# 1. PENDAHULUAN

## 1.1 Latar Belakang

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. Klasifikasi dokumen adalah proses menggolongkan suatu dokumen ke dalam suatu kategori tertentu (Manning et al 2008). Beberapa Teknik yang banyak digunakan pada klasifikasi documen teks diantaranya Naïve Bayes [Lewis, 1998; McCallum dan Nigam, 1998; Sahami, 1996], K-Nearest Neighbor [Yang,1999], Support Vector Machines [Joachims, 1998; Dumais et al., 1998], Boosting [Schapire and Singer, 1999], Algoritma Rule Learning [Cohen and Singer, 1996; Slattery dan Craven, 1998], dan Maximum Entropy [Nigam et al., 1999].

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.

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.

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), dan untuk jenis dokumen lainnya kompresi tidak akan bekerja dengan baik hanya untuk beberapa komponen. 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, mencapai hasil yang lebih baik daripada Support Vector Machine *classifier*.

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.

## 1.2 Perumusan Masalah

Berdasarkan latar belakang diatas, maka terdapat beberapa permasalahan yang akan diselesaikan dalam tugas akhir ini, permasalahan tersebut terdiri dari:

1. Bagaimana mengklasifikasi dokumen dengan menggunakan kombinasi PCA dan SVM ?
2. Bagaimana menganalisa performansi hasil dari klasifikasi tersebut sehingga bisa menghasilkan *improvement* pada performansi ?
3. Bagaimana prediksi yang didapatkan SVM dan SVM + PCA pada data dokumen online?

Adapun batasan masalah dari tugas akhir ini adalah sebagai berikut:

1. Data yang digunakan merupakan R8 *of Reuters-21578 Text Categorization Collection Data Set* yang diperoleh dari *Umas Boston Computer Science*
2. *Dataset* yang digunakan sudah melalui tahap preprocessing terlebih dahulu.
3. *Dataset* belum memiliki atribut.
4. Algoritma PCA digunakan sebagai algoritma untuk mereduksi atribut.
5. Data yang digunakan yaitu 900 *data training* dan 100 *data testing.*
6. *Sofware* yang digunakan adalah MatLAB R2015a.
7. Metode multiclass menggunakan Metode OAO (*One Against One).*
8. Kernel yang diuji yaitu kernel RBF dan *Linear*

## 1.3 Tujuan

Berdasarkan perumusan masalah diatas tujuan untuk menyelesaikan masalah tersebut adalah:

1. Menganalisis performansi yang didapatkan *Support Vector Machine (SVM)* pada klasifikasi dokumen.
2. Membandingkan tingkat akurasi dan proses komputasi antara proses SVM dan SVM + PCA

## 1.4 Sistematika Penulisan

Pada sub bab ini dijelaskan secara singkat mengenai uraian lingkup isi dari setiap bab yang ada dalam buku tugas akhir ini. Bab 1 menjelaskan tentang latar belakang masalah berdasarkan judul yang diajukan penulis serta tujuan dari pembuatan tugas akhir ini. Bab 2 menjelaskan tentang teori-teori pendukung berdasarkan masalah yang dihadapi. Bab 3 menjelaskan tentang perancangan sistem dan skenario yang dibuat untuk penyelesaian masalah yang dihadapi. Bab 4 menjelaskan tentang hasil pengujian dan analisis sistem yang telah dibuat. Bab 5 berisi kesimpulan dan saran atas selesainya tugas akhir ini