# KLASIFIKASI DOKUMEN MENGGUNAKAN KOMBINASI ALGORITMA *PRINCIPAL COMPONENT ANALYSIS DAN* SVM

## *DOCUMENT CLASSIFICATION USING COMBINATION OF PRINCIPAL COMPONENT ANALYSIS ALGORITHM AND SVM*

**Michael Freddy H.Sianturi[1], Adiwijaya[2], Said Al Faraby [3]**

[1,2,3]Prodi S1 Ilmu Komputasi Fakultas Informatika Universitas Telkom, Bandung
michaelfreddy@telkomuniversity.ac.id[1], adiwijaya@telkomuniversity.ac.id[2], saidalfaraby@telkomuniversity.ac.id[3]

**Abstrak**

**Klasifikasi dokumen teks adalah masalah yang sederhana namun sangat penting karena manfaatnya cukup besar mengingat jumlah dokumen yang ada setiap hari semakin bertambah. Dalam melakukan klasifikasi dokumen, pada tugas akhir ini digunakan algoritma. *Principal Component Analysis* merupakan suatu teknik yang dapat digunakan untuk mengekstrasi struktur dari suatu data yang berdimensi tinggi tanpa menghilangkan informasi yang signifikan pada keseluruhan data. SVM adalah metode *learning machine* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan hyperplane terbaik yang memisahkan dua buah class pada input space.**
**Hasil dari pengujian sistem menggunakan data yang direduksi oleh *Principal Component Analysis* (PCA) memiliki akurasi yang sedikit lebih rendah untuk *dataset* tertentu dibandingkan tanpa menggunakan PCA. Akurasi terbaik pada penelitian ini dihasilkan dari metode SVM dengan akurasi rata-rata 98.95%, sedangkan untuk metode SVM + PCA akurasi yang diperoleh rata-rata 96.7866%.**

**Kata kunci: Klasifikasi Dokumen, *Principal Component Analysis, Support Vector Machine.***

**Abstract**

***Classification of text documents is a simple but very important issue because the benefits are quite large considering the number of documents that exist every day is increasing. In doing the classification of documents, in this final project used algorithm. Principal Component Analysis is a technique that can be used to extract the structure of a high-dimensional data without losing any significant information to the whole data. SVM is a learning machine method that works on the principle of Structural Risk Minimization (SRM) in order to find the best hyperplane that separates the two classes in the input space. The results of system testing using data reduced by Principal Component Analysis (PCA) have slightly lower accuracy for certain datasets than without PCA. The best accuracy in this research is resulted from SVM method with an average accuracy of 98.95%, while for SVM + PCA method the accuracy obtained is 96.7866% on average.***

***Keyword: Document Classification, Principal Component Analysis, Support Vector Machine.***

## 1. PENDAHULUAN

Saat ini jumlah dokumen semakin banyak dan beragam sejalan dengan bertambahnya waktu dan teknologi. Jika jumlah dokumen semakin bertambah banyak maka proses pencarian dan penyajian dokumen semakin sukar, sehingga akan lebih mudah jika dokumen tersebut sudah tersedia sesuai dengan kategorinya masingmasing. Salah satu metode yang dapat membantu mengorganisasikan dokumen sesuai dengan ketegorinya adalah klasifikasi. Metode machine learning yang dipakai adalah menggunakan Support Vector Machine (SVM). Salah satu kelemahan dari SVM adalah tidak ada yang tahu apakah hasil klasifikasi yang dihasilkan oleh *classifier* SVM itu merupakan suatu dugaan atau suatu jawaban yang pasti, sebab classifier yang dihasilkan SVM, belajar dari pengalaman dan ekstraksi pengetahuan yang ada dalam database bertujuan untuk bisa mengklasifikasikan data baru, tetapi tidak bisa membedakan hasil jawaban apakah merupakan suatu dugaan atau suatu jawaban yang pasti, dengan kata lain hasil klasifikasinya tidak reliable. Dalam tugas akhir ini, SVM akan dikombinasikan dengan Principal Component Analysis (PCA).

Penelitian yang berkaitan dengan metode SVM telah dilakukan diantaranya oleh Ahmad Yusuf, Tirta Priambadha (2013) meneliti mengenai klasifikasi dokumen menggunakan *Support Vector Machine* yang didukung *K-Means Clustering* [2]. Peneliti mengusulkan sebuah metode untuk kategorisasi dokumen teks bahasa Inggris dengan terlebih dahulu menggunakan *K-Means* untuk melakukan pengelompokan kemudian digunakan multi-class *Support Vector Machine* untuk proses klasifikasi. Dengan adanya pengelompokan tersebut, variasi data dalam membentuk model klasifikasi akan lebih seragam. Dari hasil percobaan tersebut menunjukkan bahwa metode yang diusulkan mampu menghasilkan akurasi sebesar 88,1%, presisi sebesar 96,7% dan recall sebesar 94,4% dengan parameter jumlah kelompok sebesar 5. Dalam Tugas Akhir ini akan digunakan kombinasi dari dua

algoritma yang dapat membantu dalam melakukan klasifikasi dokumen yaitu *Principal Component Analysis* (PCA) dan *Support Vector Machine* (SVM). PCA digunakan lebih jauh dalam mereduksi dengan cara memilih dimensi data yang paling penting dan *classification* yang pada tugas akhir ini menggunakan SVM.

## 2. TINJAUAN PUSTAKA

### 2.1 Pengertian Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk memperkirakan kelas yang tidak diketahui dari suatu objek. Dalam pengklasifikasian data terdapat dua proses yang dilakukan yaitu:

1) Proses training

   Pada proses training digunakan training set yang telah diketahui label-labelnya untuk membangun model atau fungsi.

2) Proses testing

   Untuk mengetahui keakuratan model atau fungsi yang akan dibangun pada proses training, maka digunakan data yang disebut dengan testing set untuk memprediksi label-labelnya.

Klasifikasi dokumen adalah pemberian kategori yang telah didefinisikan kepada dokumen yang belum memiliki kategori (Goller, 2000). Mengklasifikasi dokumen merupakan salah satu cara untuk mengorganisasikan dokumen. Dokumen-dokumen yang memiliki isi yang sama akan dikelompokkan ke dalam kategori yang sama. Dengan demikian, orang-orang yang melakukan pencarian informasi dapat dengan mudah melewatkan kategori yang tidak relevan dengan informasi yang dicari atau yang tidak menarik perhatian (Feldman, 2004).

### 2.2 Principal Component Analysis (PCA)

*Principal Component Analysis* adalah teknik linear klasik untuk pereduksian dimensi data. Principal Component Analysis (PCA) adalah suatu teknik statistik yang secara linear mengubah bentuk sekumpulan variabel asli nienjadi kumpulan variabel yang lebih kecil yang tidak berkorelasi yang dapat mewakili informasi dari kumpulan variabel asli (Dunteman. 1989:7).

Tujuan PCA adalah untuk menjelaskan bagian dari variasi dalam kumpulan variabel yang diamati atas dasar beberapa dimensi. Dari variabel yang banyak dirubah menjadi sedikit variabel.

### 2.3 Klasifikasi

Pada Tugas Akhir ini proses klasifikasi merupakan proses di mana hasil ekstraksi ciri akan diolah menggunakan metode klasifikasi untuk memisahkan tiga kelas yaitu *acq, crude, earn, grain, interest, money-fx, ship dan trade*. Jenis *classifier* terbagi dalam tiga jenis utama yaitu:

1. *Classifier* berdasarkan keputusan *Bayes*

   *Classifier* ini melakukan klasifikasi M kelas ($w1,w2,\ldots,wm$) dan pola yang tidak diketahui yang direpresentasikan dengan vektor x dengan membentuk probabilitas kondisional sebanyak M ($P(wi|x),i=1,2,\ldots,M$). Nilai tersebut mewakili probabilitas bahwa pola yang tidak diketahui tersebut tergolong dalam kelas *wi* tertentu. *Classifier* mencari nilai kemungkinan terbesar tersebut untuk melakukan klasifikasi.

2. *Linear classifier*

   *Linear classifier* melakukan klasifikasi dengan menggunakan persamaan linear sederhana. Keunggulan dari *classifier* ini adalah simplisitasnya dan metode komputasi yang digunakan.

3. *Nonlinear classifier*

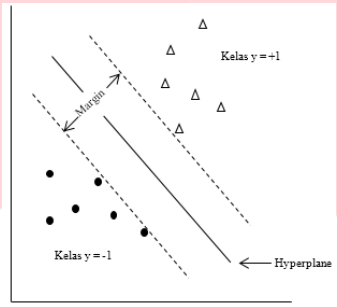
   Merupakan pengembangan dari dari *linear classifier*. Untuk kasus di mana kelas tidak dapat dipisahkan secara linier, *classifier* ini dirancang secara optimal dengan persyaratan tertentu, misalnya dengan meminimalkan kuadrat error yang dihasilkan.

### 2.4. Support Vector Machine (SVM)

*Support Vector Machine* adalah salah satu metode klasifikasi ciri yang bertujuan menemukan hyperplane terbaik yang memisahkan dua buah *class* pada input *space.* Prinsip dasarnya adalah *linier classifier*, dan selanjutnya dikembangkan agar dapat bekerja pada masalah *non-linear* dengan memasukkan *kernel trick* pada ruang kerja berdimensi tinggi.

### 2.4.1 Kasus data terpisah secara linear

Dalam kasus permasalahan data dua dimensi yang memiliki dua kelas yaitu +1 dan-1, data dapat terpisah secara *linear*. Ilustrasi data terpisah secara *linear* dapat dilihat pada gambar 2.1. Dalam ilustrasi tersebut *hyperplane* didefinisikan sebagai garis lurus dan jarak dari *margin* ke *hyperplane* didefinisikan sebagai garis putus-putus. *Margin* adalah jarak antar *hyperplane* dengan dua himpunan. Untuk mencari hyperplane terbaik dilakukanlah pencarian jarak maksimum *margin* dari *hyperplane* tersebut sehingga data dapat diklasifikasikan lebih akurat daripada *hyperplane* dengan jarak *margin* terkecil [3].




**Gambar 1. Ilustrasi Data Terpisah Secara Linear[3]**

Untuk menemukan pemisah *hyperplane* digunakan persamaan sebagai berikut sebagai berikut [8].

$$h_{(x)} = w^T x + b \quad (1)$$

Keterangan
w : bobot vector
$x$ : data latih
b : bias

bobot yang digunakan dapat disesuaikan, sehingga *hyperplane* dapat didefinisikan sebagai sisi dari *margin* [8]. Label kelas pada SVM dinyatakan sebagai $y_i \in \{-1, +1\}$.
Data yang masuk kedalam kelas -1 adalah data yang memenuhi pertidaksamaan sebagai berikut [8]:

$$w.x_i + b \leq -1 \quad (2)$$

Dan data yang masuk kedalam kelas +1 adalah data yang memenuhi pertidaksamaan sebagai berikut [8]:

$$w.x_i + b \leq +1 \quad (3)$$

*Margin hyperplane* didiberikan oleh jarak antar dua kelas diatas dengan notasi sebagai berikut [8]:

$$\|w\| \; x \; d = 2 \; atau \; d = \frac{2}{\|w\|} \quad (4)$$

Keterangan :
$\|w\|$ : vector bobot dari w.
$d$ : jarak antar dua kelas

Klasifikasi kelas data pada SVM pada persamaan (2.4) dan (2.5) dapat digabungkan dengan notasi [8]:

$$y\,(w.x_i + b) \geq 1, i = 1,2,\ldots,N \quad (5)$$

### 2.4.2 Kasus data terpisah secara *nonlinear*

Dalam permasalahan *nonlinear* tidak ada garis yang dapat memisahkan kelas [2]. Maka dari itu, permasalahan *linear* dapat diperluas agar dapat menyelesaikan permasalahan *nonlinear* dengan cara mentransformasikan data masukan kedalam dimensi yang lebih tinggi. Data ditransformasikan menjadi *nonlinear mapping* yang didefinisikan sebagai $\phi(X)$ sehingga menghasilkan fungsi *kernel* sebagai berikut [2].

$$K(x_i, x_j) = \phi(x_i) \ . \phi(x_j) \qquad (6)$$

Untuk setiap data latih yang didefinisikan sebagai $\phi(x_i) \ . \phi(x_j)$ dapat digantikan oleh $K(x_i, x_j)$.

### 2.5 Penelitian Terkait

"Ada beberapa metode klasifikasi yang biasa digunakan dalam klasifikasi data baik berupa text maupun speech, antara lain multinomial naive bayes [16, 17], bayesian network [18], dan Hidden Markov Model [19, 20].

Penelitian yang berkaitan dengan metode SVM telah dilakukan diantaranya oleh Ahmad Yusuf, Tirta Priambadha (2013) meneliti mengenai klasifikasi dokumen menggunakan *Support Vector Machine* yang didukung *K-Means Clustering* [2]. Dari hasil percobaan tersebut menunjukkan bahwa metode yang diusulkan mampu menghasilkan akurasi sebesar 88,1%, presisi sebesar 96,7% dan recall sebesar 94,4% dengan parameter jumlah kelompok sebesar 5.

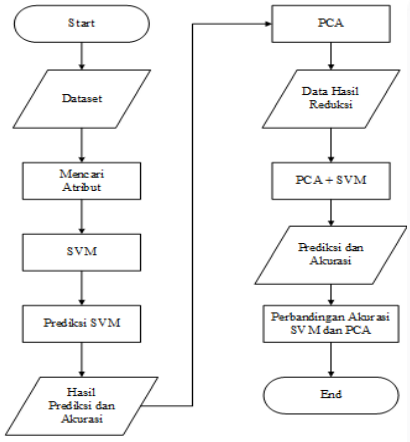
Sebelumnya pada penelitian *PCA Document Reconstruction for Email Classification* (Gomez, J.C. and Moens, M.F., 2012) [1]. Pada makalah tersebut, peneliti menyajikan dokumen *classifier* berdasarkan fitur konten teks dan aplikasi untuk klasifikasi *email*. Peneliti menguji validitas *classifier* dengan menggunakan *Principal Component Analysis Document Reconstruction* (PCADR). Dimana ide dari penelitian nya adalah *principal component analysis* (PCA) dapat mengkompres secara optimal hanya jenis dokumen, dalam eksperimen ini kelas *email* digunakan untuk menghitung *principal components* (PCs). Dengan demikian, *classifier* menghitung secara terpisah PCA untuk masing-masing kelas dokumen. Percobaan ini menunjukkan bahwa PCADR mampu untuk mendapatkan hasil yang sangat baik dengan berbeda dataset validasi yang digunakan.

## 3. Perancangan Sistem dan Simulasi

### 3.1 Flowchart Sistem

Pada tugas akhir ini akan dijelaskan mengenai perancangan untuk melakukan eksperimen dari klasifikasi dokumen teks. Dimana dalam pengklasifikasiannya digunakan *SVM*. Sedangkan untuk mereduksi atribut pada data digunakan *Principal Component Analysis* (PCA).

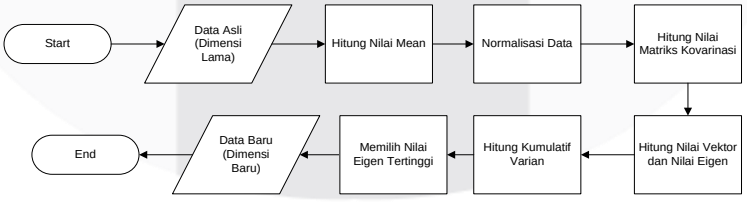
.




**Gambar 2. Flowchart Sistem**

### 3.2 Process PCA

PCA merupakan suatu sistem *preprocessing* untuk data yang berdimensi besar. Data berdimensi besar akan dimasukan ke dalam PCA, lalu user akan memilih metode apa yang akan digunakan. Dalam tugas akhir ini digunakan metode korelasi. Data kemudian di eksekusi oleh sistem, sehingga menghasilkan data baru dengan dimensi yang lebih sedikit. Berikut ini merupkan diagram proses *preprocessing* dengan menggunakan PCA:




**Gambar 3. Flowchart Proses PCA**

Berikut ini merupakan penjelasan dari Gambar 3.1-2 [10] :

1. Menghitung nilai rata-rata (mean) dari dataset.
   Sebagai contoh, terdapat dataset dengan dua dimensi,yaitu dimensi *m* dan *n* dengan hasil perhitungan mean sebagai berikut:

***Table 1 : Contoh Perhitungan***

| | *M* | *n* |
|---|---|---|
| Data | 182 | 219 |
| | 210 | 220 |
| | 5 | 100 |
| | 94 | -4 |
| | 13 | 50 |
| | 310 | 110 |
| **Mean** | **135,6** | **115,83** |

Nilai mean diperoleh dari nilai data dibagi dengan banyaknya data. Nilai mean yang diperoleh sebanding dengan jumlah dimensinya.

2. Normalisasi data, dimana pada tahap ini akan dilakukan pengurangan record pada data asli dengan menggunakan persamaan (2.1). Hasil yang diperoleh dari tahap ini merupakan Data Adjust.

3. Menghitung Kovarian
   Dilakukan perhitungan kovarian satu per satu tujuannya adalah untuk mengetahui hubungan antara dimensi yang satu dengan yang lain.
4. Menghitung nilai dan vektor eigen dari matriks kovarian.

5. Hitung kumulatif varian dari hasil variansi yang telah didapat dengan menggunakan persamaan berikut [10]:

$$cumvar = \left(\frac{\sum variansi}{n}\right) \times 100\%$$

6. Memilih nilai eigen tertinggi, dimana tahap ini nilai eigen diurutkan terlebih dahulu dari yang paling besar ke nilai yang paling kecil. Lalu dengan menggunakan kriteria pemilihan PC dari Kaiser Guttman dipilih nilai eigen yang lebih besar dari 1.

Dimensi baru,merupakan data hasil reduksi yang selanjutnya akan masuk ke tahap selanjutnya yaitu proses pengklasifikasian lagi.Pada tugas akhir ini PCA digunakan untuk mereduksi dimensi agar mempercepat waktu komputasi.

### 3.4 Hasil Simulasi

#### 3.4.1 Hasil Pengujian Menggunakan SVM Terhadap Tingkat Akurasi dan Proses Komputasi

Pada skenario ini akan dilakukan pengujian pada data latih yang bejumlah 900 dokumen serta data uji sebesar 100 dokumen. Berikut merupakan hasil dan analisis skenario 1. Pengujian dilakukan di Matlab 2015a.

**Tabel 2. Hasil Pengujian SVM Terhadap Akurasi Training dan Akurasi Testing dan Proses Komputasi**

| Kombinasi | Jenis Data | Kernel | Kernel Option | C | Akurasi (%) | Waktu (Detik) |
|---|---|---|---|---|---|---|
| 1 | | RBF | 9 | 100 | 100 | 160313.0721 |
| 2 | | | | 100 | 100 | 2866.6017 |
| 3 | | | 8 | 200 | 100 | 2861.3159 |
| 4 | | | | 100 | 100 | 2958.1296 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 5 | Training | Linear | 9 | 200 | 100 | 2919.1521 |
| 6 | | | | 100 | 100 | 2858.1228 |
| 7 | | | 11 | 200 | 100 | 2845.6687 |
| 8 | | | | 100 | 100 | 2875.7789 |
| 9 | | | 12 | 200 | 100 | 2861.8900 |
| 10 | Testing | RBF | 9 | 100 | 95 | 17933.6205 |
| 11 | | | 11 | 200 | 96 | 15793.3403 |
| 12 | | | 12 | 200 | 96 | 15014.2640 |
| 13 | | Linear | | 100 | 99 | 331.0603 |
| 14 | | | 8 | 200 | 99 | 330.5187 |
| 15 | | | | 100 | 99 | 333.5074 |
| 16 | | | 9 | 200 | 99 | 334.6041 |
| 17 | | | | 100 | 99 | 331.9844 |
| 18 | | | 11 | 200 | 99 | 331.8728 |
| 19 | | | | 100 | 99 | 333.3671 |
| 20 | | | 12 | 200 | 99 | 361.9099 |

### 3.4.2 Hasil Pengujian Menggunakan PCA Terhadap Tingkat Akurasi dan Proses Komputasi

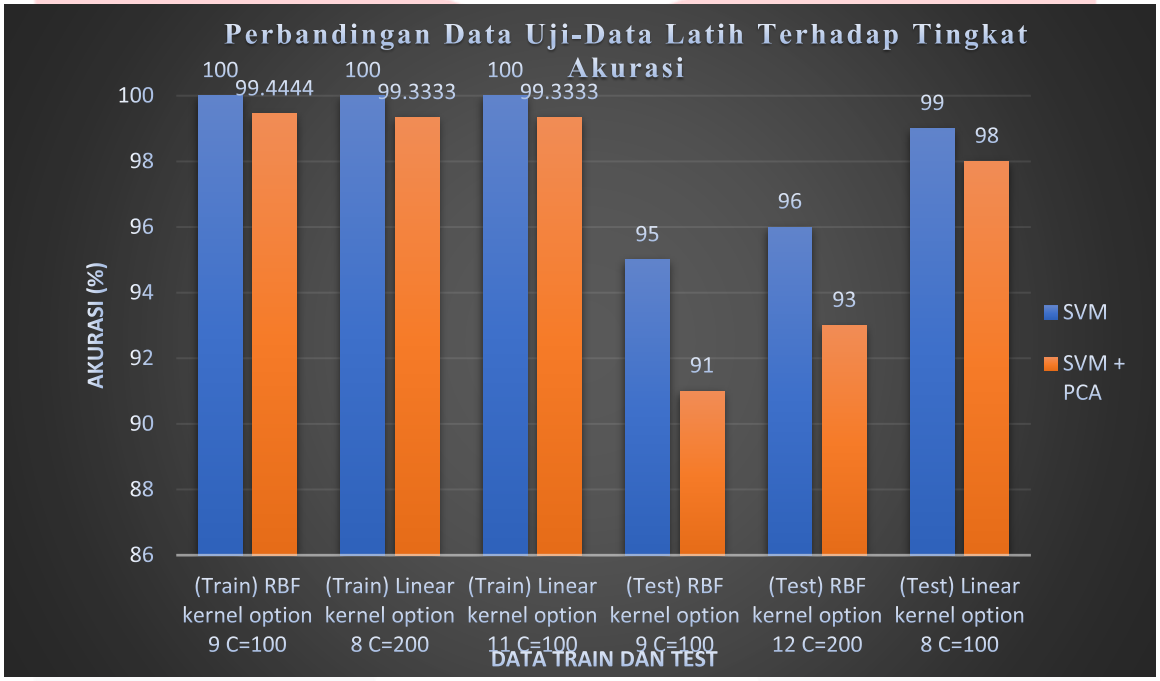
Sebelumnya diperoleh atribut sebanyak 19985 kemudian PCA melakukan pereduksian jumlah atribut menjadi 5484. Pada skenario ini akan dilakukan pengujian pada data latih dan data uji sama skenario sebelumnya tetapi pada skenario ini sudah menggunakan metode PCA. Berikut merupakan hasil dan analisis dari scenario 2:

**Tabel 3. Hasil Pengujian Menggunakan PCA Terhadap Akurasi Training dan Akurasi Testing dan Proses Komputasi**

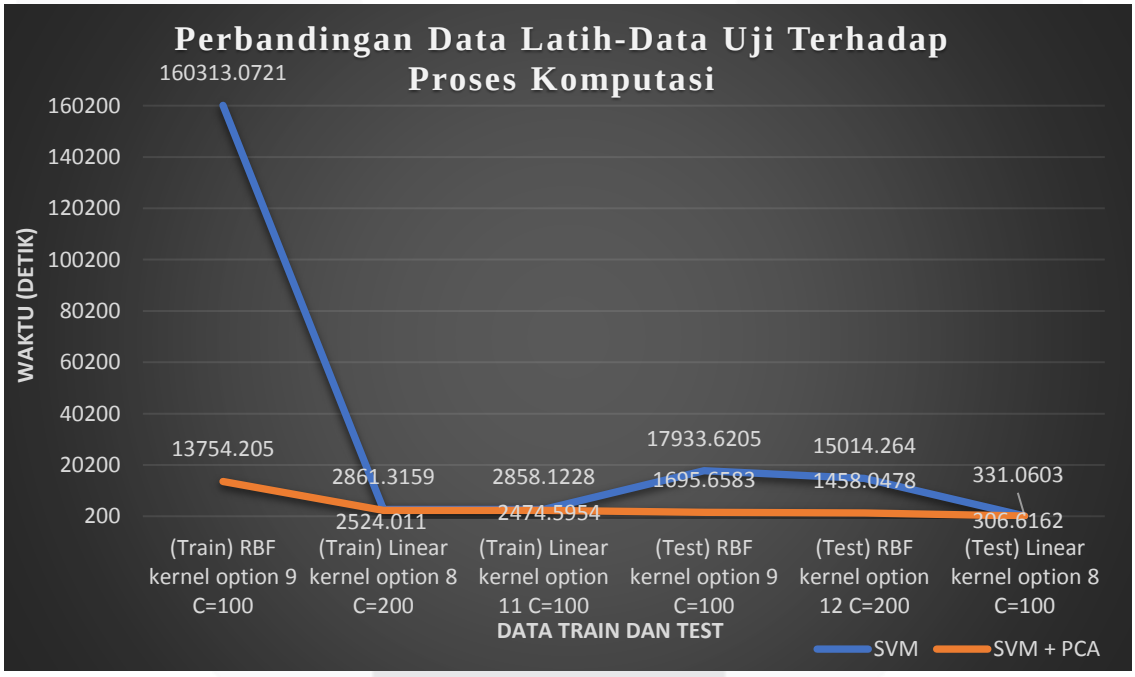
| Kombinasi | Jenis Data | Kernel | Kernel Option | C | Akurasi (%) | Waktu (Detik) |
|---|---|---|---|---|---|---|
| 1 | Training | RBF | 8 | 100 | 98.5556 | 16508.0298 |
| 2 | | | | 200 | 98.5556 | 16328.6604 |
| 3 | | | 9 | 100 | 99.4444 | 13754.2050 |
| 4 | | | 12 | 200 | 99.4444 | 12831.3306 |
| 5 | | Linear | 8 | 100 | 99.3333 | 2525.3850 |
| 6 | | | | 200 | 99.3333 | 2524.0110 |
| 7 | | | 9 | 100 | 99.3333 | 2726.2091 |
| 8 | | | | 200 | 99.3333 | 2778.8599 |
| 9 | | | 11 | 100 | 99.3333 | 2474.5954 |
| 10 | | | | 200 | 99.3333 | 2487.9871 |
| 11 | | | 12 | 100 | 99.3333 | 2468.3207 |
| 12 | | | | 200 | 99.3333 | 2486.9129 |
| 13 | Testing | RBF | 8 | 100 | 89 | 1823.5642 |
| 14 | | | | 200 | 89 | 1836.3926 |
| 15 | | | 9 | 100 | 91 | 1695.6583 |
| 16 | | | 11 | 100 | 93 | 1521.0681 |
| 17 | | | 12 | 200 | 93 | 1458.0478 |
| 18 | | Linear | 8 | 100 | 98 | 306.6162 |
| 19 | | | | 200 | 98 | 310.5841 |
| 20 | | | 9 | 100 | 98 | 307.0806 |
| 21 | | | | 200 | 98 | 307.3614 |
| 22 | | | 11 | 100 | 98 | 307.5406 |
| 23 | | | | 200 | 98 | 308.4111 |
| 24 | | | 12 | 100 | 98 | 307.6544 |
| 25 | | | | 200 | 98 | 307.3286 |

### 3.4.3 Hasil Perbandingan Pengujian SVM Tanpa Menggunakan PCA dan Menggunakan PCA Terhadap Tingkat Akurasi dan Proses Komputasi

Pada skenario ini akan dilakukan perbandingan terhadap waktu dan proses komputasi antara SVM dan SVM + PCA. Berikut merupakan hasil dan analisis dari scenario 3:




*Gambar 4 Grafik Perbandingan Data Uji-Data Latih Terhadap Tingkat Akurasi*




*Gambar 5 Grafik Perbandingan Data Uji-Data Latih Terhadap Proses Komputasi*

Berdasarkan hasil dari seluruh percobaan diatas, maka metode SVM tanpa PCA memiliki rata-rata akurasi lebih baik yaitu sebesar 98.95% pada seluruh dokumen uji dengan berbagai parameter-parameter. Tetapi proses komputasinya sangat lama jika dibandingkan dengan menggunakan PCA. Misalnya pada data *train*, fungsi kernel RBF, kernel option 9 dan C=100, perbandingan proses komputasi nya sebesar 1.6962 hari. Sedangkan dengan menggunakan metode SVM + PCA akurasi yang diperoleh tidak terlalu jauh dibandingkan hanya menggunakan SVM. Akurasi yang diperoleh rata-rata sebesar 96.7866% dan perbandingan akurasi dari kedua metode yang diuji hanya sebesar 2.1634%.

Pengujian menggunakan fungsi kernel sangat mempengaruhi seluruh proses komputasi dari data yang diuji, akurasi dan proses komputasi dari fungsi kernel linear jauh lebih baik daripada kernel RBF.

## 4. Kesimpulan

Berdasarkan hasil pengujian dan analisis dapat disimpulkan bahwa :

1. Algoritma PCA melakukan *feature reduction* pada jumlah atribut dari 19985 menjadi 5484. Penggunaan parameter yang berbeda membuat waktu komputasi yang dihasilkan bervariasi namun perbandingan akurasi yang didapat tidak terlalu jauh pada data yang diuji.
2. Berdasarkan percobaan SVM + PCA yang telah dilakukan pada data dokumen online, proses komputasi tercepat pada dokumen *training* diperoleh dari kombinasi fungsi kernel linear, kernel option 12 dan C=100 dengan waktu 2468.3207 detik dan pada dokumen *testing* diperoleh dari kombinasi fungsi kernel linear, kernel option 9 dan C=100 dengan waktu 307.0806 detik..
3. Fungsi kernel linear menghasilkan waktu komputasi lebih cepat dibandingkan fungsi kernel RBF (non-linear).
4. Penggunaan PCA + SVM pada *dataset dokumen online* menghasilkan waktu komputasi jauh lebih cepat dibandingkan menggunakan SVM. Tetapi SVM menghasilkan akurasi yang lebih baik daripada menggunakan PCA yaitu sebesar 98.95%

Saran yang dapat diberikan penulis untuk perkembangan Tugas Akhir ini antara lain :

1. Algortima SVM merupakan algoritma yang cocok untuk mengklasifikasi dokumen online, karena akurasi yang dihasilkan oleh SVM sudah sangat baik untuk pengujian dataset Penambahan pada proses *preprocessing* sehingga dapat dihasilkan hasil *preprocessing* yang lebih baik, yang mampu meningkatkan akurasi pengenalan sistem.
2. Menggunakan variasi nilai kernel option dan C yang lebih besar belum tentu mendapatkan waktu komputasi yang lebih cepat.
3. Menggunakan kapasitas RAM minimal 8 GB karena kapasitas memori komputer mempengaruhi proses komputasi, semakin besar kapasitas RAM semakin cepat juga proses komputasi nya.

## DAFTAR PUSTAKA


[1] Gomez, J.C. and Moens, M.F., 2012. PCA document reconstruction for email classification. *Computational Statistics & Data Analysis*, *56*(3), pp.741-751..

[2] Yusuf, A. and Priambadha, T., 2013. Support Vector Machines yang didukung K-Means clustering dalam klasifikasi dokumen. *JUTI: Jurnal Ilmiah Teknologi Informasi*, *11*(1), pp.15-18..

[3] Darujati, C. and Gumelar, A.B., 2012. Pemanfaatan Teknik Supervised Untuk Klasifikasi Teks Bahasa Indonesia. *Jurnal Bandung Text Mining*, *16*(1), pp.5-1..

[4] Hevi Herlina, U., Prahasto, I.T., ASc, M. and Achmad Widodo, S.T., 2013. *PROGNOSIS KERUSAKAN BANTALAN GELINDING DENGAN MENGGUNAKAN METODE SUPPORT VECTOR REGRESSION (SVR)*(Doctoral dissertation, Diponegoro University).

[5] Reuters-21578 Text Categorization Collection Datasets for single-label text categorization, http://www.cs.umb.edu/~smimarog/textmining/datasets/

[6] D. Nugraheny, "Hasil Ekstraksi Algoritma Principal Component Analysis ( PCA ) untuk Pengenalan Wajah dengan Bahasa Pemograman Java Eclipse IDE," *Sisfotek Glob.*, vol. 2, pp. 26–30, 2015.

[7] Nugroho, A.S., Witarto, A.B. and Handoko, D., 2003. Support Vector Machine. *Teori dan Aplikasinya dalam Bioinformatika, Ilmu Komputer. com, Indonesia*.

[8] Prasetyo, E., 2014. *DATA MINING Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: ANDI.

[9] Slonim, N., Friedman, N. and Tishby, N., 2002, August. Unsupervised document classification using sequential information maximization. In*Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 129-136). ACM.

[10] Smith, L.I., 2002. A tutorial on principal components analysis. *Cornell University, USA*, *51*(52), p.65.
[11] Jackson, J.E., 1991. Singular Value Decomposition: Multidimensional Scaling I. *A User's Guide to Principal Components*, pp.189-232.
[12] Gasong, A.A., Deteksi Tumor Otak Berdasarkan citra MRI dengan menggunakan metode Independent Component analysis (ICA) dan Support Vector Machines (SVM). *Tugas Akhir. Institut Teknologi Telkom, Bandung*.
[13] GFirmanto, A., 2011. Implementasi Principal Component Analysis dan Backpropagation Neural Network dalam Pengklasifikasian Terjemahan Ayat-Ayat Ilmu Pengetahuan dalam Alquran.
[14] Adiwijaya, 2014, Aplikasi Matriks dan Ruang Vektor, Yogyakarta: Graha Ilmu.
[15] Adiwijaya, 2016, Matematika Diskrit dan Aplikasinya, Bandung: Alfabeta.
[16] MS Mubarok, Adiwijaya, MD Aldhi, 2017, Aspect-based sentiment analysis to review products using Naïve Bayes, AIP Conference Proceedings 1867, 020060 (2017)
[17] Aziz, R.A., Mubarok, M.S. and Adiwijaya, 2016. Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes. In Indonesia Symposium on Computing (IndoSC) 2016.
[18] Arifin, A.H.R.Z., Mubarok, M.S. and Adiwijaya, 2016, Learning Struktur Bayesian Networks menggunakan Novel Modified Binary Differential Evolution pada Klasifikasi Data. In Indonesia Symposium on Computing (IndoSC) 2016.
[19] Yulita, I.N., Houw Liong The and Adiwijaya, 2012. Fuzzy Hidden Markov Models for Indonesian Speech Classification. JACIII, 16(3), pp.381-387.
[20] U.N. Wisesty, M.S. Mubarok, Adiwijaya, 2017, A classification of marked hijaiyah letters' pronunciation using hidden Markov model, AIP Conference Proceedings 1867 (1), 020036.
[21] Sinaga, A., Nugroho, H., and Adiwijaya, 2015. Development of word-based text compression algorithm for Indonesian language document. In Information and Communication Technology (ICoICT), 2015 3rd International Conference on (pp. 450-454). IEEE.