakhir kata-kata berbobot akan direpresentasikan ke dalam Vector Model Space (VSM).



Gambar 3.1Alur sistem secara umum

Kedua, akan diimplementasikan one pass clustering yang bertujuan untuk membangun model klasifikasi. Pembangunan model klasifikasi dilakukan dengan algoritma one pass clustering.

Ketiga, diimplementasikan algoritma k-nearest neigbour (KNN) berdasarkan model klasifikasi yang telah dibangun.

3. Uji Coba dan Analisis

Dataset yang digunakan pada uji coba ini adalah Reuters-21578 yang diunduh dari http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection. *Dataset* ini terdiri dari 21578 dokumen dan 135 kategori yang merupakan kumpulan dari berita di koran berbahasa Inggris Reuters tahun 1987. Tiap dokumen dari *dataset* ini bisa terdiri dari banyak kategori. *Dataset* dalam koleksi terdiri dari 22 berkas. Tiap berkasnya terdiri dari 1000 artikel berita yang terbentuk dengan format SGML. Dengan format SGML, tiap artikel diawali dengan *tag* `<REUTERS TOPICS=?? LEWISSPLIT=?? CGISPLIT=?? OLDID=?? NEWID=??>` dan diakhiri dengan *tag* `</REUTERS>`. Tiap teks tersebut memiliki *tag* `<TOPICS>, </TOPICS>` yang menunjukkan kategori dari artikel tersebut dan tiap kategori dipisahkan dengan *tag* `<D>, </D>`. Gambar 3.2 menampilkan contoh artikel dengan format SGML.

Uji coba diawali dengan membagi *dataset* menjadi 2 bagian, yaitu: data latih dan data uji. Dari 135 kategori yang ada, diambil tujuh kategori yang paling sering muncul dalam *dataset*. Ketujuh kategori itu adalah ACQ, *corn*, *crude*, *earn*, *interest*, *ship*, dan *trade*. Data latih digunakan untuk membangun model klasifikasi dengan algoritma one pass clustering dengan nilai ε = 8 yang digunakan untuk menginisialisasi nilai *threshold r*. Selanjutnya, data uji digunakan untuk menguji model klasifikasi yang telah dibangun dengan algoritma KNN. Uji coba kemudian akan dilakukan dengan membandingkan kinerja dari algoritma KNN berdasarkan algoritma one pass clustering dengan algoritma KNN tradisional dalam kategorisasi teks. Nilai *k* yang dipakai untuk menguji algoritma KNN tradisional adalah 10. Pengukuran kinerja atau evaluasi dilakukan menggunakan metode F_1 dan *Macro-*F_1. Nilai F_1 digunakan mengevaluasi kinerja algoritma tiap kategori. Sedangkan, *Macro-*F_1 digunakan untuk menguji algoritma secara keseluruhan. Metode yang memiliki nilai *Macro-*F_1 lebih tinggi, maka kinerja dari metode tersebut dianggap lebih baik dari yang lain.

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="5552" NEWID="9">
<DATE>26-FEB-1987 15:17:11.20</DATE>
<TOPICS><D>earn</D></TOPICS>
<PLACES><D>usa</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;F
&#22;&#22;&#1;f0762&#31;reute
r f BC-CHAMPION-PRODUCTS-&lt;CH   02-26 0067</UNKNOWN>
<TEXT>&#2;
<TITLE>CHAMPION PRODUCTS &lt;CH> APPROVES STOCK SPLIT</TITLE>
<DATELINE>    ROCHESTER, N.Y., Feb 26 - </DATELINE><BODY>Champion Products Inc said its
board of directors approved a two-for-one stock split of its
common shares for shareholders of record as of April 1, 1987.
    The company also said its board voted to recommend to
shareholders at the annual meeting April 23 an increase in the
authorized capital stock from five mln to 25 mln shares.
 Reuter
&#3;</BODY></TEXT>
</REUTERS>
```

Gambar 3.2 Contoh artikel berita dalam format SGML

4 JADWAL PENGERJAAN TUGAS AKHIR

No.	Kegiatan	Bulan														
		Oktober				November				Desember				Januari		
1	Studi literatur dan penyusunan proposal															
2	Implementasi															
	a. Preprocessing dataset															
	b. One pass clustering															
	c. K-nearest neigbour (KNN)															
	d. Evaluasi															
3	Uji coba dan analisis															
4	Penulisan buku tugas akhir															

5 DAFTAR PUSTAKA

Jiang, S., Pang, G., Wu, M. & Kuang, L., 2011. An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, Volume 39, pp. 1503-1509.

Jian, S., 2006. *Efficient Classification Method For Large Dataset.* Dalian, In Proceedings of the Fifth International Conference on Machine Learning and Cybernetics.

Manning, C. D., Raghavan, P. & Schütze, H., 2009. *An Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Porter, M., 2006. *Porter Stemming Algorithm.* [Online]
Available at: http://tartarus.org/~martin/PorterStemmer/
[Diakses 29 Agustus 2011].

Tan, P.-N., Steinbach, M. & Kumar, V., 2006. *Introduction to Data Mining*. Boston: Pearson Addison Wesley.

Tan, S., 2008. An improved centroid classifier for text categorization. *Expert Systems with Applications*, Volume 35, pp. 279-285.

LEMBAR PENGESAHAN

Surabaya, 13 Oktober 2011

Menyetujui,

Pembimbing I	Pembimbing II
Isye Arieshanti, S.Kom, M.Phil	**Dr. Agus Zainal Arifin, S.Kom, M.Kom**
NIP. 19780412 200604 2 001	NIP. 19720809 199512 1 001