KLASIFIKASI 4 BAHASA SERUMPUN MENGGUNAKAN SUPPORT VECTOR MACHINE

Skripsi

oleh
ANGELA CHANDRA SANGKALA
71150010

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2020

KLASIFIKASI 4 BAHASA SERUMPUN MENGGUNAKAN SUPPORT VECTOR MACHINE

Skripsi

Diajukan kepada Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

ANGELA CHANDRA SANGKALA
71150010

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2020

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

KLASIFIKASI 4 BAHASA SERUMPUN MENGGUNAKAN SUPPORT VECTOR MACHINE

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 9 Januari 2020

ANGELA CHANDRA SANGKALA

71150010

HALAMAN PERSETUJUAN

Judul Skripsi	:	SISTEM IDENTIFIKASI 4 BAHASA SERUMPUN MENGGUNAKAN SUPPORT VECTOR MACHINE
Nama Mahasiswa	:	ANGELA CHANDRA SANGKALA
N I M	:	71150010
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Gasal
Tahun Akademik	:	2019/2020

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 27 November 2019

Dosen Pembimbing I	Dosen Pembimbing II
Lucia Dwi Krisnawati, Dr. Phil.	Aditya Wikan Mahastama, S.Kom., M.Cs.

HALAMAN PENGESAHAN

KLASIFIKASI 4 BAHASA SERUMPUN MENGGUNAKAN SUPPORT VECTOR MACHINE

Oleh: ANGELA CHANDRA SANGKALA / 71150010

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 13 Desember 2019

Yogyakarta, 9 Januari 2020
Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, Dr. Phil.
2. Aditya Wikan Mahastama, S.Kom., M.Cs.
3. Yuan Lukito, S.Kom., M.Cs.
4. Danny Sebastian, S.Kom., M.M., M.T.

Dekan	Ketua Program Studi
(Restyandito,S.Kom.,MSIS.,Ph.D.)	(Gloria Virginia,Ph.D.)

UCAPAN TERIMA KASIH

Pertama, penulis mengucapkan puji syukur dan terima kasih kepada Tuhan Yang Maha Esa atas berkat dan rahmat-Nya, penulis mampu menyelesaikan tugas akhir berjudul "Klasifikasi 4 Bahasa Serumpun Menggunakan *Support Vector Machine*".

Meskipun dalam proses menyelesaikan tugas akhir ini penulis mengalami beberapa hambatan, Penulis mendapatkan bantuan dan dukungan dari berbagai pihak sehingga Penulis mampu menyelesaikan tugas akhir ini. Oleh sebab itu, Penulis ingin mengucapkan terima kasih kepada:

1. Kedua orangtua dan saudara dari seluruh keluarga besar yang selalu mendukung dan mendoakan penulis untuk dapat menyelesaikan tugas akhir dengan baik dan tepat waktu.
2. Bapak Restyandityo, S.Kom., MSIS, Ph.D. selaku dekan Fakultas Teknologi Informasi Universitas Kristen Duta Wacana.
3. Ibu Gloria Virginia, Ph. D. selaku ketua program studi Informatika Universitas Kristen Duta Wacana.
4. Bapak Willy Sudiarto Raharjo, S.Kom., M.Cs selaku dosen wali penulis.
5. Ibu Lucia Dwi Krisnawati, Dr. Phil. Selaku dosen pembimbing I yang telah memberikan waktu kepada penulis sehingga dapat melakukan konsultasi mengenai tugas akhir dan memberikan bimbingan, penjelasan dan masukan selama pengerjaan tugas akhir.
6. Bapak Aditya Wikan Mahastama, S.Kom., M.Cs. selaku dosen pembimbing II yang telah memberikan bimbingan kepada penulis selama menyelesaikan tugas akhir ini.
7. Teman-teman seangkatan dan seprodi terutama Maria Theresa R, Niluh Muryanti, Merla Nindya, dan Vievin Efendy yang selalu saling mendukung satu sama lain dalam menyelesaikan tugas akhir.

8. Semua pihak yang tidak dapat disebutkan satu persatu yang selalu memberikan semangat dan doa demi kelancaran pengerjaan tugas akhir baik secara langsung maupun tidak langsung untuk penulis.

Penulis menyadari bahwa masih terdapat banyak kekurangan dalam penulisan maupun pembahasan tugas akhir ini. Akhir kata penulis mengucapkan terima kasih kepada semua pihak yang ikut ambil andil dalam penelitian tugas akhir ini. Penulis juga berharap semoga penelitian tugas akhir ini dapat bermanfaat bagi pembaca.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Teknologi informasi berkembang pesat saat ini. Penggunaan internet juga sudah menjadi kebutuhan di era digital untuk membantu menyelesaikan berbagai pekerjaan. Melalui internet pula orang di seluruh dunia dapat melakukan interaksi dan saling bertukar informasi. Komunikasi antar bahasa melalui media digital pun semakin mudah dengan berkembangnya aplikasi untuk Pemrosesan Bahasa Alami. Pemrosesan Bahasa Alami memungkinkan sistem untuk memahami bahasa alami manusia sehingga sistem dapat melakukan berbagai proses yang dapat menghasilkan berbagai informasi. Pemrosesan Bahasa Alami yang termasuk dalam sistem cerdas membuat semua proses dapat dilakukan secara otomatis. Pemrosesan Bahasa Alami berkembang di banyak negara sehingga aplikasi yang berkembang juga mendukung berbagai macam bahasa yang digunakan di seluruh dunia.

Berkaitan dengan banyaknya bahasa yang digunakan dalam komunikasi dunia maya, maka dibutuhkan sistem yang dapat mengenali bahasa. Sistem Identifikasi Bahasa dapat digunakan dalam berbagai macam aplikasi seperti mesin penerjemah multi-bahasa, kategorisasi topik, temu kembali informasi, pembuatan korpus bahasa, peringkasan otomatis, mesin tanya jawab (*chatbot*) serta dapat digunakan saat tahap prapemrosesan untuk pembangunan aplikasi-aplikasi tersebut. *Google Translate* merupakan salah satu mesin penerjemah untuk menerjemahkan suatu bahasa ke bahasa yang lain dan termasuk salah satu mesin penerjemah yang populer saat ini. *Google Translate* dapat secara otomatis mengenali bahasa yang digunakan saat menerima masukan berupa teks. Selain mesin penerjemah, contoh penggunaan Sistem Identifikasi Bahasa dapat diterapkan untuk melakukan kategorisasi teks yang memiliki topik tertentu misalnya untuk mengkategorisasikan topik dari berita apakah termasuk berita olahraga, politik atau kategori lainnya. Setiap bahasa memiliki ciri khas masing-masing yang dapat digunakan

membedakan bahasa yang satu dengan yang lain baik bahasa antarnegara maupun antardaerah. Sebagai contoh masing-masing bahasa memiliki kata-kata tertentu yang banyak dipakai dalam kehidupan sehari-hari dan berbeda dengan bahasa lain meskipun memiliki arti yang sama.

Sistem identifikasi dan klasifikasi bahasa sudah banyak dikembangkan. Namun, di Indonesia sendiri yang merupakan negara dengan berbagai macam budaya dan bahasa, sistem yang dapat melakukan klasifikasi untuk bahasa daerah masih sangat minim jumlahnya. Hal ini dikarenakan bahasa Indonesia sendiri termasuk bahasa *under resource* sedangkan bahasa daerah seperti bahasa Jawa dan Sunda merupakan bahasa *critical resource* yang berarti masih minimnya sumber yang menyediakan informasi dalam bahasa-bahasa tersebut.

Melalui penelitian ini akan dibangun sistem identifikasi dan klasifikasi bahasa untuk bahasa Indonesia, Jawa, Sunda dan Minang. Penelitian ini merupakan penelitian lanjutan dari penelitian yang telah dilakukan oleh Krisnawati, L., Sentosa, F., & Mahastama, A. (2019). Jika pada penelitian sebelumnya berfokus pada identifikasi untuk bahasa Indonesia dan Jawa serta menggunakan karakter n-gram, pada penelitian ini akan digunakan metode *Support Vector Machine* untuk melakukan pengklasifikasian untuk empat bahasa yaitu bahasa Indonesia, Jawa, Sunda dan Minang.

1.2 Rumusan Masalah

Berdasarkan uraian di atas, masalah yang akan diteliti dalam penelitian ini adalah sebagai berikut

1. Bagaimana penerapan *Support Vector Machine* dalam klasifikasi bahasa ?
2. Bagaimana proses untuk menemukan fitur seminimal mungkin yang dapat menghasilkan identifikasi bahasa dengan akurasi tinggi untuk Sistem Klasifikasi Bahasa ?
3. Bagaimana cara untuk mengukur tingkat akurasi dari Sistem Klasifikasi Bahasa ?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sistem berfokus untuk mengklasifikasikan bahasa dari suatu kalimat atau teks pendek yang terdiri kurang dari 50 kata maupun teks panjang yang terdiri lebih dari 50 kata. Sistem yang akan dikembangkan untuk penelitian ini adalah sistem berbasis *desktop*. Teks yang dapat diklasifikasikan adalah teks berbahasa Indonesia, Jawa, Sunda dan Minang.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk membangun sistem yang mampu melakukan klasifikasi bahasa untuk teks berbahasa Indonesia, Jawa, Minang dan Sunda dengan tepat dari suatu masukan berupa teks yang terdiri kurang dari 50 kata dan teks yang terdiri lebih dari 50 kata.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sistem dapat melakukan klasifikasi secara tepat untuk dokumen yang menggunakan bahasa Indonesia, Jawa, Sunda dan Minang, serta dapat menjadi pustaka prapemrosesan untuk berbagai aplikasi pemrosesan bahasa natural seperti mesin penerjemah multi-bahasa otomatis, kategorisasi topik, temu kembali informasi dan mesin penanya jawab otomatis.

1.6 Metodologi Penelitian

Dalam penyelesaian penelitian ini, terdapat beberapa metodologi penelitian, antara lain

1.6.1 Pengumpulan Data

Tahap pengumpulan data merupakan tahap dimana dilakukan pengumpulan data dokumen latih serta dokumen uji berupa dokumen teks berbahasa Indonesia, Jawa, Sunda dan Minang. Dokumen latih untuk kelas bahasa Indonesia dan bahasa Jawa menggunakan korpus dari penelitian sebelumnya yaitu artikel Bahasa Indonesia dan Trawaca. Sedangkan dokumen latih untuk bahasa Sunda diambil dari halaman web Wikipedia Bahasa Sunda yang berisi kata-kata dalam bahasa Sunda, kemudian kata-kata tersebut akan disimpan ke dalam file berekstensi *.txt* dan digunakan sebagai korpus bahasa, begitu pula dengan data latih untuk bahasa

Minang. Kemudian untuk dokumen uji, dokumen uji Bahasa Indonesia diambil dari artikel di halaman web liputan6.com, bahasa Jawa dari Trawaca dan halaman web www.katapengertian.com, bahasa Minang dari Wikipedia bahasa Minang dan bahasa Sunda dari halaman web basasunda.com.

1.6.2 Prapemrosesan (*Preprocessing)*

Pada tahap prapemrosesan dilakukan beberapa proses terhadap dokumen latih dan dokumen uji sehingga dokumen dapat dapat diproses ke dalam algoritma secara efektif. Pada umumnya tahap prapemrosesan memiliki beberapa tahapan yang termasuk menghilangkan stopwords atau kata yang sering muncul, namun pada penelitian ini stopwords tidak akan dihilangkan melainkan akan digunakan sebagai fitur untuk proses klasifikasi.

1.6.2.1 Tokenisasi

Tahap tokenisasi pada prapemrosesan dimulai dengan menghapus semua tanda baca, tanda baca dan angka. Kemudian untuk dokumen yang teksnya mengandung diakritik dinormalisasi dengan standar UTF-8. Setelah itu tokenisasi dilakukan dimana teks dalam dokumen dipecah menjadi satuan kata.

1.6.3 Ekstraksi Fitur (*Feature Extraction*)

1.6.3.1 Sortir Profil

Terdapat 2 macam faktor yang dapat digunakan dalam pengembangan sistem untuk identifikasi dan klasifikasi bahasa, yaitu berdasarkan fitur atau profil. Pada penelitian ini kedua faktor tersebut digunakan secara *hybrid* yang berarti menggunakan baik profil maupun fitur. Pada penelitian ini profil dibangun untuk menentukan fitur bahasa. Sortir profil bertujuan untuk mendapatkan kata-kata yang merepresentasikan kelas bahasa dan hasil dari sortir profil akan digunakan sebagai fitur bahasa. Hasil sortir profil merupakan kata-kata yang paling banyak muncul dalam dokumen di suatu kelas bahasa. Hasil sortir profil berupa 100 kata yang memiliki frekuensi kemunculan paling banyak pada masing-masing kelas sehingga pada penelitian ini hasil sortir profil berjumlah 400 karena menggunakan 4 kelas.

1.6.3.2 Pengekstrasian Fitur Bahasa (*Feature Extraction*)

Pengekstrasian fitur bahasa dilakukan setelah mendapatkan hasil dari sortir profil. Hasil dari sortir profil yang berjumlah 400 ini kemudian dihitung bobotnya menggunakan metode pembobotan TF-CHI sehingga didapatkan nilai bobotnya. Hasil perhitungan menggunakan TF-CHI untuk masing-masing kata atau token yang termasuk hasil sortir profil digunakan sebagai bobot token yang kemudian dapat dikonversikan menjadi vektor pada klasifikator. Fitur dibutuhkan untuk menjadi faktor yang digunakan sebagai pembanding dalam proses klasifikasi.

1.6.4 Pelatihan Sistem (*System Training*)

Support Vector Machine (SVM) termasuk algoritma *supervised learning* yang termasuk dalam algoritma *machine learning* yang berarti sistem membutuhkan proses belajar dari fitur data latih yang sudah diberi label untuk masing-masing kelas. Tahap pelatihan sistem merupakan proses dimana sistem dilatih untuk memahami dan membedakan karakteristik dari masing-masing kelas sehingga dapat membedakan teks masukan yang berbeda. Pelatihan sistem dilakukan dengan membangun model SVM dan pada penelitian ini penulis menggunakan library SVM dari *scikit-learn*. *Scikit-learn* menyediakan berbagai tools untuk implementasi sistem yang menggunakan algoritma SVM. SVM memiliki beberapa macam jenis kernel yang dapat digunakan sesuai jenis data yang digunakan untuk melakukan klasifikasi dan *scikit-learn* juga menyediakan pilihan kernel yang siap digunakan. Model yang telah dibangun kemudian dilatih menggunakan fitur yang sudah dihasilkan dari prapemrosesan.

1.6.5 Pengujian (*Testing*)

Tahap pengujian atau *testing* merupakan proses dimana sistem yang sudah dibangun diuji untuk melakukan klasifikasi pada dokumen uji yang telah dikumpulkan. Pengujian terhadap dokumen uji dilakukan menggunakan model klasifikator yang sudah melewati tahap pembelajaran berdasarkan fitur pada dokumen latih. Pada penelitian ini digunakan 40 dokumen uji dimana masing-masing dokumen uji untuk masing-masing kelas berjumlah 10 dokumen. Pengujian dilakukan 3 kali dengan perbedaan pada rata-rata jumlah token atau panjang

dokumen uji. Pengujian pertama dilakukan dengan variasi panjang dokumen uji 500 sampai kurang lebih 3000 token, pengujian kedua dilakukan dengan panjang dokumen 6 sampai dengan kurang lebih 200 token dan pengujian ketiga 6 sampai dengan kurang lebih 1000 token.

1.6.5.1.1 Evaluasi

Tahap evaluasi merupakan proses terakhir dalam implementasi sistem klasifikasi bahasa menggunakan SVM dimana pada proses ini akan dilakukan perhitungan menggunakan metode akurasi. Setelah dilakukan perhitungan, kemudian nilai akurasi menjadi tolok ukur apakah sistem sudah dapat melakukan klasifikasi dengan akurat atau belum. Pada penelitian ini, sistem klasifikasi dianggap sudah layak untuk digunakan apabila nilai akurasi mencapai angka 50% atau lebih.

1.7 Sistematika Penulisan

Penulisan laporan ini terdiri dari beberapa bab:

BAB 1 PENDAHULUAN, pada bab ini dijelaskan latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, metodologi penelitian dan sistematika penulisan.

BAB 2 TINJAUAN PUSTAKA, pada bab ini terdapat landasan teori dan tinjauan pustaka yang berisi penjelasan mengenai teori-teori serta penelitian yang telah dilakukan terkait klasifikasi bahasa menggunakan Support Vector Machine.

BAB 3 PERANCANGAN SISTEM, pada bab ini dibahas mengenai perancangan sistem serta tahapan pengembangan sistem.

BAB 4 IMPLEMENTASI DAN ANALISIS SISTEM, pada bab ini dibahas mengenai hasil implementasi dan analisis dari pengembangan sistem yang telah dibuat.

BAB 5 KESIMPULAN, pada bab ini dibahas mengenai kesimpulan dan saran dari keseluruhan hasil penelitian yang telah dilakukan.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil penelitian yang dilakukan, maka didapatkan nilai akurasi sebesar 70% untuk pengujian dokumen panjang, 40% untuk pengujian dokumen pendek dan 62.5% dari pengujian sistem klasifikasi untuk 4 kelas bahasa menggunakan *Support Vector Machine*. Hal ini membuktikan bahwa *Support Vector Machine* dengan kernel RBF yang menggunakan pendekatan *hybrid* antara profil dan fitur dapat digunakan untuk melakukan klasifikasi multi-kelas yang dalam penelitian ini berfokus pada klasifikasi multi-bahasa. Panjang dokumen berpengaruh jika dilihat dari hasil pengujian dimana dokumen panjang dapat diklasifikasi secara lebih tepat oleh sistem. Berdasarkan hasil pengujian yang telah dilakukan, sistem klasifikasi menggunakan *Support Vector Machine* (SVM), hasil klasifikasi terkadang masih kurang tepat karena beberapa faktor seperti kurang seimbangnya dokumen latih antara kelas bahasa yang satu dengan yang lain. Selain itu sumber dokumen latih untuk bahasa Minang dan Sunda masih sangat terbatas sehingga variasi *stopwords* juga kurang sehingga beberapa kata yang bukan *stopwords* dan tidak merepresentasikan bahasa tersebut ikut menjadi fitur yang digunakan dalam proses klasifikasi. Normalisasi diakritik menggunakan standar UTF-8 ternyata kurang tepat digunakan dan membuat beberapa kata tidak dapat diklasifikasikan oleh sistem dengan tepat. Penggunaan nilai TF-CHI dan nilai ASCII dapat digunakan untuk melakukan klasifikasi meskipun beberapa nilai antara fitur yang satu dengan fitur dari kelas bahasa yang lain hanya memiliki selisih yang sedikit sehingga memengaruhi hasil klasifikasi. Namun untuk dokumen panjang dengan nilai 70% menunjukkan bahwa sistem sudah mampu melakukan klasifikasi yang apabila beberapa faktor diperbaiki akan dapat meningkatkan kinerja sistem.

5.2 Saran

Dari kekurangan yang terdapat pada sistem seperti yang telah disebutkan di kesimpulan, maka saran yang dapat diberikan penulis adalah dengan cara memperbanyak dokumen latih terutama dalam konteks penelitian ini yaitu Bahasa

Minang dan Bahasa Sunda sehingga hasil sortir profil bahasa dapat menghasilkan *stopwords* yang lebih bervariasi. Pada normalisasi diakritik juga bisa dilakukan normalisasi manual sehingga huruf yang mengandung diakritik dapat disesuaikan dengan penggunaan. Selain itu pada proses ekstraksi fitur, dibutuhkan metode untuk menghasilkan nilai yang membedakan lebih jelas antara fitur dari kelas yang satu dengan yang lain sehingga saat diproses ke dalam klasifikator tidak terlalu bias dengan nilai fitur dari kelas yang lain atau dengan nilai fitur yang memiliki selisih nilai sedikit.

DAFTAR PUSTAKA

Baldwin, T., & Lui, M. (2010). Language Identification: The Long and the Short of the Matter. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, (hal. 229-237).

Mahastama, A. W., & Krisnawati, L. D. (2017). Histogram Peak-Based Binarization for Historical Documents. *International Conference on Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS) 2017)* (hal. 93-98). Yogyakarta: IEEE.

Man, L., Sam-Yuan, S., Hwee-Boon, L., & Chew-Lim, T. (2005). A Comparative Study on Term Weighting Schemes for Text Categorization. *Proceedings of the International Joint Conference on Neural Network*, (hal. 546-551).

Selamat, A., & Akosu, N. (2014, December). Word-Length Algorithm for Language Identification of Under-Resourced Languages. *Journal of King Saud University – Computer and Information Science*, 457-469.

Sukma, A., Santoso, B. P., Ramadhan, D., Wiraswari, N. M., & Sari, T. R. (t.thn.). Klasifikasi Dokumen Bahasa Jawa Menggunakan Metode N-gram.

Takçı, H., & Ekinci, E. (2011). Minimal Feature Set in Language Identification and Finding Suitable Classification Method with It. 444-448.