ISSN 2580-2801

JUS TEKNO
Jurnal Sains & Teknologi

SUPPORT VECTOR MACHINE BERBASIS GRID SEARCH ALGORITHM UNTUK MENINGKATKAN AKURASI PENENTUAN NILAI SENTIMEN PADA TWITTER (Studi Kasus Pada Dataset Sanders)

Dedi Wirasasmita

Program Studi Teknik Informatika
Sekolah Tinggi Teknologi Duta Bangsa
dedi.wiro2013@gmail.com url: http://dediwssttdb.wordpress.com

Abstrak

Twitter dapat mengekspresikan opini yang objektif tentang topik yang berbeda, serta dapat membatu dalam bidang pemasaran untuk menyediakan opini terhadap konsumen mengenai merk dan produk yang populer. Dataset twitter Sanders adalah data yang sudah diakuisisi berupa empat kategori merk atau produk @appel, @google, @mcrosoft, dan @twitter.
Pada penelitian ini akan dilakukan pengklasifikasian data *tweet Sanders* untuk review ke empat kata dari data sanders yang paling populer menggunakan metode *support vector machine* (SVM) berbasis Grid Search Algorithm (GSA), agar *tweet* yang ada tidak bercampur antara tweet negative, positive dan netral. SVM salah satu metode yang dapat melakukan pengklasifikasi data dengan baik, karena proses yang akan dilakukan bersifat *non linier* maka parameter yang akan digunakan adalah nilai C dan γ. Dari hasil uji coba, aplikasi menunjukkan akurasi stabil Dengan menggunakan penambahan metode *preprocessing* akurasi dapat mencapai hingga 69,22%. Dengan metode *stopwordremoval* akurasi dapat mencapai hingga 73,74%. Serta metode *Grid Search dan Two Step GridSearch* akurasi dapat mencapai hingga 73,89%. Sedangkan keseluruhan optimasi tersebut dapat memberikan hasil dengan akurasi hingga 80,33%. Dengan pencapaian nilai akurasi yang baik maka, hasil ini dapat diterapkan untuk membantu pengguna *twitter* untuk melakukan filter terhadap *tweet* informasi yang dibutuhkan yang terdapat pada akun Twitter mereka.

Kata Kunci: Klasifikasi, , *Support Vector Machine, Grid Search Algorithm, Tweet,* Twitter *akurasi*.

1. PENDAHULUAN

Twitter adalah layanan jejaring sosial yang mengalami pertumbuhan sangat pesat dan dengan cepat meraih popularitas di seluruh dunia [Karen Wickre,2013]. Twitter dapat mengekspresikan opini yang objektif tentang topik yang berbeda, serta dapat membatu dalam bidang pemasaran untuk menyediakan opini terhadap konsumen mengenai merk dan produk yang populer. [B.j. Jansen, 2009]. Sentimen mempengaruhi penjualan volume produk atau jasa sampai batas tertentu. perusahaan melihat bahwa deteksi sentimen tentang produk tertentu bisa mempengaruhi daya saing perusahaan, di mana mereka menangkap peluang, kelemahan dan ancaman (Serrano-Guerrero, Olivas, Romero, & Herrera-Viedma 2015; Wijnhoven & Bloemen 2014).

Analisis sentimen bertujuan untuk mengetahui pendapat, emosi, dan sikap Yang berasal dari berbagai sumber seperti dokumen, teks pendek, kalimat dari ulasan, [P.D.Turney, B.Pang, 2003], blog [M.Hu, P.He, 2008,2009], berita [Balahur, 2010]. Secara umum Ada dua pendekatan utama untuk melakukan analisis sentimen - machine learning base dan Lexicon base [Medhat, Hassan, & Korashy 2014; Serrano Guerrero et al., 2015]. pembelajaran mesin didasarkan pada seleksi dan ekstraksi set sesuai fitur yang digunakan untuk mendeteksi sentimen [Medhat et al., 2014]. Sementara Lexicon base bergantung pada pra-definisikan kamus leksikon dan / atau dataset [Kang & Park, 2014].

Pada dasarnya analisis sentimen merupakan klasifikasi, tetapi kenyataannya tidak semudah proses klasifikasi biasa karena terkait penggunaan bahasa. [Bing Liu. 2012]. Analisis sentimen atau opinion mining dapat digunakan untuk memperoleh gambaran umum persepsi masyarakat terhadap kualitas layanan, apakah cenderung positif, negatif atau netral. Sentimen biasanya bernilai positif atau negatif tetapi dapat dikategorisasikan juga menjadi baik, sangat baik, buruk, dan sangat buruk. [Shelby.M.I, 2013].
Sebelumnya telah dilakukan beberapa penelitian analisis sentimen terhadap Twitter dengan berbagai metode diantaranya:

- Naïve Bayes (NB), Decission Tree (DT), k-Nearest Neighbor (KNN) pada bahasa Urdu. [Muhammad Bilal, 2015],
- Support Vector Machine (SVM), NB, Maximun Entropy (ME) pada stock harga dan korelasinya. [Linhao Zhang, 2013],
- NB, SVM dan DT pada Reputasi Merk Mobile Phone Provider [Nur Azizah Vidya, 2015]
- NB terhadap Calon Presiden 2014. [Faishol , 2013].

NB mempunyai kelebihan kokoh terhadap atribut yang tidak relevan serta Cepat dan Efesiensi ruang. [Gholamreza, 2007] Tapi memiki kelemahan Harus mengasumsi bahwa antar fitur tidak terkait (independent) alam realita, keterkaitan itu ada dan Keterkaitan tersebut tidak dapatdimodelkan oleh NB. [Dyarsa, 2014] ,[Taufik Fuadi Abidin, 2013]. DT memiliki kelebihan mudah dalam proses pengelompokan data teks, relatif transfaran, mudah dimengerti dan memiliki sejumlah kualitas yang tinggi. Tapi memiliki kelemahan mengeksploitasi fitur yang relatif independen satu sama lain sehingga membatasi kemampuan DT [Mitchell, 1996]. KNN memiki kelebihan mudah direpresentasikan,memiliki ketangguhan terhadap training data yangmemiliki banyak noise, dan cukup efektif untuk proses pengelompokan. Tapi memliki kelemahan pada akurasi, karena nilai k ditetapkan sama pada semua kategori tanpa memperhitungkan jumlah dokumen yang dimiliki masing-masing kategori. [TSK, 2006]. SVM mempunyai kelebihan kinerja dalam hasil eksperimen sangat baik serta tingkat ketergantungan rendah pada dataset yang berdimensi. Tapi memiliki kelemahan pada pengkategorian atribut yang kurang optimal jika ada missing value dan memerlukan proses ulang . [Pravesh Kumar Singh, 2014].

Didapatkan kesimpulan bahwa penggunaan metode SVM memberikan hasil paling baik dibandingkan metode lainnya, yaitu dengan tingkat keakuratan lebih tinggi. [A. Go, R. Bhayani, and L. Huang, Sentiment Classification using, 2009]. Tetapi SVM memiliki kelemahan pada sulitnya pemilihan fitur yang sesuai dan optimal pada bobot atribut yang digunakan sehingga menyebabkan tingkat akurasi menjadi rendah. [Pravesh Kumar Singh, 2014]. metode Grid Search (GS) adalah metode sebagai pemilihan fitur yang terbukti efektif yang bisa membantu pemilihan fitur pada metode klasifikasi seperti SVM. [Abdul Razak Naufal, 2015], maka pada penelitian ini GS akan diterapkan untuk pemilihan fitur SVM yang sesuai dan optimal, sehingga hasil analisis nilai sentimen lebih akurat.

2. TINJAUAN PUSTAKA DAN TEORI

Dari beberapa kajian penelitian sebelumnya yang sudah dijelaskan diatas, maka dalam penelitian ini akan dibuatkan table perbandingan kajian penelitian sebelumnya berdasarkan akurasi pada Metode machine Learning yang digunakan.

Peneliti	Algoritma	Fitur	Domain	Dataset	Hasil
(Linhao Zhang, 2013)	• NB • Maximum Entropy • SVM	• Unigram • N-gram • External lexion • Part of speech tagging • Neutral label • Chi square • TF-IDF •	Stock harga	Acuisisi dataset dari Twitter API dengan python menhasilkan MongoDB	Akurasi : • NB= 0.8173 • Maximum Entropy=0.8058 • SVM=0.8374
(Pravesh Kumar	• NB • SVM	• N-gram	• Product • Film	• Produk amazon.com	Akurasi : • NB=0.799 % • SVM=82.9 %

Singh, 2014)	• Multi-Layer Perceptron (MLP) • Clustering			• Film review dari pang & lee (2004)	• MLP=83.25% • Clustering = 65.33%
(M. Bilal, Huma Israr, M. Shahid, Amin Khan, 2015)	• NB • DT • KNN	• StringToWordvector filter • Reorder filter • Numeric to binary filter • Bag of Words Mode	Opini bahasa urdu	ditulis dalam Roman-Urdu diekstrak dari Blog menggunakan Easy software web Extractor	Akurasi : • NB= 82.33% • DT=81.67% • KNN=80.67%
(Nur Azizah Vidya, 2015)	• NB • SVM • D_Tree	• *Part-Of-Speech tagging*	Reputasi Merk pada Mobile Phone Provider	Melalui Acuisisi Twitter API	Akurasi : • NB= 0.8373 • SVM=0.8374 • D_TREE=0.840
(Dinar Ajeng, 2015)	• SVM	• *PSO*	Review Produk kosmetik	Melalui www.amazon.com	Akurasi : 82.00%

Tabel 1 Kajian penelitian sebelumnya berdasarkan metode machine Learning

Twiiter

Twitter adalah sebuah situs web yang dimiliki dan dioperasikan oleh Twitter Inc., yang menawarkan jaringan sosial berupa mikroblog sehingga memungkinkan penggunanya untuk mengirim dan membaca pesan *Tweets* (Twitter, 2013).

Analisis Sentimen

Analisis sentimen adalah proses yang bertujuan untuk memenentukan isi dari dataset yang berbentuk teks (dokumen, kalimat, paragraf, dll) bersifat positif, negative atau netral (Kontopoulos 2013)

Text Mining

Text mining dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang merupakan komponen-komponen dalam data mining.

Pre-processing

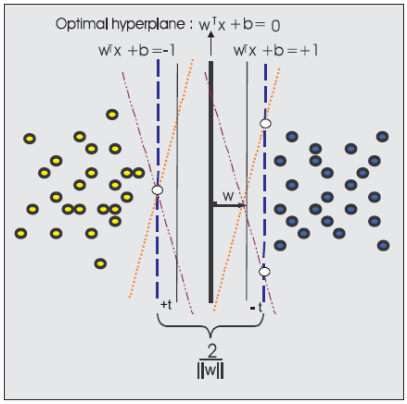
Preprocessing merupakan tahapan awal dalam mengolah data input sebelum memasuki proses tahapan utama dari metode *Machine learning*. *Preprocessing text* dilakukan untuk tujuan penyeragaman dan kemudahan pembacaan serta proses Machine Learning selanjutnya (Aji P., Baizal SSi. and Firdaus S.T., 2011).

Algoritma Support Vector Machine

Support Vector Machine (SVM) yaitu metode yang mencari fungsi pemisah (hyperplane) terbaik untuk memisahkan data-data dengan kelas-kelas yang berbeda. (Nugroho 2008).

SVM adalah seperangkat metode yang terkait untuk suatu metode pembelajaran, untuk kedua masalah klasifikasi dan regresi. Dengan berorientasi pada tugas, kuat, sifat komputasi yang mudah dikerjakan, SVM telah mencapai sukses besar dan dianggap sebagai *state of the art classifier* saat ini (Huang, 2008)

Secara sederhana konsep SVM adalah sebagai usaha mencari *hyperlane* terbaik yang berfungsi sebagai pemisah dua buah *class* pada *input space*, dimana dapat dilihat pada gambar dibawah ini:



Gambar 1. Konsep SVM untuk mencari *hyperlane* terbaik (Santosa, 2007)

Pada gambar diatas memperlihatkan beberapa *pattern* yang merupakan anggota dari dua buah *class*: +1 dan -1. *Pattern* yang tergabung pada *class* -1 disimbolkan dengan warna kuning. Sedangkan *pattern* pada *class* +1, disimbolkan dengan warna biru. Problem klasifikasi dapat diterjemahkan dengan usaha menemukan garis (*hyperplane*) yang memisahkan antara kedua kelompok tersebut. Berbagai alternatif garis pemisah (*discrimination boundaries*) ditunjukkan garis berwarna orange. *Hyperplane* pemisah terbaik antara kedua *class* dapat ditemukan dengan mengukur margin *hyperplane* tersebut. dan mencari titik maksimalnya.

Margin adalah jarak antara *hyperplane* tersebut dengan *pattern* terdekat dari masing-masing *class. Pattern* yang paling dekat ini disebut sebagai *support vector*. *Hyperplane* yang terbaik yaitu yang terletak tepat pada tengah-tengah kedua *class,* sedangkan titik putih yang berada dalam garis bidang pembatas *class* adalah *support vector*. Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari proses pembelajaran pada SVM.

Data yang tersedia dinotasikan sebagai $x \in R^d$ sedangkan label masing masing dinotasikan $y_i \in \{-1+1\}$ untuk i = 1,2,....,1 yang mana l adalah banyaknya data. Diasumsikan kedua class –1 dan +1 dapat terpisah secara sempurna oleh hyperplane berdimensi d , yang didefinisikan:

- $w.x + b = 0$ (2.1)

Sebuah pattern xi yang termasuk class –1 (sampel negatif) dapat dirumuskan sebagai pattern yang memenuhi pertidaksamaan:

- $w.x + b = -1$ (2.2)

sedangkan pattern yang termasuk class +1 (sampel positif):

- $w.x + b = +1$ (2.3)

Margin terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara hyperplane dan titik terdekatnya, yaitu 1/||w||. Hal ini dapat dirumuskan sebagai Quadratic Programming (QP) problem, yaitu mencari titik minimal persamaan 2.4, dengan memperhatikan constraint persamaan 2.5.

$$\min t(w) = \frac{1}{2}||w||^2 \quad (2.4)$$

$$y_i (x_i.w+b) -1 \geq 0, \quad \forall i \quad (2.5)$$

Problem ini dapat dipecahkan dengan berbagai teknik komputasi, diantaranya Lagrange Multiplier sebagaimana ditunjukkan pada persamaan 2.6:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 w\sum_{i=1} \alpha i(yi((xi.w+b)-1)) \quad (i=1,2,\ldots,1) \quad (2.6)$$

αi adalah Lagrange multipliers, yang bernilai nol atau positif ($\alpha i \geq 0$). Nilai optimal dari persamaan (2.6) dapat dihitung dengan meminimalkan L terhadap w dan b , dan memaksimalkan L terhadap αi. Dengan memperhatikan sifat bahwa pada titik optimal gradient L =0, persamaan langkah 2.6 dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung saja αi, sebagaimana persamaan 2.7.

Maximize: $\sum_{i=1} \alpha i - \frac{1}{2} \sum_{i,j=1} \alpha i\, \alpha j y_i.y_j,\ x_i,x_j$ (2.7)

Subject to:

$\alpha i \geq 0$ (i=1,2,…,1) $\sum_{i=1}^{1} \alpha i y_i = 0$ (2.8)

Dari hasil dari perhitungan ini diperoleh αi yang kebanyakan bernilai positif. Data yang berkorelasi dengan αi yang positif inilah yang disebut sebagai *support vector.*

Sebagai contoh digunakan problem AND. Problem AND adalah klasifikasi dua kelas dengan empat data (lihat Tabel 2.1). Karena ini problem linier, kernelisasi tidak diperlukan.

Table 2. AND Problem

X1	X2	y
1	1	1
-1	1	-1
1	-1	-1
-1	-1	-1

dapatkan formulasi masalah optimisasi sebagai berikut:

min $\frac{1}{2}$ ($w_1^2 + w_2^2$) + C($t_1 + t_2 + t_3 + t_4$)
Subject to:

w1 + w2 + b + t1 = 1
w1 - w2 - b + t2 = 1
-w1 + w2 - b + t3 = 1
w1 + w2 - b + t4 = 1
t1, t2, t3, t4 = 0

Karena fungsi AND adalah kasus klasifikasi linier, maka bisa dipastikan nilai variable slack ti = 0. Jadi Kita bisa masukkan nilai C = 0. Setelah menyelesaikan problem optimasi di atas didapat solusi w1 = 1,w2 = 1, b = -1
Persamaan fungsi pemisahnya adalah f(x) = x1 + x2 - 1.

Untuk menentukan output atau label dari setiap titik data/obyek kita gunakan fungsi g(x) = sign(x). Dengan fungsi sign ini semua nilai f(x) < 0 diberi label -1 dan lainnya diberi label +1.

Pemilihan Fitur (*Feature Selection*)

Seleksi fitur adalah salah satu faktor yang paling penting yang dapat mempengaruhi tingkat akurasi klasifikasi karena jika dataset berisi sejumlah fitur, dimensi ruang akan menjadi besar, merendahkan tingkat akurasi klasifikasi (Liu, 2011).

Algoritma Pemilihan Fitur Grid Search Algorthm (GSA)

Grid-search adalah salah satu prosedur pemilihan fitur model yang direkomendasikan pada metode classification, karena untuk melakukan grid-search secara lengkap membutuhkan waktu yang lama, maka proses ini dapat dilakukan dalam dua tahap (two-step grid search) (C.-W. Hsu, 2013), yaitu:

- Pertama-tama lakukan loose grid-search dengan menggunakan interval yang cukup besar, missal 2^2, yaitu:
  C = 2^{-5}, 2^{-3}, …, 2^{15} dan © = 2^{-15}, 2^{-13}, …, 2^3

- Setelah diketahui daerah „kandidat“, lakukan finer grid search pada sekitar daerah “kandidat“ tersebut dengan menggunakan interval yang lebih kecil, misal $2^{0.25}$

- K-fold cross-Validation adalah prosedur yang direkomendasikan untuk mengestimasi akurasi model (C.-W. Hsu, 2013), yaitu:

1. Bagi data menjadi k bagian dengan ukuran yang sama
2. Gunakan satu bagian sebagai data testing, sementara k-1bagian lainnya sebagai data training
3. Ulangi sebanyak k iterasi sehingga semua bagian pernah menjadi data testing

- Jumlah fold yang umum digunakan adalah k = 10
- Stratified adalah proses untuk membuat masing-masing fold terdiri dari data dengan jumlah kelas yang seimbang
- Model selection dapat dilakukan dengan menggunakan dua metode pencarian berikut ini:
- GridSearchCV, yaitu pencarian nilai akurasi terbaik pada semua kemungkinan nilai parameter
- RandomizedSearchCV, yaitu percarian nilai akurasi terbaik secara acak pada beberapa kemungkinan nilai parameter. Metode ini lebih cepat, akan tetapi belum tentu mendapatkan nilai akurasi terbaik dari semua kemungkinan nilai parameter.

Aplikasi Bahasa Python

Python merupakan bahasa pemrograman yang freeware atau perangkat bebas dalam arti sebenarnya, tidak ada batasan dalam penyalinannya atau mendistribusikannya. Lengkap dengan source codenya, debugger dan profiler, antarmuka yang terkandung di dalamnya untuk pelayanan antarmuka, fungsi sistem, GUI (antarmuka pengguna grafis), dan basis datanya.

Python Menggunakan Enthought Canopy

Untuk menulis program, dibutuhkan lingkungan pemprograman untuk meningkatkan produktifitas. Begitu juga Python, Enthought Canopy membawa paket terintegrasi distribusi berbagai tool Python. Enthought Canopy menyediakan *one-click Python installation*, *user-friendly Package Manager*, *integrated analysis environment* yang menyediakan 100+ paket Python termasuk di dalamnya paket utama *scientific* dan *analytic* seperti **NumPy, SciPy, Pandas, Matplotlib, IPython** dan lain sebagainya.

3. METODOLOGI PENELITIAN

Menurut (Dawson, 2009) ada empat metode penelitian yang umum digunakan yaitu tindakan penelitian, eksperimen, studi kasus dan survey. Dalam konteks penelitian, metode yang dilakukan mengacu kepada pemecahan masala yang meliputi mengumpulkan data, merumuskan hipotesis atau proposisi, pengujian hipotesis, menafsirkan hasil, dan kesimpulan (Berndtssom, Hansson, Olsson, & Lundell, 2008).

Dalam penelitian ini dilakukan beberapa langkah yang dilakukan dalam proses penelitian.

1. Pengumpulan data

Pada tahap ini ditentukan data yang akan diproses. Mencari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan semua data kedalam data set, termasuk variabel yang diperlukan dalam proses.

2. Pengolahan data awal

Ditahap ini dilakukan penyeleksian data, data dibersihkan dan ditransformasikan kebentuk yang diinginkan sehingga dapat dilakukan persiapan dalam pembuatan model.

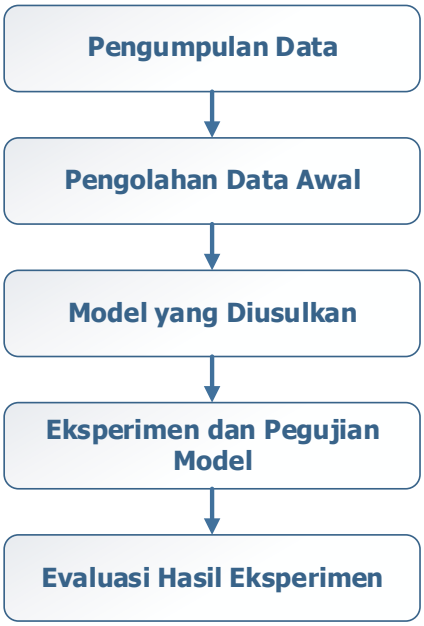
3. Metode yang diusulkan

Pada tahap ini data dianalisis, dikelompokan variabel mana yang berhubungan dengan satu sama lainnya. Setelah data dianalisis lalu diterapkan model-model yang sesuai dengan jenis data. Pembagian data kedalam data latihan (*training data*) dan data uji (*testing data*) juga diperlukan untuk pembuatan model.

4. Eksperimen dan pengujian metode

Pada tahap ini model yang diusulkan akan diuji untuk melihat hasil berupa *rule* yang akan dimanfaatkan dalam pengambilan keputusan.

5. Evaluasi dan validasi



Gambar 1. Langkah langkah penelitian

4. HASIL PENELITIAN DAN PEMBAHASAN

Nilai akurasi sentiment twitter dalam penelitian ini ditentukan dengan cara melakukan uji coba memasukkan C, epsilon. Pengujian ini dilakukan dengan metode SVM terhadap nilai akurasi ***Weigthing, Filtering, Preprosessing(skala),*** serta ***selection feature*** terhadap dataset yang sudah disediakan diantaranya STS-Test, STS_Gold, Sanders, dan Capres2014. Tools yang digunakan dalam pengujian ini menggunakan Bahasa Pemrograman Python Canopy. Berikut ini adalah hasil dari percobaan yang telah dilakukan untuk penentuan nilai akurasi sentiment.

Untuk loading data menggunakan Dataset STS-Test disimpan dalam format csv, sehingga untuk membaca data tersebut menggunakan pustaka csv:

```
# Data Loading
print "Data Loading ..."
data_file = csv.reader(open("D:\kuliah\KULIAH S2\data minning\DATA/STS-Test.csv"))
data_file.next()
for column in data_file:
data.append(column[5])
target.append(column[0])
```

4.1. Hasil Pengujian Weighting

Weighting, Yaitu Proses Pembobotan Masing-Masing Token Yang Dapat Berupa Term Count, Term Frequency (Tf), Atau Term Frequency Inversed Document Frequency (Tfidf).

```
import csv
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

# Feature Extraction
# -----------------------
#vectorizer = CountVectorizer()
#vectorizer = TfidfVectorizer(use_idf=False)
#vectorizer = TfidfVectorizer()
vectorizer = TfidfVectorizer(stop_words='english')
#vectorizer = TfidfVectorizer(min_df=2)

X = vectorizer.fit_transform(data)
t = np.asarray(target,dtype=np.int)
```

Hasil akurasi Term Count :

```
In [1]:

In [2]: %run "D:\kuliah\KULIAH S2\data minning\Program Python\Program Python\sk-AnalisisSentimen-Svm.py"
Data Loading ...
Learning ...
Nilai C          : 9.51365692002
Nilai gamma      : 0.011048543456
Estimasi akurasi : 0.659959758551
```

Hasil akurasi Term TF :

```
In [3]: %run "D:\kuliah\KULIAH S2\data minning\Program Python\Program Python\sk-AnalisisSentimen-Svm.py"
Data Loading ...
Learning ...
Nilai C          : 92681.9000237
Nilai gamma      : 0.001381067932
Estimasi akurasi : 0.665995975855

In [4]:
```

Hasil akurasi Term TFIDF :

```
In [4]: %run "D:\kuliah\KULIAH S2\data minning\Program Python\Program Python\sk-AnalisisSentimen-Svm.py"
Data Loading ...
Learning ...
Nilai C          : 27554.493735
Nilai gamma      : 0.25
Estimasi akurasi : 0.661971830986

In [5]:
```

Sehingga dari hasil pengujian pada Metode SVM pada proses Weighting dapat dilihat pada tabel berikut :

	Dataset STS-Test		
Term	Count	TF	TFIDF
Akurasi	65.99%	66.60%	68.20%

4.1.2. Hasil Pengujian Filtering

Filtering, yaitu proses penyaringan fitur yang potensial. Misal Token yang bukan stopwords, token yang bukan merupakan noise, dll serta menggunakan weighting TFIDF.

Hasil Akurasi Filtering

```
In [8]: %run "D:\kuliah\KULIAH S2\data minning\Program Python\Program Python\sk-AnalisisSentimen-Svm.py"
Data Loading ...
Learning ...
Nilai C          : 38967.9387444
Nilai gamma      : 1.52587890625e-05
Estimasi akurasi : 0.682092555332

In [9]: %run "D:\kuliah\KULIAH S2\data minning\Program Python\Program Python\sk-AnalisisSentimen-Svm.py"
Data Loading ...
Learning ...
Nilai C          : 8.0
Nilai gamma      : 0.0525560259534
Estimasi akurasi : 0.69416498994

In [10]: %run "D:\kuliah\KULIAH S2\data minning\Program Python\Program Python\sk-AnalisisSentimen-Svm.py"
Data Loading ...
Learning ...
Nilai C          : 1024.0
Nilai gamma      : 0.000410593952761
Estimasi akurasi : 0.66800804829
```

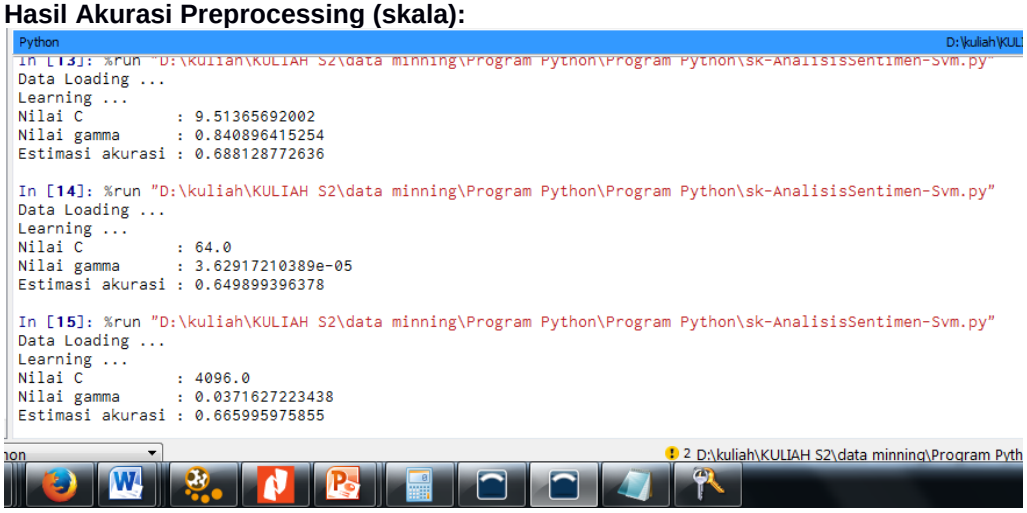
Sehingga dari hasil pengujian pada Metode SVM pada proses Filtering dapat dilihat pada tabel berikut :

	Dataset STS-Test		
Term	TFIDF	TFIDF	TFIDF
Filtering	None	Stopword	Min df=2
Akurasi	68.20%	69.41%	66.80%

4.1.3. Hasil Pengujian Preprocessing (Skala)

Preprocessing adalah pra pengolahan data, misal normalisasi, standarisasi, reduksi dimensi. Pada metode machine learning tertentu, nilai fitur data diharapkan berada pada suatu skala tententu, seperti bernilai [0,1], [-1,1], atau memiliki mean nol dan standar deviasi satu.
Pada pengujian Preprocessing ini proses yang digunakan pada Weighting menggunakan TFIDF, proses Filtering menggunakan Stopword dan Skala yang digunakan adalah standar, Stddev=1 dan Min=0 Max=1

Hasil Akurasi Preprocessing (skala):



Sehingga dari hasil pengujian pada Metode SVM pada proses Preprocessing dapat dilihat pada tabel berikut :

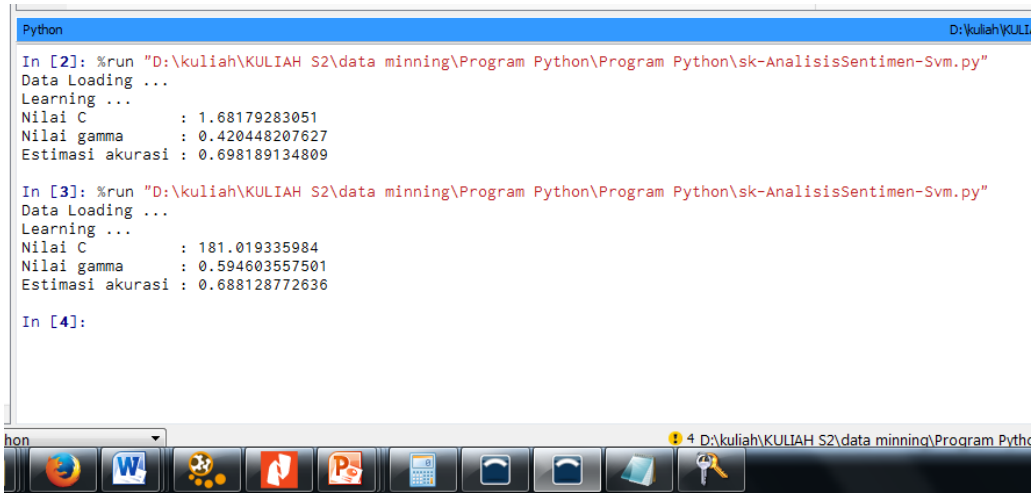
	Dataset STS-Test		
Term	TFIDF	TFIDF	TFIDF
Filtering	Stopword	Stopword	Stopword

Skala	None	Stddev=1	Min=0 Max=1
Akurasi	68.81%	64.98%	66.59%

4.1.4. Hasil Pengujian Model Selection

Model Selection: penentuan parameter model yang tidak dipelajari secara langsung pada tahapan learning, seperti orde polinomial M dan parameter regularisasi λ.
Pada pengujian model seleksi yang digunakan adalah untuk Weighting menggunakan TFIDF, Filtering menggunakan Stopword dan model seleksi yang digunakan adalah Grid Serach dan Two Step Grid Search.

Hasil akurasi model seleksi:



Sehingga dari hasil pengujian pada Metode SVM pada proses model seleksi fitur dapat dilihat pada tabel berikut :

Dataset STS-Test		
Term	TFIDF	TFIDF
Filtering	Stopword	Stopword
Model seleksi	Grid Search	Two Step GS
Akurasi	69.98%	68.81%

4.1.5. Hasil Pengujian Metode klasifikasi

Pada hasil pengujian nilai sentiment twitter berdasarkan dataset STS-Tes diantaranya menggunakan SVM yang dikomparasi dengan metode Logistic Regression dan Naïve Bayes adalah sebagai berikut:
%run "D:\kuliah\KULIAH S2\data minning\Program Python\Program Python\sk-AnalisisSentimen-Lreg.py"
Data Loading ...
Learning ...
Nilai C : 2.0
Estimasi akurasi : 0.682092555332

%run "D:\kuliah\KULIAH S2\data minning\Program Python\Program Python\sk-AnalisisSentimen-Nbayes.py"

Data Loading ...
Learning ...
Estimasi akurasi : 0.585231492597

%run "D:\kuliah\KULIAH S2\data minning\Program Python\Program Python\sk-AnalisisSentimen-Svm.py"
Data Loading ...
Learning ...
Nilai C : 1448.15468787
Estimasi akurasi : 0.686116700201
Sehingga dari hasil pengujian pada Metode SVM pada proses klasifikasi yang dibandingkan dengan metode Logistic Regresion serta Naïve Bayes dapat dilihat pada tabel berikut :

	Dataset STS-Test		
Metode	Log. Reg	Naïve Bayes	SVM
Akurasi	68.20%	58.52%	70.80%

5. KESIMPULAN DAN SARAN

KESIMPULAN

Berdasarkan hasil penelitian yang telah didapat, maka dapat ditarik beberapa simpulan antara lain:

- Dengan menggunakan 497 data *training* yang terdiri dari 177 negatif, 182 positip dan 139 netral. Disimpan dalam format csv dan terdiri dari 6 field: 0 : polarity (0 = negative, 2 = netral, 4 = positive), 1 : id. 2 tanggal. 3 : query, 4 : user, 5 : teks dari tweets untuk mengklasifikasi data *testing* yang terdiri dari kfold cross = 10 maka SVM berbasis GSA memberikan hasil dengan akurasi hingga 70,30%.
- Dengan menggunakan penambahan metode *preprocessing* akurasi dapat mencapai hingga 69,22%. Dengan metode *stopwordremoval* akurasi dapat mencapai hingga 73,74%. Dengan metode *Grid Search dan Two Step GridSearch* akurasi dapat mencapai hingga 73,89%. Sedangkan keseluruhan optimasi tersebut dapat memberikan hasil dengan akurasi hingga 80,33%.

Saran

Dari hasil pengujian yang telah dilakukan dan hasil kesimpulan yang diberikan maka ada saran atau usul yang di berikan antara lain:

1. Untuk meningkatkan hasil optimasi dapat dilakukan metode pemilihan parameter dengan metode *Genetic Algorithm* dan lain-lain.
2. Untuk mengetahui kehandalan metode maka pada penelitian selanjutnya dapat dilakukan penggunaan data set lebih dari satu.
3. Mencoba menerapkan metode optimasi yang lain sebagai bahan perbandingan, misalkan dengan metode *Neural Network, C4.5* dan *K- Nearest Neighbor*.

DAFTAR PUSTAKA

- Alexandra Balahur, Ralf Steinberger, (2010), Sentiment Analysis in the News: arXiv preprint arXiv:1309.6202

- B. Liu. (2010) *Handbook of Natural Language Processing, chapter Sentimen Analysis, 2nd Edition.*
- Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J.(2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Engineering, 53, 453–462. doi:10.1016/j.proeng.2013.02.059.
- Bernard J. Jansen, (2009), Twitter Power: Tweets As Electronic Word Of Mouth, Journal Of The American Society For Information Science And Technology Volume 60, Issue 11, Pages 2169–2188
- Bing Liu. (2012), Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers.
- Chou, J.-S., Cheng, M.-Y., Wu, Y.-W., & Pham, A.-D. (2014). Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification. Expert Systems with Applications, 41(8), 3955–3964. doi:10.1016/j.eswa.2013.12.035.
- Dave, Kushal; Lawrence, Steve; Pennock, David M. (2003) Mining the peanut gallery: Opinion Extraction and semancitc classification of product reviews.
- Dyarsa Singgih Pamungkas, (2015), Analisis Sentiment Pada Sosial Media Twitter Menggunakan Naive Bayes Classifier Terhadap Kata Kunci "Kurikulum 2013",*Techno.COM, Vol. 14, No. 4, November 2015: 299-314*.
- Faishol Nurhuda, (2013), Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari witter Menggunakan Metode Naive Bayes Classifier, JURNAL ITSMART Vol 2. No 2. Desember 2013 ISSN : 2301–7201
- Feldman, R & Sanger, J. (2007) *The Text Mining Handbook-Advanced Approaches in Analyzing Unstructured Data*, USA: New York.
- Gholamreza Nakhaeizadeh, (2007), Application of Bayesian Statistics in Classification Naïve Bayes, Statistical Data Mining.
- Go, Alec; Bhayani, Richa; Huang, Lei. (2009), *Twitter Sentimen Classification using Distant Supervision.*
- Huang, K., Yang, H., King, I., & Lyu, M. (2008). Machine Learning Modeling Data Locally And Globally. Berlin Heidelberg: Zhejiang University Press, Hangzhou And Springer-Verlag Gmbh.
- Karen Wickre (2013), The Art Of Social Selling: Finding And Engaging Customers On Twitter.
- Linhao Zhang, (2013), Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation, Department of Computer Science, The University of Texas at Austin.
- Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S. (2011). An Improved Particle Swarm Optimization for Feature Selection. Journal of Bionic Engineering, 8(2), 191–200. doi:10.1016/S1672-6529(11)60020-6.
- M. Tohar. (2000). *Membuka Usaha Kecil.* Kanisius: Yogyakarta.

- Muhammad Bilal, (2015), Sentiment classification of Roman-Urdu opinions using Naive Bayesian, Decision Tree and KNN classification techniques, Journal of King Saud University – Computer and Information Sciences.

- Nur Azizah Vidya, (2015), Twitter Sentiment to Analyze Net Brand reputation of Mobile Phone Providers, The Third Information Systems International Conference, Procedia Computer Science 72 (2015) 519 – 526.

- Pak, A., dan Paurobek, P., 2010, *Twitter as a Corpus for Sentimen Analysis and Opinion Mining*, Universite de Paris-Sud, Laboratoire LIMSI-CNRS. Batiment 508, F-91405 Orsay Cedex, France.

- Pang, Bo & Lilian, Lee., 2008, *Opinion Mining and Sentimen Analysis*. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135.

- Qin Li (2015), Examining the accuracy of sentiment analysis by brand monitoring companies, *IBA Bachelor Thesis Conference*, July 2nd , Enschede, The Netherlands.

- Serrano-Guerrero, Olivas, Romero, & Herrera-Viedma (2015), E: Sentiment analysis : a review and comparative Analysis of web service. Inf. Sci. 311, 18-38

- Triawati, Candra; Bijaksana, M.Arif; Indrawati, Nur; Saputro, Widyanto Adi. (2009) Pemodelan Berbasis Konsep Untuk Kategorisasi Artikel Berita Berbahasa Indonesia,Dalam Seminar Nasional Aplikasi Teknologi Informasi 2009.

- Yates, Baeza R. & Neto, Ribero B. (1999) *Modern Information Retrieval. New York*: ACM Press.

- Zhao, M., Fu, C., Ji, L., Tang, K., & Zhou, M. (2011). Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. Expert Systems with Applications, 38(5), 5197–5204. doi:10.1016/j.eswa.2010.10.041.