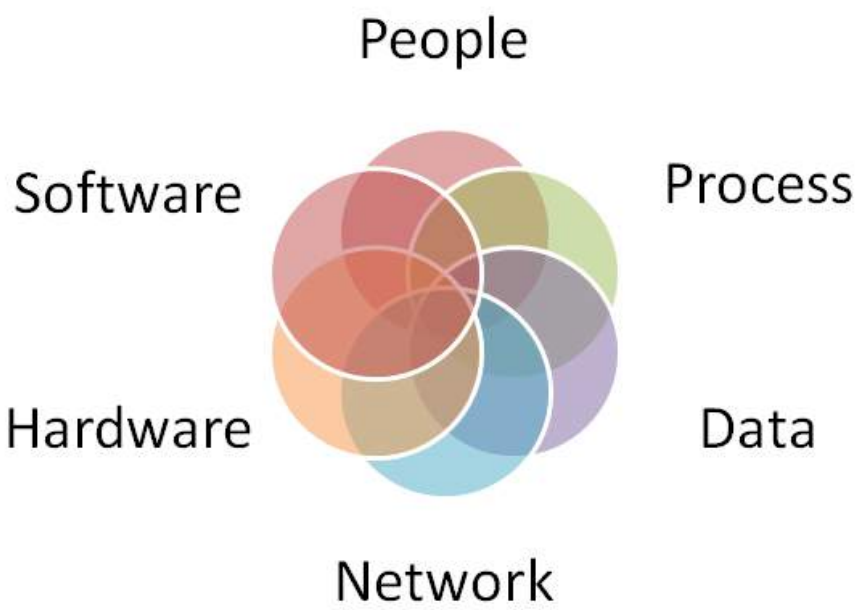
OAJIS
Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

ISSN 1979-3979

jurnal sisfo

Inspirasi Profesional Sistem Informasi



Vol. 07 No. 02 Januari 2018

OAJIS
Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

jurnal sisfo

Jurnal Sisfo Vol. 07 No. 02 (2018) i–ii

Sekretariat

Departemen Sistem Informasi – Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember (ITS) – Surabaya
Telp. 031-5999944 Fax. 031-5964965
Email: editor@jurnalsisfo.org
Website: http://jurnalsisfo.org

Jurnal SISFO juga dipublikasikan di *Open Access Journal of Information Systems* (OAJIS)

Website: http://is.its.ac.id/pubs/oajis/index.php

OAJIS
Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

jurnal sisfo

Jurnal Sisfo Vol. 07 No. 02 (2018) i–ii

Mitra Bestari

Ahmad Mukhlason, S.Kom, M.Sc, Ph.D (Institut Teknologi Sepuluh Nopember)

Dr. Darmawan Napitupulu, S.T, M.Kom (Lembaga Ilmu Pengetahuan Indonesia)

Faizal Johan Atletiko, S.Kom, M.T (Institut Teknologi Sepuluh Nopember)

Ir. Dana Indra Sensuse, MLIS, Ph.D (Universitas Indonesia)

Nur Aini Rakhmawati, Ph.D (Institut Teknologi Sepuluh Nopember)

Nurul Khaqiqi, S.Pi, M.P (Laboratorium Perikanan Banyuwangi)

Radityo Prasetianto.W, S.Kom, M.Kom (Institut Teknologi Sepuluh Nopember)

Retno Aulia Vinarti, S.Kom, M.Kom (Institut Teknologi Sepuluh Nopember)

Rully Agus Hendrawan, S.Kom, M.Eng (Institut Teknologi Sepuluh Nopember)

Satria Fadil Persada, S.Kom, M.BA, Ph.D (Institut Teknologi Sepuluh Nopember)

Wayan Firdaus Mahmudy, S.Si., M.T., Ph.D (Universitas Brawijaya)

OAJIS
Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

jurnal sisfo

Jurnal Sisfo Vol. 07 No. 02 (2018) iii

Daftar Isi

Halaman ini sengaja dikosongkan

OAJIS
Open Access
Journal of
Information
Systems
is.its.ac.id/pubs/oajis/

jurnal sisfo

Jurnal Sisfo Vol. 07 No. 02 (2018) 83–98

Klasifikasi Data Twitter Pelanggan Berdasarkan Kategori myTelkomsel Menggunakan Metode *Support Vector Machine* (SVM)

Sila Prayoginingsih, Renny Pradina Kusumawardani*

Departemen Sistem Informasi, Fakultas Teknologi Informasi dan Komunikasi, Institut Teknologi Sepuluh Nopember

Abstract

This research performs classification on social media text, specifically for the case of customer complaint in the telecommunication industry. To represent complaint criteria relevant to telecommunication services, we use the categories used in myTelkomsel, a web application of Telkomsel. Although this application enables customers to file in their complaints directly in a self-service manner, many customers opt to post their complaints in the social media such as Twitter. Therefore, in this research we create a classification model using Support Vector Machines (SVMs) to enable the automatic categorization of such customer complaints. As the input for the training and testing process, we crawl Twitter using the Streaming API. The data is then filtered to get tweets containing information, complaints, criticisms, suggestions, and questions about Telkomsel's products or services. Using RBF kernels optimized with grid search, the resulting classifier gives good accuracy and f-measure of 84.84% and 84.88%, respectively.

Keywords: Classification, Twitter, Customer, Support Vector Machine, myTelkomsel, Telkomsel, Grid Search

Abstrak

Penelitian ini melakukan klasifikasi pada teks media sosial kasus aduan pelanggan untuk layanan telekomunikasi. Untuk merepresentasikan kriteria aduan pelanggan yang relevan bagi perusahaan telekomunikasi, dipergunakan kategori pada aplikasi myTelkomsel, yaitu layanan web dari Telkomsel. Meskipun aplikasi ini memungkinkan keluhan untuk disampaikan secara *self-service*, namun masih banyak pelanggan yang memilih untuk meyampaikan aduannya melalui media sosial seperti Twitter. Pada penelitian ini dilakukan pembentukan model klasfikasi dengan algoritma *Support Vector Machines* (SVM), sehingga pengaduan dapat dikategorikan secara otomatis. Sebagai masukan untuk pembentukan pengklasifikasi, dipergunakan data hasil *crawling* Twitter dengan menggunakan Streaming API. Data kemudian difilter untuk mendapatkan *tweet* terkait dengan informasi, keluhan, kritik, saran, dan pertanyaan seputar layanan atau produk Telkomsel. Menggunakan kernel RBF yang dioptimasi dengan metode *grid search*, didapatkan pengklasifikasi dengan performa yang baik, yaitu dengan akurasi dan *f-measure* sebesar 84.84% dan 84.88%.

Kata kunci: Klasifikasi, Twitter, Pelanggan, Support Vector Machine, myTelkomsel, Telkomsel, Grid Search

Histori Artikel: Disubmit 24 Oktober 2017; Diterima 13 Desember 2017; Tersedia online 2 Januari 2018

*Corresponding Author
Email address: renny.pradina@gmail.com (Renny Pradina Kusumawardani)

1. Pendahuluan

Perkembangan teknologi informasi mempengaruhi pertumbuhan penggunaan internet di Indonesia yang semakin meningkat setiap tahun [1]. Pengguna internet tidak hanya *browsing* saat mengakses internet, tetapi juga menggunakan *social media* [2]. Menurut survei *Nielsen On Device Meter* (ODM) yang dilakukan oleh Nielsen pada Februari 2014, sebanyak 36% pengguna *smartphone* menggunakan media sosial twitter [3]. Penggunaan twitter yang semakin pesat membuat organisasi memanfaatkan twitter sebagai *customer service* dengan menyediakan media untuk memberikan keluhan, saran atau pertanyaan. Namun, belum banyak organisasi di Indonesia yang memfasilitasi pengguna untuk aduan melalui twitter.

Telkomsel merupakan perusahaan operator layanan telepon seluler terbesar di Indonesia [4]. Perusahaan ini menyediakan layanan yaitu "*myTelkomsel*" dalam memberikan saran, kritik maupun pertanyaan secara privasi dengan memasukkan *username* berupa nomor *handphone* dan *password* berupa PIN T-Care [5]. Program layanan myTelkomsel ini merupakan *self-service web* [6] yang memiliki beberapa kategori yaitu Fitur, Jaringan, Jaringan 4G, Penipuan, Program Khusus, *Value Added Services*, Voucher dan Isi Ulang dan Pertanyaan Umum [5].

Namun, tidak semua pengguna Telkomsel mengetahui layanan *myTelkomsel* atau memiliki nomor dan pin *myTelkomsel*. Menurut data statistik dari SocialBakers dan MediaBistro, menunjukkan bahwa Indonesia khusunya Jakarta menyumbang 2.4% dari 10.6 milyar *twitter post* di seluruh dunia dari bulan Januari hingga Maret 2015 [7]. Berdasarkan informasi tersebut, masyarakat Indonesia lebih *prefer* menggunakan twitter untuk menyampaikan opini mereka. Selain itu, akan lebih memudahkan pelanggan Telkomsel jika menyampaikan opini melalui Twitter sehingga tidak perlu repot memasukkan *username* dan *password* pada myTelkomsel.

Maka dari itu, PT. Telekomunikasi Selular perlu memanfaatkan twitter lebih intens. Pemanfaatan twitter berguna untuk memetakan *tweet* kedalam beberapa kategori. Hal inilah yang akhirnya menimbulkan kebutuhan untuk melakukan klasifikasi tweet pengguna Telkomsel sesuai kategori pada myTelkomsel. Metode yang digunakan adalah *Support Vector Machine* (SVM). Metode SVM terbukti efektif [8] untuk klasifikasi teks.

Dalam penelitian ini, dilakukan proses klasifikasi teks menggunakan metode SVM yang memetakan tweet sesuai kategori *myTelkomsel* serta kategori tambahan jika diperlukan. Dengan adanya *paper* ini, diharapkan dapat dijadikan referensi untuk penelitian selanjutnya serta membantu pihak Telkomsel mengetahui kategori yang memiliki pengaruh terhadap pelayanan menurut pelanggan. Dengan demikian, dapat memberikan *value added* sehingga dapat meningkatkan pelayanan sesuai kebutuhan pelanggan.

2. Tinjauan Pustaka

2.1 Klasifikasi

Klasifikasi merupakan suatu proses pencarian dari kumpulan model atau fitur data yang dapat digunakan untuk membedakan label kelas data dengan tujuan agar model yang terbentuk dapat memprediksi kelas dari suatu objek dengan tepat. Klasifikasi merupakan salah satu tugas dari pembelajaran mesin (*machine learning*) yang membutuhkan ketersediaan data berlabel sehingga disebut sebagai *supervised learning* yang termasuk kedalam model yang prediktif, yaitu model yang menghasilkan output berupa nilai variable target.

Evaluasi performa pada task klasifikasi sangat dibutuhkan untuk melihat seberapa tepat prediksi label kelas atau target yang dihasilkan oleh model yang telah dibentuk. Pada klasifikasi, data dibagi menjadi 2 bagian untuk melihat secara tepat performa model, yaitu sebagai data pelatihan (*training set*) yang

digunakan untuk membentuk model dan data pengujian (*test set*) yang digunakan untuk menguji performa model. Pembagian data ini bertujuan untuk mengevaluasi performa model agar dapat terukur secara obyektif.

2.2 Text Preprocessing

Text Prepocessing merupakan proses melakukan pengubahan bentuk dari informasi-informasi yang strukturnya 'sembarang' menjadi data yang terstruktur [9]. *Prepocessing* dilakukan untuk mengetahui dan menghindari data yang kurang sempurna, gangguan dan tidak konsisten [10]. Oleh karena itu, dibutuhkan teknik dalam *text preprocessing* yaitu:

1) *Case Folding*
2) *Tokenizing*, yaitu tahapan pemotongan string input berdasarkan kata yang menyusunnya [11] sehingga menjadi kumpulan kata yang berdiri sendiri.
3) *Filtering*
4) *Stoplist*, berisi kumpulan kata yang tidak relevan, namun sering kali muncul di dalam suatu dokumen.
5) *Stopwords Removal*
6) *Stemming*, merupakan tahapan dalam mencari akar kata dari setiap kata hasil dari filtering atau mencari kata dasar dari kata yang berimbuhan.

2.3 Support Vector Machine

Support Vector Machine (SVM) adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur (*feature space*) berdimensi tinggi, dipelatihan dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan *learning bias* yang berasal dari teori pembelajaran statistik [12] [13]. Metode pembelajaran ini ialah *supervised*, membutuhkan ketersediaan data berlabel [14]. *Supervised learning* merupakan suatu pendekatan yang memiliki data yang telah dilatih sebelumnya (*train data*) dan variabel target, sehingga tujuan dari pembelajaran ini adalah mengelompokkan data ke data yang sudah ada. SVM adalah salah satu teknik yang baru dibandingkan dengan teknik lain, tetapi memiliki performansi yang lebih baik di berbagai bidang aplikasi seperti *bioinformatics*, pengenalan tulisan tangan, klasifikasi teks dan lain sebagainya [13].

Pada *machine learning*, terdapat istilah *kernel trick* yang merupakan metode yang menggunakan algoritma *linier classifier* untuk menyelesaikan permasalahan nonlinier. Menurut C.W Hsu dkk, persamaan (1) hingga (4) berikut ini adalah beberapa fungsi kernel yang umum digunakan antara lain [15]:

$$Linear: K(x_i, x_j) = X_i^T X_j \quad (1)$$

$$Polynomial: K(x_i, x_j) = (\gamma . X_i^T X_j + r)^d, \gamma > 0 \quad (2)$$

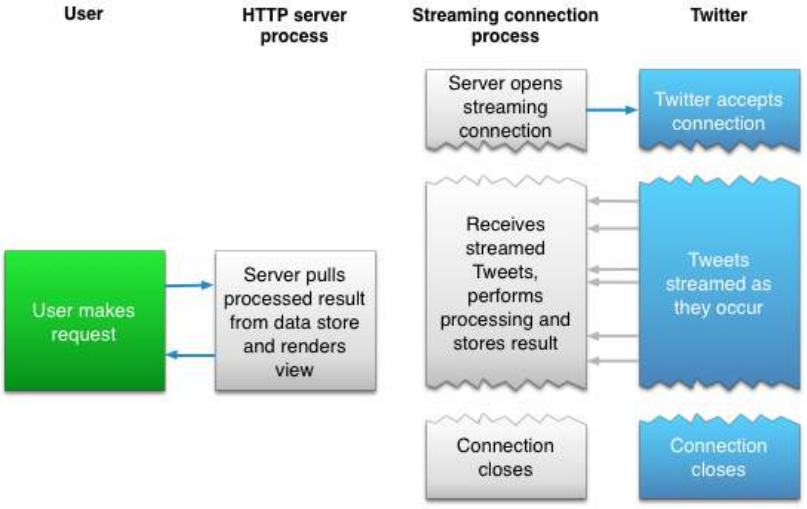
$$RBF: K(x_i, x_j) = exp\left(-\gamma \left| X_i - X_j \right|^2\right), \gamma > 0 \quad (3)$$

$$Sigmod: K(x_i, x_j) = tanh(\gamma . X_i^T X_j + r) \quad (4)$$

Proses klasifikasi teks pada twitter Telkomsel hanya menggunakan dua fungsi kernel yaitu linear dan RBF. Pemilihan 2 fungsi kernel tersebut disebabkan oleh beberapa faktor, diantaranya kernel linear sering digunakan dan cocok pada proses klasifikasi teks, memiliki peforma yang cepat, dapat digunakan dengan baik pada *linear separable data* dan yang memiliki banyak fitur, sedikit parameter yang dioptimalkan, yaitu C (*cost)*. Kernel RBF menggunakan 2 parameter yakni C (*cost)* dan γ (*gamma*) yang dapat melihat seberapa jauh pengaruh dari contoh pelatihan data tunggal tercapai. Konfigurasi 2 kernel serta parameter didalamnya dilakukan untuk mengetahui kernel manakah yang paling mempengaruhi performa klasifikasi SVM.

2.4 Twitter API Stream

Twitter API Stream merupakan salah satu layanan yang diberikan Twitter kepada *developer* untuk mengakses data twitter baik *public* maupun *protected* . *Streaming API* ini menyediakan akses pengguna untuk menerima tweet secara *real-time* dari twitter [16]. Pengguna diijinkan untuk mengambil data pada *local database* dan menggunakannya untuk ditampilkan pada website atau aplikasi. Kelebihan *Streaming API* ini adalah pengguna dapat melacak peristiwa yang masuk secepat mungkin dan mengolahnya disuatu aplikasi untuk mendapatkan data yang lebih mendalam. Proses *API Stream* bisa dilakukan dengan memberikan *keyword* yang diinginkan misalkan berdasarkan lokasi geografis, *user ID* atau tanggal tweet dan lain-lain. Setelah *crawling* data dirasa sudah memenuhi, program *Twitter API Stream* dapat dihentikan, sehingga didapatkan data pada database MySQL. Gambar 1 memberikan visualisasi dari mekanisme *Twitter API Stream.*



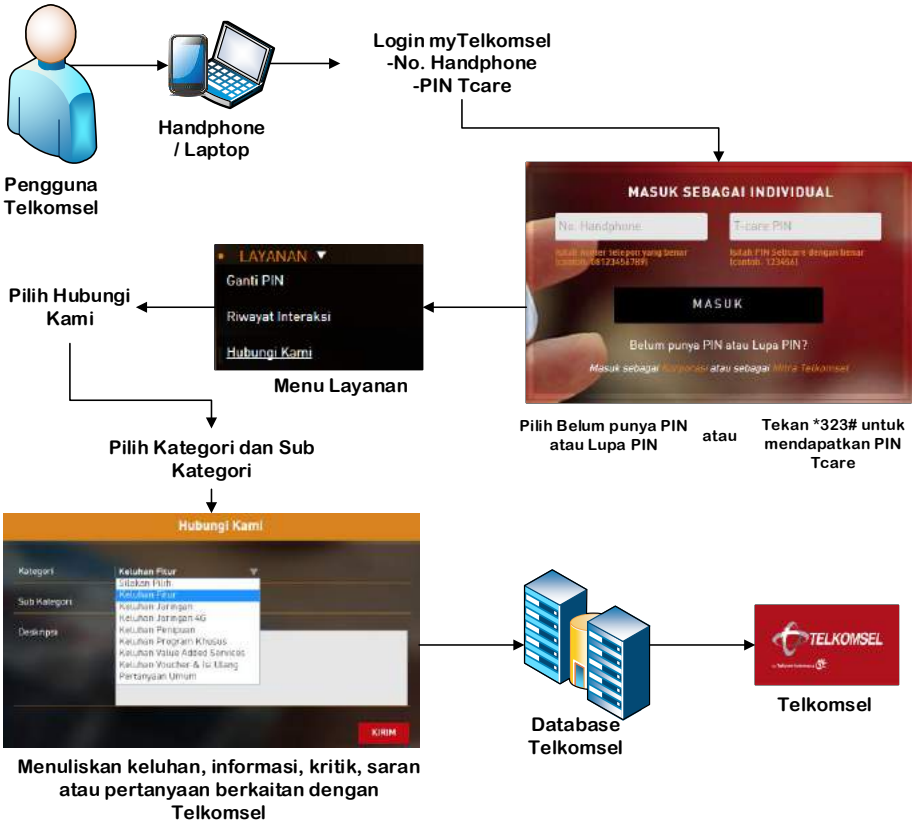
Gambar 1 Mekanisme twitter API stream [16]

2.5 myTelkomsel

Aplikasi myTelkomsel merupakan program *Seft Service Web My Telkomsel* yang berbasiskan web atau aplikasi yang dapat di install di *personal device* [5]. Layanan *customer service* milik Telkomsel ini, dapat mengakomodasi keinginan dan kebutuhan pelanggan. Kebutuhan pelanggan yang dimaksud diantaranya adalah untuk mengetahui berbagai macam informasi terkait akun yang dimiliki pelanggan, mengajukan permintaan untuk mengaktifkan fitur, perubahan paket dan memberikan *feedback* (kritik, saran, keluhan dan pertanyaan) kepada Telkomsel [5]. Pengguna yang ingin *login* myTelkomsel harus memiliki nomor dari produk Telkomsel dan memiliki PIN T-Care. Jika pengguna belum memiliki PIN T-Care, maka pengguna bisa mendapatkan PIN tersebut dengan memasukkan nomor *handphone*, tanggal lahir dan *captcha*. Setelah menunggu beberapa detik, pengguna akan mendapatkan sms resmi dari Telkomsel mengenai PIN T-Care.

Program myTelkomsel yang bersifat *private* ini memiliki menu yaitu Hubungi Kami. Menu Hubungi Kami merupakan bagian penting dalam penelitian karena menu ini memiliki kategori yang digunakan sebagai label kelas pada klasifikasi kategori. Beberapa kategori yang ada pada myTelkomsel yaitu Fitur, Jaringan, Jaringan 4G, Penipuan, Program Khusus, *Value Added Services*, *Voucher* dan Isi Ulang dan Pertanyaan Umum [5]. Pada penelitian ini, label yang dibutuhkan untuk klasifikasi adalah 8 kategori myTelkomsel dan satu kategori yang menjadi issue perbincangan selama proses pengumpulan data yaitu SSH-Netflix.

Alur penyampaian keluhan, kritik, saran, informasi dan pertanyaan melalui aplikasi myTelkomsel diilustrasikan pada Gambar 2.



Gambar 2 Alur penyampaian feedback via myTelkomsel

2.6 Evaluasi Performa Klasifikasi

Ketepatan prediksi dari pengklasifikasi multikelas dapat dilihat pada tangkapan layer dari tabel kontingensi multikelas pada Gambar 3 [17]:

	Prediksi c = A	**Prediksi c = B**	**Prediksi c = C**	**Prediksi c = D**	
Aktual c = A	**True A**	(FP B|c = A)	(FP C|c = A)	(FP D|c = A)	Total Aktual c = A
Aktual c = B	(FP A|c = B)	**True B**	(FP C|c = B)	(FP D|c = B)	Total Aktual c = B
Aktual c = C	(FP A|c = C)	(FP B|c = C)	**True C**	(FP D|c = C)	Total Aktual c = C
Aktual c = D	(FP A|c = D)	(FP B|c = D)	(FP C|c = D)	**True D**	Total Aktual c = D
	Total Prediksi c = A	Total Prediksi c = B	Total Prediksi c = C	Total Prediksi c = D	**TOTAL**

Gambar 3 Tangkapan layar tabel kontingensi multikelas

Tabel 1 berikut memberikan penjelasan mengenai beberapa terminologi yang digunakan dalam makalah ini, diantaranya yaitu [18]:

Tabel 1. Terminologi tabel kontingensi multikelas

Data	Terminologi
Aktual	Label atau kelas sesungguhnya dari suatu titik data
Prediksi	Label atau kelas dari titik data berdasarkan prediksi dari model
True Positives (TN)	Titik data yang diklasifikasikan model sebagai positif dan label yang sebenarnya memang positif (prediksi benar)
False Positives (FP)	Titik data yang diklasifikasikan model sebagai positif dan label yang sebenarnya memang negatif (prediksi salah)

Untuk permasalahan dalam klasifikasi yang melibatkan klasifikasi multikelas, ukuran performa yang biasa digunakan adalah uji akurasi, uji *precision*, uji *recall* dan *F-measure* [19].

Akurasi merupakan presentase dari total data yang benar teridentifikasi. Persamaan (5) berikut merupakan rumus untuk perhitungan akurasi.

$$Accuration = \frac{True\ i + \cdots + True\ n}{Total} x100\% \quad (5)$$

Keterangan:
Accuration = tingkat akurasi
True = jumlah klasifikasi benar dari kategori *ke-i* hingga *ke-n*
Total = jumlah data uji coba

Precision adalah ukuran ketepatan prediksi pengklasifikasi dengan benar. *Precision* dapat diformulasikan sebagaimana dalam persamaan (6) berikut.

$$Precision(i) = \frac{\Sigma x(c = i|\hat{c} = i)}{\Sigma x_{(\hat{c}=i)}}\ x\ 100\% \quad (6)$$

Recall adalah perbandingan jumlah item yang relevan dengan item yang ada. Persamaan untuk *recall* adalah sebagaimana ditunjukkan pada persamaan (7) berikut.

$$Recall\ (i) = \frac{\Sigma x(\hat{c} = i|c = i)}{\Sigma x_{(c=i)}}\ x\ 100\% \quad (7)$$

Keterangan:
Precision = tingkat ketepatan
Recall = tingkat ketepatan
C = jumlah data uang relevan
$\hat{C}$ = jumlah data yang ditemukan

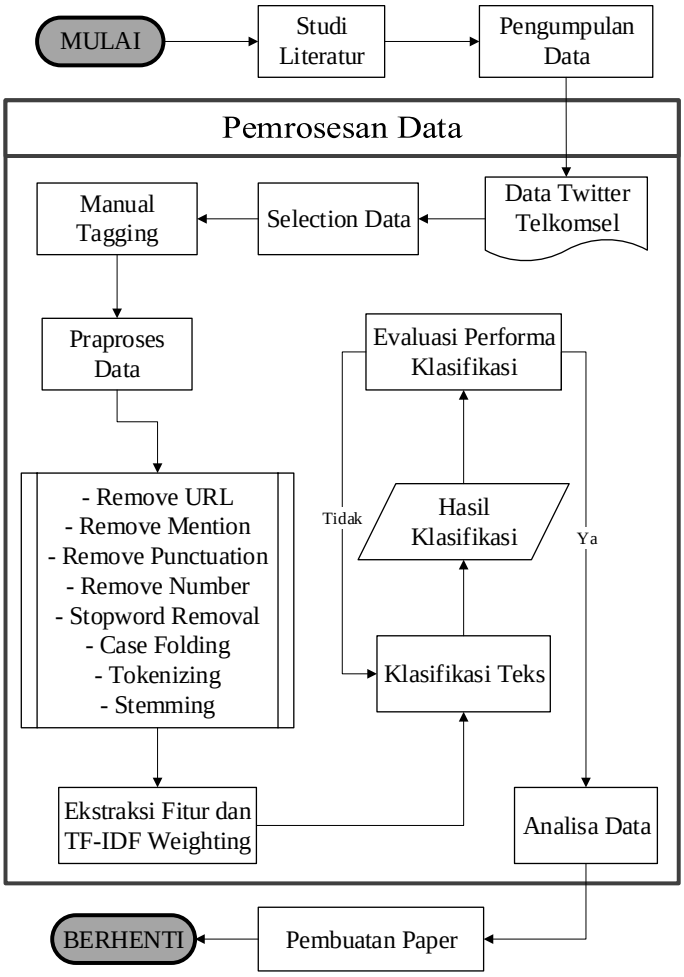
Parameter tunggak untuk ukuran keberhasilan *informational retrieval* adalah parameter *F-Measure.* Persamaan *F-Measure* sebagaimana dalam persamaan (8) berikut.

$$F - Measure = \frac{2*precision*recall}{precision+recall} \quad (8)$$

3. Metodologi

Permasalahan pada penelitian ini akan diselesaikan dengan metode penelitian yang tergambar dalam diagram alir pada Gambar 4.

Langkah pertama ialah studi literatur dengan meninjau dan memahami dasar-dasar teori yang berkaitan. Proses pengumpulan data twitter Telkomsel diperoleh dengan layanan *Twitter API Stream*. Proses pengumpulan data dengan Twitter API Stream ini dilakukan selama 14 hari (25 Januari – 7 Februari 2016) sehingga menghasilkan sekitar 52023 data.



Gambar 4 Diagram alir penelitian

Proses pemilihan data (*selection data*) dilakukan dengan membagi data menjadi kategori relevan dan tidak relevan. Kategori tidak relevan merupakan tweet tidak berkaitan dengan informasi, keluhan, kritik, saran dan pertanyaan seputar layanan atau produk Telkomsel begitu sebaliknya. Proses pembagian data yang tidak relevan dan relevan, data awal yang berjumlah 52023 data berubah menjadi 9651 data yang relevan dan 42371 data yang tidak relevan. Data sampel kategori tidak relevan yang digunakan adalah 9690 data. Hal ini dimaksudkan agar hasil dari perhitungan seimbang atau tidak berat sebelah (akurasi besar pada data yang lebih besar).

Oleh karena itu, pembagian data yang digunakan adalah 9690 data untuk kategori tidak relevan dan 9652 untuk kategori relevan. Untuk mengetahui keakuratan pembagian data secara manual ini maka dilakukan proses klasifikasi menggunakan metode SVM. Data klasifikasi dibagi lagi menjadi 2 bagian yaitu data pelatihan dan pengujian dengan proporsi 70% dan 30%. Data tersebut diuji coba menggunakan metode SVM dengan parameter kernel linear dan RBF. Hasil akurasi yang memiliki performa bagus dapat digunakan untuk proses klasifikasi lebih lanjut. Hanya data yang masuk kedalam kategori relevan yang akan dilakukan klasifikasi teks.

Proses memberikan label manual berdasarkan kategori myTelkomsel dan satu kategori tambahan yaitu SSH-Netflix. Langkah selanjutnya adalah praproses data, dengan tahapan *case folding, filtering, stopword removal, tokenize* dan *stemming.* Pada saat tahapan *filtering* dilakukan proses *remove url* (menghapus url atau link yang ikut pada tweet), *remove mention & hastag* (menghilangkan nama *user, account mention* dan *hastag*), *remove punctuation* (menghilangkan tanda baca yang tidak diperlukan) dan *remove number* (menghilangkan angka-angka yang tidak perlu seperti nomor *handphone*, alamat rumah dan lain-lain). Proses *stopword removal* menggunakan *stopword* yang merujuk pada thesis Fadillah Z. Tala dengan judul "*A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*" [20].

Selanjutnya adalah membagi data menjadi 2 bagian yaitu fase pelatihan (*training data*) dan fase pengujian (*testing data*). Kedua fase tersebut memiliki bobot masing-masing 70% dan 30%. Ekstraksi fitur dan pembobotan dengan menggunakan TF (*term frequency*) – IDF (*inverse document frequency*) dengan *min_df* = 75 karena nilai tersebut menghasilkan jumlah fitur yang konstan. Mekanisme data teks berubah menjadi vektor bertujuan untuk memberikan bobot berdasarkan seberapa penting teks tersebut didalam tweet. Pembobotan teks dilakukan menggunakan metode *Count Vectorizer* dan *TF-IDF Transformer.* Vektor input dengan metode *Count Vectorizer* menghasilkan angka *binary* yaitu hanya 0 dan 1, sedangkan jika menggunakan *TF-IDF Transformer* menghasilkan luaran angka yang memiliki nilai *decimal* atau *float.*

Implementasi *Support Vector Machine* pada penelitian ini hanya akan menguji dua kernel yaitu kernel linear dan RBF (*Radial Basis Function*). Tahapan ini melakukan berbagai macam percobaan atau penemuan-penemuan baru seperti tertera pada Tabel 2.

Tabel 2. Percobaan klasifikasi

Skenario				Algorithma	
Percobaan	PP + SR	SS	ST	SVC (Kernel)	
				Linear	RBF
Percobaan 1	V	-	-	V	-
Percobaan 2	V	V	-	V	-
Percobaan 3	V	V	V	V	-
Percobaan 4	V	-	-	-	V
Percobaan 5	V	V	-	-	V
Percobaan 6	V	V	V	-	V

Keterangan:
PP = Praproses
SR = Stopword Removal
SS = Special Stopword
ST = Stemming
SVC = Support Vector Classifier

Evaluasi performa dilakukan dengan uji akurasi, uji *precision* dan uji *recall* dan *f-measure.* Jika hasil akurasi belum maksimal maka menggunakan algoritma *grid search.* Dengan *grid search,* akan memperhitungkan parameter-parameter yang mungkin untuk mendapatkan hasil yang optimal atau terbaik.

4. Hasil dan Pembahasan

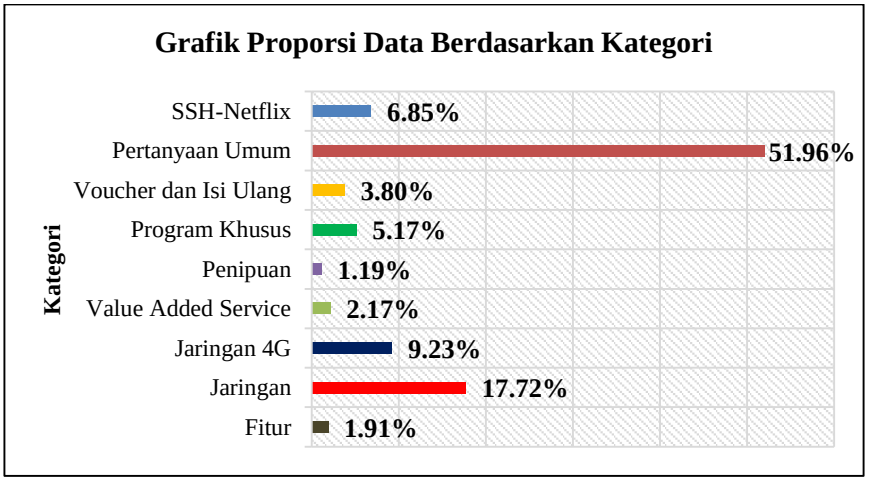
4.1 Data Tweet Telkomsel

Data tweet Telkomsel ialah data yang telah diperoleh dari proses *crawling twitter* selama 2 minggu dari 25 Januari hingga 7 Februari 2016. Data yang diperoleh sebesar 52013 data, kemudian dilakukan *cleansing data* dari *tweet* yang tidak relevan sehingga menjadi 9651 data relevan. Data tersebut kemudian dilakukan pelabelan manual (*tagging*) berdasarkan kategori yang ada pada aplikasi myTelkomsel dan SSH-Netflix. Berdasarkan data yang sudah di *tagging manual,* dapat dihitung jumlah jenis kategori. Dataset yang diperoleh dibagi menjadi 2 bagian yaitu *data training* (70%) dan *data testing* (30%) seperti pada Tabel 3.

Tabel 3. Dataset klasifikasi

Jenis Kategori			
Fitur	184		
Jaringan	1710		
Jaringan 4G	891		
Value Added Service	209		
Penipuan	115	Data Training (70%)	Data Testing (30%)
Program Khusus	499		
Voucher dan Isi Ulang	367		
Pertanyaan Umum	5015		
SSH-Netflix	661		
Total	9651	6755	2896

Gambar 5 berikut menampilkan proporsi data tweet yang telah dikategorisasikan secara manual.



Gambar 5 Proporsi data berdasarkan kategori

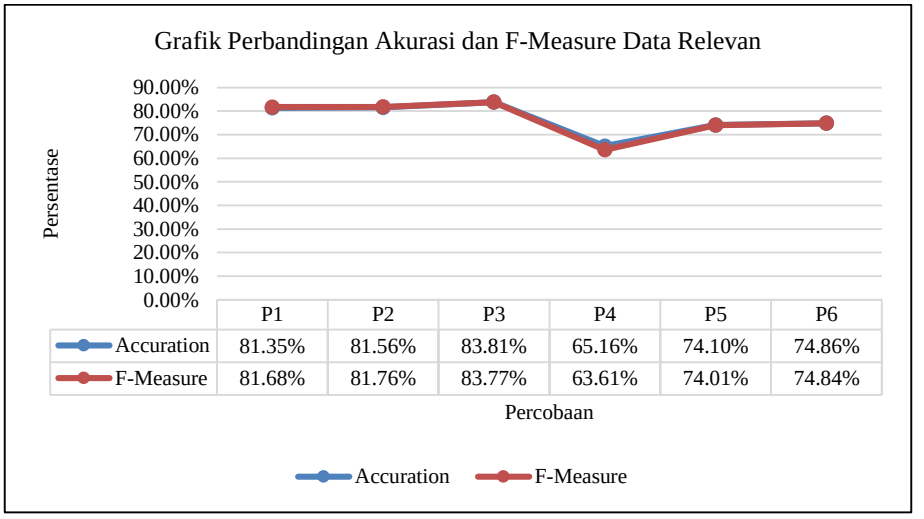
4.2 Hasil Percobaan

Penelitian ini melakukan semua percobaan, baik melalui tahapan *stopword removal*, *special stopword* dan *stemming* maupun tidak dengan 2 konfigurasi kernel. Berdasarkan hasil percobaan diperoleh nilai akurasi, *precision*, *recall* dan *f-measure* seperti Tabel 4. Percobaan kernel linear memiliki nilai akurasi dan *f-measure* tertinggi. Selain itu, tahapan proses *stemming* memberikan pengaruh positif performa klasifikasi. Optimasi parameter fitur *grid search* dilakukan pada klasifikasi kernel Linear dengan *special stopword* dan *stemming* (3) dan klasifikasi kernel RBF dengan *special stopword* dan *stemming* (6). Percobaan tersebut dipilih karena nilai akurasi dan *f-measure* tertinggi berdasarkan konfigurasi kernel.

Tabel 4. Perbandingan evaluasi performa klasifikasi

Percobaan	Uji Validasi			
	Akurasi	Precisi-on	Recall	F-Measure
Percobaan 1	81.35%	82.02%	81.35%	81.68%
Kernel Linear, Stopword Removal				
Percobaan 2	81.56%	81.96%	81.56%	81.76%
Kernel Linear, Stopword Removal, Special Stopword				
Percobaan 3	83.81%	83.75%	83.80%	83.77%
Kernel Linear, Stopword Removal, Special Stopword + Stemming				
Percobaan 4	65.15%	62.13%	65.16%	63.61%
Kernel RBF, Stopword Removal				
Percobaan 5	74.10%	73.92%	74.10%	74.01%
Kernel RBF, Stopword Removal, Special Stopword				
Percobaan 6	74.86%	74.83%	74.86%	74.84%
Kernel RBF, Stopword Removal, Special Stopword + Stemming				

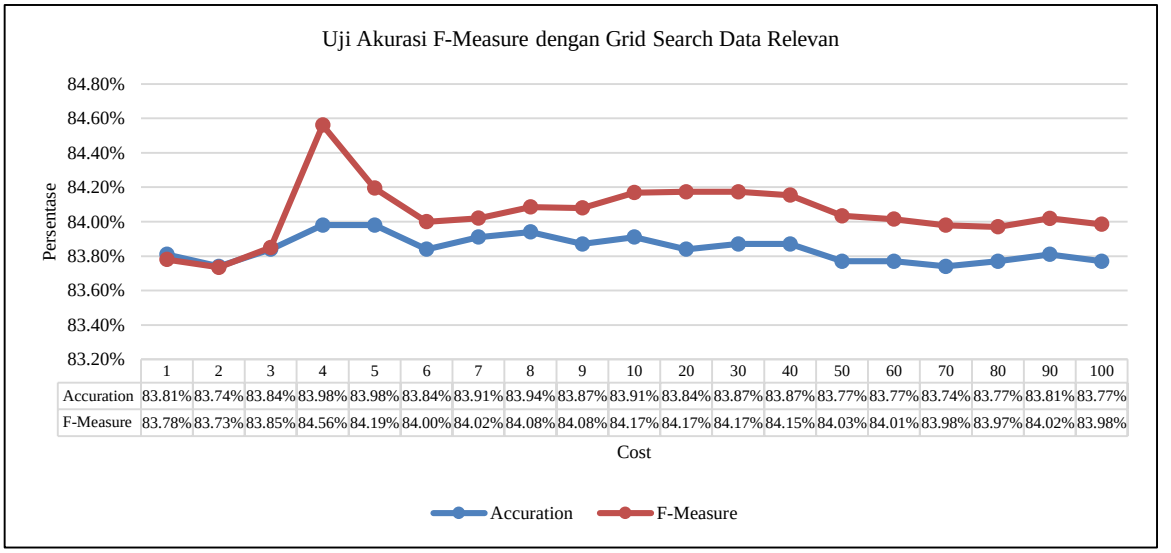
Gambar 6 berikut menampilkan perbandingan antara nilai akurasi dan *f-measure* dari setiap percobaan.



Gambar 6 Perbandingan akurasi dan *f-measure* data relevan

4.2.1 Hasil Klasifikasi Kernel Linear dengan Stopword Removal, Special Stopword dan Stemming dengan Grid Search

Percobaan ini menghasilkan nilai akurasi 83.97% dan *f-measure* 84.56% dengan menggunakan parameter terbaik yaitu *cost* = 4. Gambar 7 menunjukkan hasil uji akurasi *f-measure* dengan *grid search* dalam penelitian ini. Nilai tersebut menghasilkan nilai ketepatan prediksi seperti yang ditampilkan pada Tabel 5. Dengan akurasi 83.97% dan *f-measure* 84.56%, klasifikasi mampu mengklasifikasi dengan baik walaupun masih terdapat kesalahan prediksi seperti pada Tabel 6.



Cost	1	2	3	4	5	6	7	8	9	10	20	30	40	50	60	70	80	90	100
Accuration	83.81%	83.74%	83.84%	83.98%	83.98%	83.84%	83.91%	83.94%	83.87%	83.91%	83.84%	83.87%	83.87%	83.77%	83.77%	83.74%	83.77%	83.81%	83.77%
F-Measure	83.78%	83.73%	83.85%	84.56%	84.19%	84.00%	84.02%	84.08%	84.08%	84.17%	84.17%	84.17%	84.15%	84.03%	84.01%	83.98%	83.97%	84.02%	83.98%

Gambar 7 Uji akurasi f-measure dengan *grid search* data relevan

Tabel 5. Hasil evaluasi performa klasifikasi linear

Kategori	Precision	Recall	F1-Score	Support
Fitur	56.00%	25.45%	35.00%	55
Jaringan	94.43%	72.85%	82.25%	512
Jaringan 4G	97.71%	85.20%	91.03%	250
Penipuan	100%	03.03%	05.88%	33
Pertanyaan Umum	79.97%	97.43%	87.84%	1516
Program Khusus	92.31%	38.22%	54.05%	157
SSH-Netflix	87.43%	85.11%	86.25%	188
Value Added S.	80.95%	53.97%	64.76%	63
Voucher Isi Ulang	83.33%	81.97%	82.64%	122
AVG / TOTAL	85.15%	83.98%	82.53%	2896

4.2.2 Hasil Klasifikasi Kernel RBF dengan Stopword Removal, Special Stopword dan Stemming dengan Grid Search

Pada percobaan menggunakan kernel, klasifikasi menghasilkan akurasi lebih baik dengan nilai 84.84% dengan *F-measure* 84.88%. Parameter optimal berada pada nilai *cost = 10* dan *gamma =0.3017.* Tabel 7

menunjukkan hasil evaluasi performa klasifikasi RBF. Sedangkan Tabel 8 menampilkan *Confusion Matrix* dari *RBF Kernel*.

Tabel 6. Confusion matrix linear kernel

		Prediksi									Total Aktual
		A	B	C	D	E	F	G	H	I	
Aktual	A	14	0	0	0	40	1	0	0	0	55
	B	1	373	0	0	118	0	17	3	0	512
	C	0	4	213	0	28	0	0	3	2	250
	D	0	0	0	1	30	0	0	0	2	33
	E	3	15	0	0	1477	3	6	2	10	1516
	F	7	1	5	0	78	60	0	0	6	157
	G	0	0	0	0	28	0	160	0	0	188
	H	0	2	0	0	27	0	0	34	0	63
	I	0	0	0	0	21	1	0	0	100	122
Total Prediksi		25	395	218	1	1847	65	183	42	120	2896

Keterangan:
A = Fitur, B = Jaringan, C = Jaringan 4G, D = Penipuan, E = Pertanyaan Umum,
F = Program Khusus, G = SSH-Netflix, H = Value Added Service, I = Voucher dan Isi Ulang

Tabel 7. Hasil evaluasi performa klasifikasi RBF

Kategori	Precision	Recall	F1-score	Support
Fitur	58.54%	43.64%	50.00%	55
Jaringan	89.21%	77.54%	82.97%	512
Jaringan 4G	98.15%	84.80%	90.99%	250
Penipuan	30.00%	09.09%	13.95%	33
Pertanyaan Umum	82.05%	96.17%	88.55%	1516
Program Khusus	85.19%	43.95%	57.98%	157
SSH-Netflix	97.48%	82.45%	89.34%	188
Value Added Service	83.33%	55.56%	66.67%	63
Voucher & Isi Ulang	83.20%	85.25%	84.21%	122
AVG / TOTAL	84.91%	84.84%	83.93%	2896

Perbedaan hasil evaluasi performa klasifikasi pada kategori satu dengan lainnya terjadi karena nilai prediksi pada kategori lain lebih tinggi dibandingkan dengan nilai prediksi kategori aktualnya. Dengan kata lain bahwa *feature* pada kategori aktual menjadi *feature* pada kategori lain dengan proporsi yang lebih banyak.

Tabel 8. Confusion matrix RBF kernel

		Prediksi									Total Aktual
		A	B	C	D	E	F	G	H	I	
Aktual	A	24	1	0	3	25	2	0	0	0	55
	B	2	397	0	0	110	0	1	2	0	512
	C	0	5	212	0	27	1	0	3	2	250
	D	1	2	0	3	23	2	0	0	2	33
	E	8	25	0	3	1458	6	3	2	11	1516
	F	6	5	4	0	67	69	0	0	6	157
	G	0	6	0	0	27	0	155	0	0	188
	H	0	4	0	1	23	0	0	35	0	63
	I	0	0	0	0	17	1	0	0	104	122
Total Prediksi		41	445	216	10	1777	81	159	42	125	2896

Keterangan:
A = Fitur, B = Jaringan, C = Jaringan 4G, D = Penipuan, E = Pertanyaan Umum,
F = Program Khusus, G = SSH-Netflix, H = Value Added Service, I = Voucher dan Isi Ulang

4.3 Analisis Hasil Keseluruhan Percobaan

Hasil uji performa klasifikasi teks dengan menggunakan metode SVM meliputi pengujian akurasi, *precision*, *recall* dan *F-measure*. Tak hanya itu, optimasi parameter juga dilakukan dengan menggunakan fitur *grid search*.

Hasil perhitungan akurasi pada data tweet awal jika menggunakan parameter *default* maka kernel yang terbaik ialah kernel linear pada percobaan 3 dengan nilai *cost* 1 menghasilkan akurasi sebesar 98.79%. Percobaan 1 yakni klasifikasi dengan kernel linear menghasilkan akurasi 98.76% dan percobaan 2 yakni klasifiksi kernel linear dengan *special stopword* menghasilkan akurasi 98.66%. Sedangkan jika menggunakan kernel RBF dengan parameter *default* pula yakni nilai *cost = 1* dan nilai *gamma* = '*auto*' maka akurasi terbaik ialah pada percobaan 4 dan 6 dengan nilai akurasi 98.05%. Percobaan 5 menghasilkan akurasi 97.93%.

Hasil perhitungan akurasi pada data relevan jika menggunakan parameter *default* maka kernel yang terbaik ialah kernel linear pada percobaan 3 dengan nilai *cost* 1 menghasilkan akurasi sebesar 83.81%. Sedangkan jika menggunakan kernel RBF dengan parameter *default* pula yakni nilai *cost = 1* dan nilai *gamma* = '*auto*' maka akurasi terbaik ialah pada percobaan 6 dengan nilai akurasi 74.86%. Untuk percobaan tanpa menggunakan *special stopword* dan tanpa proses *stemming*, hasil akurasi menggunakan kernel linear dengan parameter *default* menghasilkan akurasi sebesar 81.35%. Jika menggunakan kernel RBF maka menghasilkan akurasi sebesar 65.16%. Sedangkan untuk percobaan melalui tahapan *praproses data*, *stopword removal*, *special stopword* tanpa menggunakan stemmer dengan menggunakan kernel linear memiliki akurasi sebesar 81.56%. Dengan menggunakan kernel RBF, hasil akurasi hanya mencapai 74.10% dengan parameter *default*.

Dari keenam percobaan, pada data tweet dipilih percobaan 3 untuk dioptimalkan parameternya. Untuk data relevan dipilih 2 percobaan yakni percobaan 3 dan 6 karena memiliki nilai akurasi tertinggi dari masing-

masing kernel dan juga sama-sama melalui proses *stemming.* Selanjutnya melakukan optimasi parameter terbaik menggunakan *grid search.*

Berdasarkan hasil *grid search,* data tweet pada percobaan 3 menghasilkan akurasi 98.85% dengan nilai *cost=10.* Data relevan pada percobaan 3 dengan kernel linear memiliki performa akurasi sebesar 83.98% dengan parameter *cost = 4.* Sedangkan percobaan 6 dengan kernel RBF menghasilkan performa klasifikasi sebesar 84.84% dengan pengaturan parameter yakni nilai *cost = 10* dan nilai *gamma = 0.3017.* Oleh karena itu, percobaan 6 merupakan percobaan yang memiliki performa terbaik dengan nilai *precision, recall* dan *F-measure* secara berurutan adalah 84.91%, 84.84% dan 83.93%.

Pada tahap uji validasi, pengklasifikasi teks menghasilkan beberapa kesalahan klasifikasi. Kesalahan klasifikasi ini dapat terjadi karena nilai prediksi pada kategori lain lebih tinggi dibandingkan dengan nilai prediksi kategori aktualnya. Nilai prediksi yang dimaksud didapatkan dari hasil perhitungan kata atau *feature* dalam suatu *tweet* berdasarkan model klasifikasi yang telah dibuat dari data pelatihan. Hal ini juga berarti bahwa *feature* pada kategori aktual menjadi *feature* pada kategori lain dengan proporsi yang lebih banyak. Kesalahan klasifikasi dapat diminimalkan dengan menambahkan data latih dengan fitur lebih representative dan menggunakan metode penggabungan *feature* seperti *bi-gram.*

Proses tidak menghapus *hashtag* dan menghapus semua angka yang ada pada data menimbulkan penurunan hasil klasifikasi. Selain itu, berdasarkan perbandingan kernel yang digunakan untuk klasifikasi teks, kernel terbaik pada penelitian tugas akhir ini adalah kernel RBF. Namun, terlepas menggunakan kernel apapun, proses klasifikasi teks akan lebih baik kinerjanya jika melalui proses *praproses data,* menghapus *stopword* dengan tambahan *special stopword* yang diperlukan serta menggunakan *stemmer.* Praproses data yang dimaksud ialah *case folding, filtering* yang meliputi proses menghapus *link url, punctuation, mention hashtag* dan *selection numbers* dan *tokenizing.*

5. Kesimpulan

5.1 Simpulan

Berdasarkan proses – proses pengerjaan yang telah dilakukan, dihasilkan beberapa kesimpulan yang dapat diambil, yaitu kesimpulan proses dan kesimpulan hasil, diantaranya sebagai berikut:

1) Pengklasifikasi teks ini telah berhasil mengklasifikasikan *tweet* sesuai dengan kategori myTelkomsel menggunakan metode SVM dengan performa yang baik. Berdasarkan 6 percobaan yang telah dilakukan, setiap percobaan memiliki performa SVM yang baik dalam mengklasifikasikan *tweet* melalui proses *stemming* maupun tidak. Jika dilihat dari uji performa, klasifikasi SVM menghasilkan akurasi terbaik ialah konfigurasi menggunakan kernel Linear melalui *stemming* dan *special stopword*
2) Selama penelitian menggunakan metode SVM, proses *stemming* mempengaruhi performa klasifikasi teks. Jika dibandingkan, setiap percobaan klasifikasi yang menggunakan *stemmer* akan memberikan nilai akurasi yang lebih tinggi dibandingkan percobaan tanpa *stemming.* Hal ini membuktikan bahwa proses *stemming* dapat meningkatkan performa klasifikasi teks dengan metode SVM.
3) Proses *not remove hashtag* dan *remove all number* memiliki pengaruh terhadap performa klasifikasi. Pengaruh yang diberikan merupakan pengaruh negatif karena performa klasifikasi mengalami penurunan setelah melakukan *remove all number* dan *not remove hashtag*. *Hashtag* bisa berupa inti *tweet* namun bisa juga sekedar kicauan. Pada klasifikasi teks ini, *hashtag* tidak terlalu berpengaruh sehingga untuk mendapatkan akurasi yang tinggi harus menghapus *hashtag*. Namun, jika semua angka dihilangkan maka pengklasifikasi teks ini tidak dapat menemukan fitur kata utama pada kategori Jaringan 4G, kategori Voucher dan Isi Ulang seperti fitur kata 4G, 100ribu, 1,3GB dan sebagainya.
4) Optimasi parameter dengan fitur *grid search* pada percobaan yang melalui proses *stemming* dan *special stopword* menggunakan kernel Linear dan RBF. Apabila dibandingkan, penggunaan kernel

RBF akan lebih baik daripada menggunakan kernel Linear setelah optimasi parameter dengan *grid search* dalam kasus klasifikasi teks (*tweet*) ini.

5.2 Saran

Berikut beberapa saran yang dapat dipertimbangan untuk pengembangan penelitian selanjutnya dan peningkatan pelayanan oleh pihak Telkomsel.

1) Pada penelitian ini, penulis menggunakan 2 jenis kernel yaitu kernel Linear dan RBF. Maka dari itu, diperlukan pengaturan kernel lain yang mungkin dapat meningkatkan performa klasifikasi.
2) Metode yang digunakan adalah SVM, untuk penelitian selanjutnya bisa menggunakan metode Naïve Bayes, Decision Tree, K-Nearest-Neighbourhood atau Artificial Neural Network sehingga dapat dibandingkan performa klasifikasi dengan SVM.
3) Diperlukan proses utnuk melakukan *rebalance* data dengan metode *Sampling* atau menggunakan pengaturan *cost* yang berbeda pada setiap kelas pada *soft margin classifier.*
4) Untuk penelitian selanjutnya, pembobotan tidak hanya menggunakan TF-IDF sehingga hasil menjadi bervariasi seperti menggunakan TF dibandingan dengan IDF, *information gain*, *chi square* dan lain-lain. Selain itu, penelitian ini menggunakan metode ekstraksi fitur dengan *n-gram* sehingga penelitian selanjutnya dapat menggunakan penggabungan *feature* seperti *bi-gram.*
5) Berdasarkan hasil penelitian yang dilakukan pada data *tweet* relevan periode 25 Januari hingga 7 Februari 2016, terdapat *frequent words* pada setiap kategori yang banyak diadukan oleh pengguna Telkomsel. Terdapat kata yang sering muncul pada kategori Jaringan ialah *koneksi.* Hal ini mengindikasikan bahwa *koneksi* harus menjadi prioritas utama bagi Telkomsel untuk meningkatkan pelayanan bagi pengguna.
6) Untuk meningkatkan pelayanan kepada pelanggan dalam menikmati layanan data, pihak Telkomsel dapat melakukan peningkatan kualitas layanan *mobile broadbrand* yang prima dengan optimasi jaringan. Optimasi jaringan *broadband* dilakukan pada seluruh kota di seluruh Indonesia untuk memastikan kecepatan akses layanan data dan stabilitas koneksi jaringan. Pihak Telkomsel dapat menambahkan BTS (*Base Transceiver Station*) baru dan kapasitas jaringan serta mengimplementasi HSPA+ (*High-Speed Packet Access*) sehingga kecepatan dapat meningkat hingga mencapai 42 Mbps.

6. Daftar Rujukan

[1] "Number of internet users in Indonesia from 2013 to 2018 (in millions)," Statista - The Statistics Portal, 2015. [Online]. Available: http://www.statista.com/statistics/254456/number-of-internet-users-in-indonesia/. [Accessed 4 Oktober 2015].
[2] APJII, "Profil Pengguna Internet Indonesia 2014," Asosiasi Penyedia Jasa Internet Indonesia, Jakarta, Maret 2015.
[3] "Blackberry Messenger, Aplikasi Chat Paling Banyak Dipilih DI Indonesia," Neilsen, 6 September 2014. [Online]. Available: http://www.nielsen.com/id/en/press-room/2014/blackberry-messenger-aplikasi-chat-paling-banyak-dipilih-di-indonesia.html. [Accessed 12 Mei 2015].
[4] Telkom, "Data Perusahaan," [Online]. Available: http://www.telkom.co.id/UHI/assets/pdf/ID/09_Data%20Perusahaan.pdf. [Accessed 5 September 2015].
[5] "my Telkomsel," Telkomsel, [Online]. Available: https://my.telkomsel.com/GTConnect/index.jsp. [Accessed 3 Mei 2015].
[6] "Customer Service Link eCare," Telkomsel, [Online]. Available: http://www.telkomsel.com/customer-service/ecare. [Accessed 6 Mei 2014].
[7] S. Millward, "Indonesia is Social: 2.4% of World's Twitter Posts Come From Jakarta (INFOGRAPHIC)," Techinasia, 13 Maret 2013. [Online]. Available: https://www.techinasia.com/indonesia-social-jakarta-infographic/. [Accessed 7 Oktober 2015].
[8] E. E. Pratama and B. R. Trilaksono, "Klasifikasi Keluhan Pelanggan Berdasarkan Tweet dengan Menggunakan Metode Support Vector Machine (SVM)," Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung, Bandung, 2014.
[9] C. Trianawati, Metode Pembobotan Statictical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia, 2009.
[10] I. V. P. d. G. A. Hemalatha, " Preprocessing the Informal Text for Efficient Sentiment," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), vol. I, no. ISSN 2278-6856, 2012.
[11] E. Nugroho, "Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoritma Rabin-Karpi," Program Studi Ilmu Komputer, Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan, Universitas Brawijaya Malang, 2011.
[12] V. V., The Nature of Statistical Learning Theory, New York: Springer-Verlag, 1995.
[13] N. Cristianini and J. S. Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge: Cambridge University Press, 2000.

[14] S.-l. Developers, "Support Vector Machine," Scikit Learn, 2014. [Online]. Available: http://scikit-learn.org/stable/modules/svm.html. [Accessed 19 Mei 2016].
[15] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "A Practical Guide to Support Vector Classification," Department of Computer Science National Taiwan University, Taipei, Taiwan, 2016.
[16] "The Streaming APIs Overview," Dev Twitter, Inc., 2016. [Online]. Available: https://dev.twitter.com/streaming/overview. [Accessed 15 Mei 2016].
[17] R. P. Kusumawardani, "Machine Learning Taks: Classification and Beyond for Sistem Cerdas Course," Jurusan Sistem Informasi ITS, Surabaya, 2014.
[18] S. Kusumadewi, Artificial Intellegence (Teknik dan Aplikasinya), Yogyakarta: Graha Ilmu, 2003.
[19] P. Flach, Machine Learning: The Art and Science of Algorithms that Makes Sense of Data, New York: Cambridge University Press, 2012.
[20] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," Master of Logic Project Institute for Logic, Language and Computation, Universiteit van Amsterdam The Netherlands, 2003.