MINI PROPOSAL TUGAS AKHIR

Program Studi Pendidikan Teknik Informatika dan Komputer
Fakultas Keguruan dan Ilmu Pendidikan - Universitas Sebelas Maret Surakarta

---

Identitas Mahasiswa

Nama Mahasiswa	:	Arruya Hiyarani
NIM	:	K3516061
Nomor Handphone / WA	:	082244543983
IPK Terakhir	:	3.58
Jumlah SKS Kumulatif	:	146

Deskripsi Rencana Tugas Akhir

Judul Rencana Tugas Akhir

PERBANDINGAN KINERJA METODE SUPPORT VECTOR MACHINE DAN NAIVE BAYES UNTUK KLASIFIKASI DOKUMEN SKRIPSI

Jenis Penelitian ☐ Kualitatif ☑ Kuantitatif ☐ PTK ☐ Research and Development

☐ Lain-Lain (Sebutkan: .....)

Latar Belakang

Revolusi teknologi telah memfasilitasi jutaan orang melalui peningkatan penggunaan beragam perangkat digital dan terutama sensor jarak jauh yang menghasilkan aliran data digital terus menerus dengan jumlah data yang besar, yang kemudian disebut sebagai big data (Che, Safran, & Peng, 2013). Big data saat ini tersebar secara global dan diterima secara luas dalam bentuk data terstruktur maupun data tidak terstruktur. Sejumlah besar data tersebut dikumpulkan dan disimpan oleh organisasi, dengan harapan berguna di masa depan. McKinsey Global Institute memperkirakan bahwa volume data tumbuh 40% per tahun, dan akan tumbuh 44x antara 2009 dan 2020 (Singh, 2014). Untuk industri modern, data yang dihasilkan oleh mesin dan perangkat, solusi berbasis cloud, manajemen bisnis, dll. telah mencapai volume total lebih dari 1000 Exabyte setiap tahun dan diperkirakan akan meningkat 20 kali lipat dalam sepuluh tahun ke depan (R & K R, 2017).

Big data sering kali datang dalam bentuk aliran dari berbagai jenis. Che dkk. (2013) berpendapat waktu adalah dimensi integral dari aliran data yang menunjukkan bahwa data harus diproses secara tepat waktu atau mendekati real time. Namun data-data tersebut biasanya terstruktur secara longgar dan seringkali tidak lengkap, dikarenakan pengguna tidak dapat mengakses data esensial. Ini menimbulkan tantangan dalam mengelola data dan mengekstrak informasi yang tepat untuk mendukung keputusan. Dengan jumlah data yang besar serta heterogen membutuhkan teknologi dan alat untuk menemukan, mengubah, menganalisis, dan memvisualisasikan data agar dapat digunakan untuk pengambilan keputusan yang efektif.

Pertumbuhan komputer dan kemajuan teknologi komputasi awan yang sangat cepat mendorong penggunaan data elektronik, salah satu yang paling banyak digunakan adalah dokumen teks elektronik dalam pertukaran informasi internet (Choo et al., 2019). Perkembangan teknologi informasi dan internet berdampak pada meningkatnya jumlah data elektronik yang mengandung dokumen teks (Dogan & Uysal, 2019). Hal ini menyebabkan penyimpanan, pengelolaan serta pengambilan dokumen teks ini menjadi sangat penting. Ketersediaan dan aksesibilitas dalam hal ini untuk memperoleh informasi telah mengarah pada pengembangan sistem klasifikasi teks dan dokumen otomatis yang mampu mengatur dan mengklasifikasikan dokumen secara otomatis.

Klasifikasi merupakan teknik machine learning untuk menetapkan label-label pada data yang tidak terlihat berdasarkan pada model yang dirancang menggunakan algoritma dan data berlabel (Mocherla, Danehy, & Impey, 2017). Klasifikasi teks adalah proses pengelompokkan dokumen ke dalam seperangkat kategori yang telah ditentukan berdasarkan konten mereka (Gharib, Habib, & Fayed, 2009). Dogan dan Uysal (2019) berpendapat, selain untuk mendapatkan informasi yang bermanfaat, klasifikasi teks adalah sebuah metode yang penting untuk mengelola dan mengatur data. Klasifikasi teks sangat dibutuhkan dalam melakukan kegiatan pengarsipan terutama melibatkan dokumen dengan jumlah yang cukup besar. Dalam hal ini seperti kegiatan pengarsipan dokumen tugas akhir atau skripsi di perguruan tinggi, seringkali terjadi kesulitan saat menentukan klasifikasi tema dari judul skripsi yang telah diajukan oleh mahasiswa. Umumnya klasifikasi jenis tema dari skripsi tersebut hanya diperkirakan berdasarkan isi konten yang diteliti oleh mahasiswa sehingga kesesuaian antara judul dan tema sering diabaikan.

Poin utama dalam klasifikasi teks yaitu membangun struktur data yang dapat mewakili dokumen, dan membangun classifier yang dapat digunakan untuk predikat label kelas

dokumen dengan akurasi tinggi (Khan, Baharudin, & Lee, 2010). Classifier menggunakan algoritma machine learning untuk mempelajari model prediksi kelas berdasarkan dokumen berlabel (Jadon & Sharma, 2017). Pengklasifikasian dokumen skripsi perlu dilakukan dalam upaya memisahkan atau mengelompokkan berdasarkan jenis atau kategori tertentu sehingga data yang terkumpul dapat memberikan informasi yang tepat. Banyaknya dokumen yang akan diproses tidak mungkin dilakukan secara manual dikarenakan memerlukan banyak waktu dan tenaga.

Gharib, dkk. (2009) menyatakan dokumen dalam sistem klasifikasi teks harus melewati serangkaian langkah: konversi dokumen yang mengubah berbagai jenis dokumen menjadi teks biasa, menggunakan stop word removal untuk menghapus kata-kata yang tidak penting, stemming untuk mengelompokkan kata dengan struktur dasar yang sama, feature selection/extraction, super vector construction, pembobotan fitur (feature weighting), kontruksi classifier, klasifikasi, evaluasi classifier. Menurut Mahender (2012), dokumen dapat diklasifikasikan dengan tiga cara, yaitu tanpa pengawasan (unsupervised), terpandu (supervised) dan semi terpandu (semi-supervised). Algoritma klasifikasi teks saat ini telah banyak berkembang, antara lain Support Vector Machines (SVM), Naïve Bayesian (NB), pohon keputusan (Decision Tree), K-Nearest Neighbor (KNN), dan lainnya (Somantri, Wiyono, & Dairoh, 2016).

Chakrabarti, dkk. menyatakan klasifikasi teks menggunakan Naïve Bayes telah banyak digunakan di antara algoritma yang berkembang saat ini dikarenakan kesederhanaannya dalam tahap training dan klasifikasi (Ting, Ip, & Tsang, 2011). Hal tersebut juga didukung oleh pernyataan Xu yang menyebutkan Naïve Bayes cepat dan mudah diimplementasikan, sehingga menjadi dasar dalam klasifikasi teks (Aliwy & Ameer, 2017). Dalam penelitiannya, Ting, dkk. (2011) menjelaskan algoritma Naïve Bayes memungkinkan setiap atribut untuk berkontribusi terhadap keputusan akhir secara setara dan independen dari atribut lainnya, yang mana lebih efisien secara komputasi jika dibandingkan dengan algoritma classifier teks lainnya. Naïve Bayes adalah algoritma dengan kesederhanaannya yang berakar pada asumsi bahwa fitur-fitur dari data yang mendasarinya tidak tergantung satu sama lain. Terlepas dari kesederhanaannya dan asumsi gagal dalam banyak kasus, hasilnya mengejutkan secara akurat. Dalam konteks klasifikasi teks, ini berarti bahwa semua kata dalam dokumen tidak bergantung satu sama lain.

Sebaliknya, Rennie, dkk. (Aliwy & Ameer, 2017) berpendapat bahwa metode menggunakan Naïve Bayes cukup efektif untuk mengklasifikasikan teks dalam banyak domain, meskipun kurang akurat dibandingkan metode klasifikasi lainnya seperti Support Vector Machine. Beberapa penulis berpendapat, penggunaan Support Vector Machine (SVM) sebagai classifier dianggap bekerja dengan baik untuk menangani ruang input dengan dimensi tinggi dan memiliki akurasi yang baik dalam klasifikasi teks (Utomo & Sibaroni, 2019). Penelitian yang dilakukan oleh Gharib, dkk. (2009) membandingkan algoritma K-Nearest Neighbor, Naive Bayes, Rocchio, dan Support Vector Machine menunjukkan hasil bahwa classifier Rocchio memberikan hasil yang lebih baik ketika ukuran set fitur kecil sedangkan SVM mengungguli classifier lain ketika ukuran set fitur cukup besar. Tingkat klasifikasi melebihi 90% ketika menggunakan lebih dari 4000 fitur. Dalam algoritma Support Vector Machine, dokumen teks direpresentasikan sebagai vektor dan dimensi adalah jumlah dari kata kunci yang unik.

Kedua algoritma yaitu Naive Bayes dan Support Vector Machine masing-masing menunjukkan keunggulannya dalam klasifikasi dokumen teks. Secara umum, kinerja suatu algoritma klasifikasi sangat dipengaruhi oleh kualitas sumber data serta algoritma yang berbeda bekerja secara berbeda tergantung pada pengumpulan data. Dari penelitian yang telah disebutkan sebelumnya, penulis menerapkan text mining dengan menggunakan dua algortima untuk

penelitian yang berjudul "PERBANDINGAN KINERJA METODE SUPPORT VECTOR MACHINE DAN NAIVE BAYES UNTUK KLASIFIKASI DOKUMEN SKRIPSI".

Penelitian ini ditujukan untuk mengklasifikasikan dokumen skripsi sesuai dengan kategorinya dengan menggunakan algoritma klasifikasi Support Vector Machine dan Naive Bayes dengan tujuan membandingkan kinerja algoritma klasifikasi tersebut dan optimalisasinya untuk mendapatkan algoritma dengan kinerja dan tingkat akurasi yang baik.

Rumusan Masalah

Bagaimana perbandingan kinerja algoritma Naïve Bayes dan Support Vector Machine dalam melakukan klasifikasi dokumen skripsi Program Studi Pendidikan Teknik Informatika dan Komputer?

Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini adalah untuk mengetahui perbandingan kinerja algoritma Naïve Bayes dan Support Vector Machine dalam melakukan klasifikasi dokumen skripsi Program Studi Pendidikan Teknik Informatika dan Komputer.