SKRIPSI

ANALISIS KLASIFIKASI TEKS MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE*

Alwy Bathia Ramadhan

NPM: 6161801017

PROGRAM STUDI MATEMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2022

FINAL PROJECT

TEXT CLASSIFICATION ANALYSIS USING SUPPORT VECTOR MACHINE METHOD

Alwy Bathia Ramadhan

NPM: 6161801017

DEPARTMENT OF MATHEMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2022

LEMBAR PENGESAHAN

ANALISIS KLASIFIKASI TEKS MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE*

Alwy Bathia Ramadhan

NPM: 6161801017

Bandung, 5 Agustus 2022

Menyetujui,

Pembimbing 1	Pembimbing 2
	
Agus Sukmana, M.Sc.	**Dr. Erwinna Chendra**

Ketua Tim Penguji	Anggota Tim Penguji
	
Dr. Livia Owen	**Robyn Irawan, M.Sc.**

Mengetahui,

Ketua Program Studi

Dr. Livia Owen

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

ANALISIS KLASIFIKASI TEKS MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE*

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 5 Agustus 2022

Alwy Bathia Ramadhan
NPM: 6161801017

ABSTRAK

Perkembangan teknologi mengakibatkan bertambahnya jumlah pengguna internet, hal tersebut diiringi dengan peningkatan jumlah data teks atau dokumen. Data teks tersebut bisa berupa ulasan, opini, dan berita yang dapat dimanfaatkan untuk berbagai kebutuhan dengan mengambil informasi dari data teks tersebut. Salah satu cara untuk mendapatkan informasi dari data teks dalam jumlah besar adalah dengan mengklasifikasikan teks tersebut ke dalam sentimen yang berbeda. Kendala yang dialami dalam mengklasifikasikan teks adalah jumlah data yang sangat besar sehingga sulit untuk diklasifikasikan secara manual. Di sini peran algoritma pembelajaran mesin untuk mempermudah penelitian klasifikasi teks. Menurut penelitian sebelumnya dari beberapa algoritma klasifikasi seperti *Logistic Regression* (LR), Pohon Keputusan, *Naive Bayes*, *Random Forest*, dan *Support Vector Machine* (SVM), dimana SVM mampu memberikan kinerja prediktif tekstual yang terbaik di antara metode lainnya. Hal tersebut dikarenakan SVM memiliki keunggulan dalam menangani data yang tidak terstruktur dan memiliki dimensi tinggi. Pada skripsi ini digunakan dua himpunan data, yaitu pesan Twitter mewakili data tidak terstruktur dan ulasan Shopee mewakili data terstruktur yang masing-masing berjumlah 3000 data teks. Dari hasil yang didapat pada penelitian ini, algoritma SVM mampu mengklasifikasikan data teks terstruktur dan tidak terstruktur dengan cukup baik. Hal ini dapat dilihat dari nilai f_1-*score* yang didapat, yaitu 61% untuk data pesan Twitter dan 81,43% untuk data ulasan Shopee. Penggunaan fungsi kernel dan jumlah data juga mempengaruhi perform SVM. Untuk kedua himpunan data teks ini didapat fungsi kernel RBF dan Linear memiliki hasil performa yang lebih baik dibandingkan kernel Polinomial dan performa akan semakin baik jika memperbanyak jumlah data latih.

Kata-kata kunci: Klasifikasi teks, *Support Vector Machine*, fungsi kernel, data terstruktur, data tidak terstruktur.

ABSTRACT

Technological developments have resulted in an increase in the number of internet users, this is accompanied by an increase in the number of text data or documents. The text data can be in the form of reviews, opinions, and news that can be used for various needs by taking information from the text data. One way to get information from a large amount of data text is to classify the text into different sentiments. The obstacle experienced in classifying text is the large amount of data that makes it difficult to classify manually. Here the role of machine learning algorithms is to facilitate text classification research. According to previous research from several classification algorithms such as Logistic Regression (LR), Decision Tree, Naive Bayes, Random Forest, and Support Vector Machine (SVM), SVM is able to provide the best predictive textual performance among other methods. This is because SVM has the advantage of handling unstructured and high-dimensional data. In this thesis, two data sets are used, namely Twitter messages representing unstructured data and Shopee reviews representing structured data, each of which performs 3000 text data. From the results obtained in this study, the SVM algorithm is able to classify structured and unstructured text data quite well. This can be seen from the f_1-*score* obtained, which is 61% for Twitter message data and 81,43% for Shopee data reviews. The use of kernel functions and the amount of data also affects running SVM. For these two text data sets, the RBF and Linear kernel functions have better performance results than the Polynomial kernel and the performance will improve if you increase the amount of training data.

Keywords: Text classification, support vector machine, kernel function, structured data, unstructured data.

KATA PENGANTAR

Alhamdulillahi rabbil 'alamin, puji dan syukur kepada Allah SWT. atas nikmat, rahmat, dan hidayah-Nya berupa akal pikiran dan kesehatan yang diberikan sehingga penulis dapat menyelesaikan skripsi dengan judul "Analisis Klasifikasi Teks Menggunakan Metode *Support Vector Machine*" sebagai salah satu syarat untuk menyelesaikan studi di Program Studi Matematika, Fakultas Teknologi Informasi dan Sains, Universitas Katolik Parahyangan. Dengan harapan skripsi ini dapat memberikan manfaat bagi pembaca.

Pada kesempatan kali ini, penulis ingin mengucapkan terima kasih kepada pihak-pihak yang telah memberikan dukungan serta bantuan secara langsung maupun tidak langsung, yaitu:

1. Bapak Agus Sukmana, M.Sc selaku dosen pembimbing utama yang telah memberikan ilmu, motivasi, bantuan, serta senantiasa memberikan waktu dan arahan yang bermanfaat dalam penyusunan skripsi ini.
2. Ibu Dr. Erwinna Chendra selaku dosen pembimbing pendamping yang telah memberikan ilmu, motivasi, bantuan, serta senantiasa memberikan waktu dan arahan yang bermanfaat dalam penyusunan skripsi ini.
3. Ibu Dr. Livia Owen selaku dosen penguji yang telah memberikan saran serta kritik untuk skripsi ini.
4. Bapak Robyn Irawan, M.Sc selaku dosen penguji yang telah memberikan saran serta kritik untuk skripsi ini.
5. Ayah, Ibu, Naya, dan Indi yang selalu memberikan dukungan, bantuan, nasihat, dan doa.
6. Novaldi Dwi Putra yang telah membantu, mendukung, dan berjuang bersama semasa kuliah hingga lulus bersama.
7. Orlin Monica Kencana yang telah menyediakan kos miliknya sebagai tempat belajar bersama dan istirahat sepulang kuliah.
8. Akmal dan Nadia yang telah menjadi tempat cerita, mendukung, dan menasehati penulis pada saat penulisan skripsi.
9. Novaldi, Orlin, Dimas, dan Adrian selaku keluarga besar segiti-gay yang selalu memberikan dukungan, nasihat, dan selalu menemani penulis
10. Odi dan Mei selaku BURT MG 2018 yang telah menemani penulis semasa awal perkuliahan.
11. Teman-teman matematika 2018 yang selalu mendukung satu sama lain dan memberikan pengalaman hidup selama perkuliahan.
12. Seluruh dosen, tata usaha, dan pekarya FTIS terima kasih atas ilmu serta bantuannya selama perkuliahan.
13. Seluruh pihak lainnya yang tidak dapat disebutkan satu per satu. Terima kasih banyak atas segala dukungan dan bantuannya.

Penulis menyadari bahwa penulisan skripsi ini masih terdapat suatu kekurangan. Oleh karena itu, penulis menerima segala kritik dan saran yang membangun dari pembaca untuk penyempurnaan skripsi ini. Semoga skripsi ini dapat bermanfaat bagi pembaca.

Bandung, Agustus 2022

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi menjadikan internet semakin mudah diakses dan digunakan oleh masyarakat, yang mengakibatkan data seperti data teks atau dokumen bertambah dengan sangat cepat. Data teks tersebut dapat dimanfaatkan untuk mendapatkan berbagai informasi dan salah satu cara untuk mendapatkan informasi dari data teks adalah dengan melakukan klasifikasi ke dalam kelas-kelas tertentu agar informasi bisa disimpulkan dari hasil klasifikasi tersebut. Penelitian klasifikasi teks menjadi sangat penting seiring berkembangnya laman seperti blog, forum diskusi, dan berbagai macam media sosial [1].

Penelitian terkait klasifikasi teks dapat diterapkan pada berbagai bidang. Pada bidang politik, seperti pemilihan umum, kita dapat memahami opini masyarakat terhadap calon dengan melakukan klasifikasi terhadap data teks terkait calon tersebut. Contohnya seperti opini positif masyarakat yang mengakibatkan masyarakat ingin memlilih calon tersebut atau sebaliknya. Pemerintahan juga dapat memanfaatkan klasifikasi teks untuk mengetahui permasalahan umum yang sedang terjadi di masyarakat seperti opini masyarakat terkait kebijakan, konflik sosial, dan berbagai macam masalah lainnya. Pada bidang bisnis klasifikasi teks dapat digunakan agar perusahaan dapat mengetahui kelebihan dan kekurangan produknya dengan melihat klasifikasi dari pendapat masyarakat terkait produk atau barang yang dijual.

Aplikasi pada telepon genggam merupakan salah satu produk yang sering mendapatkan ulasan dari penggunanya. Ulasan ini biasanya terdiri dari data teks dan nilai bintang dari 1 sampai 5. Ada beberapa kasus di mana nilai bintang yang diberikan tidak sesuai dengan apa yang ditulis pada ulasan, seperti memberikan bintang 5 namun memberikan beberapa keluhan terkait aplikasi yang digunakan. Media sosial yang bisa digunakan untuk memberikan pendapat atau opini berupa teks adalah Twitter. Twitter adalah platform publikasi yang cepat dan mudah, dengan begitu menjadikannya medium komunikasi bagi banyak orang, Twitter juga telah memainkan peran utama dalam peristiwa sosial politik [2]. Media sosial seperti Twitter dapat digunakan untuk mendapatkan berbagai macam data teks dengan jumlah yang besar.

Tujuan dari klasifikasi teks adalah untuk mengetahui sentimen dari setiap data teks. Nilai ulasan yang tidak sesuai dengan isi ulasan dan data teks dalam jumlah besar merupakan contoh dari permasalahan dalam klasifikasi teks yang mengakibatkan klasifikasi teks akan sangat rumit jika dilakukan secara manual. Disinilah peran algoritma pembelajaran dapat diterapkan untuk mempermudah dalam penelitian klasifikasi teks.

Pembelajaran mesin memiliki beberapa algoritma yang bisa digunakan untuk klasifikasi teks seperti *Logistic Regression* (LR), Pohon Keputusan, *Naive Bayes*, *Random Forest* dan *Support Vector Machine* (SVM). Di antara semua algoritma klasifikasi tersebut SVM memiliki keunggulan dalam menangani data yang tidak terstruktur dan memiliki dimensi yang tinggi, maka dari itu analisis tekstual menggunakan SVM memiliki kinerja prediktif terbaik [3]. Data teks berupa opini dan ulasan cenderung memiliki dimensi yang cukup tinggi, karena semakin banyak opini atau ulasan yang ditulis maka akan meningkatkan dimensi dari data yang digunakan. Pada skripsi ini SVM akan digunakan untuk memprediksi sentimen dari masing-masing data teks tanpa harus membaca

secara keseluruhan dari data teks yang ada.

1.2 Rumusan Masalah

Berdasarkan hal-hal yang dibahas pada latar belakang, terdapat beberapa permasalahan yang akan diselesaikan, yaitu:

1. Bagaimana perbandingan performa SVM terhadap fungsi kernel yang berbeda?
2. Bagaimana pengaruh jumlah data latih yang berbeda terhadap performa SVM?
3. Bagaimana analisis performa SVM terhadap data terstruktur dan tidak terstruktur?

1.3 Tujuan

Berdasarkan rumusan masalah tersebut, tujuan dari skripsi ini adalah sebagai berikut:

1. Mengetahui performa SVM dalam klasifikasi teks menggunakan fungsi kernel berbeda.
2. Membandingkan performa SVM berdasarkan jumlah data latih yang digunakan.
3. Membandingkan kemampuan algoritma SVM dalam melakukan klasifikasi teks terhadap data teks terstruktur dan tidak terstruktur.

1.4 Batasan Masalah

Batasan masalah dari makalah ini:

1. Data teks pesan Twitter hanya berfokus pada topik PJJ dan sejumlah 3000 pesan berupa teks.
2. Data teks ulasan aplikasi Shopee sejumlah 3000 ulasan berupa teks.
3. Pesan Twitter dan ulasan aplikasi Shopee hanya diklasifikasikan ke dalam dua kelas yaitu sentimen positif dan negatif.

1.5 Metodologi

Penyusunan makalah ini menerapkan metodologi sebagai berikut.

1. Melakukan pengambilan data teks pesan Twitter dan ulasan aplikasi Shopee.
2. Melakukan prapengolahan data agar bisa digunakan untuk algoritma SVM.
3. Mengimplementasikan algoritma SVM untuk klasifikasi teks.
4. Menganalisis performa algoritma SVM untuk klasifikasi teks.

1.6 Sistematika Pembahasan

Bab 1 Pendahuluan
Bab ini berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penelitian, dan sistematika pembahasan.

Bab 2 Landasan Teori
Bab ini berisi teori-teori klasifikasi teks, metode prapengolahan data teks, *Support Vector Machine*, fungsi kernel, dan analisis performa algoritma klasifikasi.

Bab 3 Metode Penelitian
Bab ini berisi data yang digunakan, hal apa saja yang akan dianalisis, langkah-langkah dari prapengolahan data sampai implementasi algoritma SVM, dan penjelasan evaluasi performa SVM.

Bab 4 Hasil dan Pembahasan
Bab ini berisi hasil klasifikasi teks menggunakan data terstruktur dan tidak terstruktur. Hasil-hasil tersebut meliputi fungsi kernel yang digunakan, perbandingan jumlah terhadap performa, dan perbandingan antara data terstruktur dan tidak terstruktur.
Bab 5 Kesimpulan dan Saran
Berisi terkait kesimpulan dari penerapan algoritma SVM untuk klasifikasi teks dan saran untuk penelitian lebih lanjut.