BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Penyebaran dan kebutuhan akan informasi digital di Indonesia dalam bentuk teks atau dokumen semakin meningkat dan setiap waktu terus mengalami pertumbuhan seiring dengan perkembangan teknologi. Salah satu sumber informasi tersebut adalah portal berita elektronik. Suatu portal berita elektronik mengklasifikasi artikel – artikel ke dalam kategori. Selama ini pengklasifikasian berita masih menggunakan tenaga manusia atau manual. Kategori yang banyak beserta waktu yang cepat akan menyulitkan editor untuk mengklasifikasikan artikel, terutama pada artikel yang isinya tidak terlalu berbeda secara jelas. Beberapa kategori yang penggunaan bahasanya tidak berbeda terlalu jauh seperti olahraga, sains, ekonomi, teknologi dan kesehatan mengharuskan seorang editor mengetahui isi artikel secara keseluruhan untuk selanjutnya dimasukkan ke dalam kategori yang tepat. Akan lebih efesien apabila kategori berita dimasukkan secara otomatis dengan menggunakan metode tertentu. Pengklasifikasian juga dilakukan untuk mempermudah para pengguna dalam mengakses artikel.

Support Vector Machine (SVM) adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur (*feature space*) berdimensi tinggi, dilatih dengan algoritma pembelajaran yang didasarkan pada teori optimasi dengan mengimplementasikan learning bias yang berasal dari teori pembelajaran statistik [1]. SVM memiliki kelebihan yaitu mampu menemukan fungsi pemisah (klasifier) yang optimal yang bisa memisahkan dua set data dari dua kelas yang berbeda. Tetapi SVM pada saat pertama kali diperkenalkan oleh Vapnik hanya dapat mengklasifikasikan data ke dalam dua kelas (klasifikasi biner). Sementara masalah di dunia nyata umumnya mempunyai banyak kelas. Salah satu cara untuk mengimplementasikan *Multiclass* SVM yaitu dengan menggabungkan beberapa SVM biner [1]. Pada pengaturan *Multiclass*, sebuah dokumen teks masuk dalam tepat salah satu kelas dari beberapa kelas. Pendekatan *Multiclass* SVM yang digunakan dalam penelitian adalah *One-vs-One*, *One-vs-All*,

dan *error correcting output code* (ECOC).

Pada penelitian sebelumnya sudah ada yang melakukan perbandingan pendekatan *Multiclass* SVM. Pendekatan yang digunakan adalah *one vs one*, *one vs all* [20]. Data yang digunakan dalam penelitian sebelumnya adalah data aroma yang terdiri dari 3 jenis aroma. Dari hasil percobaan memberikan hasil bahwa metode *One-vs-One* telah mampu 100% mengklasifikasikan data aroma berdasarkan kelas yang tepat. Semakin banyak data training yang digunakan, metode *One-vs-One* akan lebih cepat mengklasifikasikan data dibandingkan dengan metode *One-vs-Rest* [20].

Selain metode *One vs One* dan *One vs All*, ada metode *error correcting output code* (ECOC). Pendekatan *error correcting output code* (ECOC) merupakan pendekatan yang terinspirasi dari pendekatan teori informasi untuk mengirimkan pesan melalui saluran yang ber-noise [1]. Metode *error correcting output code* (ECOC) dianggap sebagai varian dari metode *One vs All classification* [21]. Tetapi belum ada penelitian yang membandingkan performansi pada metode *One vs One* dan *One vs All* dengan metode ECOC khususnya pada data berbahasa Indonesia, padahal ECOC merupakan salah satu metode yang paling popular [10].

Berdasarkan uraian diatas dapat disimpulkan bahwa penggunaan ECOC pada data berbahasa Indonesia belum ada, maka perlu diadakan penelitian untuk membandingkan metode *One-vs-One*, *One-vs-All*, dan *error correcting output code* (ECOC) untuk mengetahui akurasi dari masing – masing metode agar dapat mengetahui metode mana yang lebih sesuai untuk kasus pengklasifikasian artikel dan diharapkan dapat membantu penelitian lebih lanjut dalam pengembangan pendekatan *Multiclass SVM.*

1.2 Identifikasi Masalah

Berdasarkan uraian latar belakang yang telah dikemukakan sebelumnya, maka identifikasi masalah pada penelitian ini adalah belum diketahuinya akurasi, kelebihan dan kekurangan antara metode *One vs One, One vs All*, dan *Error Correcting Output Code* pada data artikel berbahasa Indonesia.

1.3 Maksud dan Tujuan

Penelitian ini bermaksud untuk melakukan perbandingan beberapa pendekatan Multiclass SVM pada klasifikasi artikel berbahasa Indonesia. Adapun tujuan dari penelitian ini adalah untuk mengetahui perbedaan nilai akurasi antara *multiclass* svm *one vs one, one vs all* dan *error correcting output code* , dari aspek penggunaan *precision, recall* dan *F-measure.*

1.4 Batasan Masalah

Batasan masalah dalam penelitian yang dilakukan antara lain :

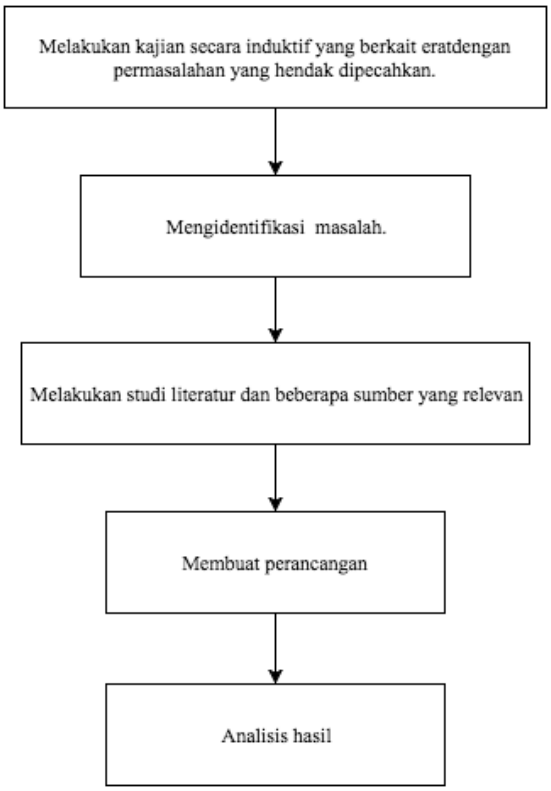
1. Input
    a. Dokumen yang digunakan sebagai inputan adalah dokumen berbahasa indonesia dengan ekstensi *.txt, *.csv dan bersifat *plain text*.
    b. Dokumen yang digunakan diambil dari kompas (www.kompas.com), tempo (www.tempo.com), merdeka (www.merdeka.com) dan tribunnews (www.tribunnews.com).
    c. Teks artikel tidak disertai gambar.
    d. Teks artikel tidak disertai tag html.
2. Proses
    a. Proses *preprocessing* yang dilakukan adalah *case folding*, *filtering*, tokenisasi, *stopword* dan *feature extraction*.
    b. Library yang digunakan untuk implementasi SVM mwnggunakan Sklearn.
3. Output
    a. Kelas target pada berita ada 5 yaitu : Ekonomi, Teknologi, Kesehatan, Pariwisata dan Olahraga. Pemilihan kelas target dilihat dari segi persoalan menurut Sedia Willing Barus[4].
    b. Kernel yang digunakan adalah kernel RBF dengan gamma 0.5.

1.5 Metodologi Penelitian

Metodologi penelitian yang digunakan dalam penelitian ini adalah metode penelitian ekperimental. Metode ekperimental adalah metode yang mempunyai

tujuan untuk menjelaskan hubungan sebab – akibat antara satu variabel dengan lainnya [16]. Hasil akhir dari penelitian ini adalah berupa grafik persentase klasifikasi teks pada artikel berbahasa Indonesia.

Langkah-langkah yang dilakukan selama melakukan penelitian dapat dilihat pada Gambar 1.1



Gambar 1.1 Langkah – langkah penelitian

Metode yang digunakan dalam penulisan tugas akhir ini menggunakan dua metode, yaitu metode pengumpulan data dan pembangunan perangkat lunak.

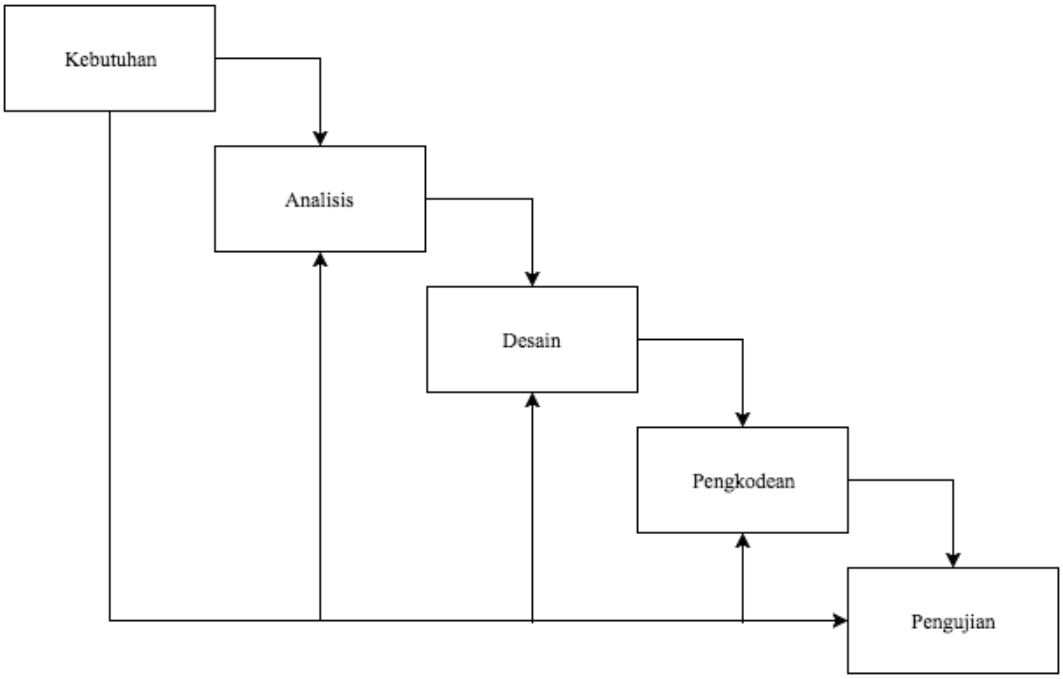
1.5.1 Metode Pengumpulan Data

Metode pengumpulan data yang digunakan dalam penelitian ini adalah studi literatur, yaitu dengan mencari jurnal, e-book, buku, serta artikel-artikel mengenai *Text Mining,* metode Multiclass SVM, *kernel – based methods*, serta penelitian-penilitian sebelumnya yang terkait dengan topik penelitian.

1.5.2 Metode Pembangunan Perangkat Lunak

Metode pembangunan perangkat lunak yang digunakan untuk implementasi klasifikasi teks dengan *Multiclass* SVM pada artikel berbahasa indonesia

menggunakan model *waterfall sommerville* [3] yang melakukan pendekatan secara sistematis dan berurutan dalam pembangunan perangkat lunak yang dirubah sesuai dengan kebutuhan penelitian meliputi proses sebagai berikut.



Gambar 1.2 Diagram Waterfall

a. Kebutuhan

Langkah ini merupakan analisa terhadap kebutuhan sistem. Pada tahap ini dilakukan pengumpulan kebutuhan penelitian secara lengkap tentang klasifikasi teks menggunakan *Multiclass* SVM. Data yang digunakan dalam format *.txt atau *.csv. Semua hal tersebut akan ditetapkan secara rinci dan berfungsi sebagai spesifikasi sistem.

b. Analisis

Setelah kebutuhan data dan pemroses telah dikumpulkan, maka tahap selanjutnya adalah melakukan *analysis*. Analisa yang dilakukan pada penelitian ini adalah analisa pendekatan metode *multiclass* SVM melalui alur diagram (*flowchart*), analisa kebutuhan perangkat keras dan perangkat lunak sistem dan analisa kebutuhan pengguna sistem.

c. Desain

Pada tahapan ini dilakukan penuangan pikiran dan perancangan sistem terhadap solusi dari permasalahan yang ada dengan menggunakan perangkat pemodelan sistem seperti diagram alir data (*Data Flow Diagram*), perancangan struktur menu, perancangan antarmuka dan perancangan jaringan semantik.

d. Pengkodean

Pada tahap ini desain program yang telah dibuat kemudian diimplementasikan ke dalam bentuk kode bahasa pemrograman diantaranya adalah tahap *preprocessing* (case folding, filtering,tokenisasi,dan stopwords) yang kemudian akan dilakukan *feature extraction* menggunakan TF-IDF. Hasil dari pembobotan TF-IDF ini akan dirubah ke dalam bentuk vektor untuk dilakukan pelatihan. Dari pelatihan ini akan menghasilkan model prediksi yang akan digunakan dalam klasifikasi teks. Setiap yang telah dirancang pada DFD akan diimplementasikan pada *coding.* Pada penelitian ini menggunakan bahasa pemrograman *python* 2.7

e. Pengujian

Dalam tahap ini, sistem sudah siap digunakan. Selain itu juga memperbaiki error yang tidak ditemukan pada tahap pembuatan. Dalam tahap ini juga dilakukan pengembangan sistem seperti penambahan fitur dan fungsi baru yang mungkin muncul kemudian sesuai dengan kebutuhan pengguna.

1.6 Sistematika Penulisan

Sistematika penulisan laporan ini disusun untuk memberikan gambaran umum tentang penelitian dalam tugas akhir yang dilaksanakan. Sistematika penulisan tugas akhir ini adalah sebagai berikut :

BAB I PENDAHULUAN

Pada bab ini dijelaskan latar belakang dari penelitian klasifikasi teks pada artikel berita berbahasa Indonesia. Tujuan dan ruang lingkup dari tugas akhir memberikan penjelasan mengenai hasil yang ingin diketahui dan batasan-batasan yang ada dalam melakukan penelitian.

BAB II LANDASAN TEORI

Pada bab ini berisi tentang landasan teori dan metode yang digunakan pada tugas akhir ini dalam melakukan klasifikasi teks pada artikel berbahasa Indonesia. Pembahasan dimulai dengan penjelasan mengenai berita dan jenis – jenis berita, dilanjutkan dengan penjelasan tentang klasifikasi teks dan metode – metode yang digunakan dalam melakukan klasifikasi teks.

BAB III ANALISIS DAN PERANCANGAN

Pada bab ini berisi mengenai perancangan untuk melakukan klasifikasi teks pada artikel. Klasifikasi dilakukan dengan menentukan kategori dari semua dokumen artikel testing yang ada. Perancangan klasifikasi teks ini meliputi persiapan dokumen artikel, tahap case *folding*, *filtering*, tokenisasi, *stopword*, pembuatan term documents matrix dan klasifikasi teks menggunakan machine learning *Multiclass* SVM. Pendekatan *Multiclass* SVM yang digunakan, yaitu *one-against-all, one-against-one*, dan *error correcting output code*.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Pada bab ini berisi tentang tahapan yang dilakukan dalam penelitian secara garis besar dimulai dari tahap persiapan sampai pada penarikan kesimpulan, pada tahap ini klasifikasi teks akan diuji dengan menggunakan data testing.

BAB V KESIMPULAN DAN SARAN

Pada bab ini menjelaskan tentang kesimpulan dari penelitian dan saran yang dapat dijadikan masukan untuk pengembangan penelitian pada kasus yang sama selanjutnya.