KLASIFIKASI PROPOSAL TUGAS AKHIR BERDASARKAN BIDANG KEAHLIAN DAN PEMINATAN DOSEN PEMBIMBING MENGGUNAKAN SUPPORT VECTOR MACHINE PADA JURUSAN TEKNIK INFORMATIKA DI STMIK INDONESIA BANJARMASIN

Amrul Hadiyanoor
Jl Pangeran Hidayatullah, Banua Anyar, Banjarmasin
Email : amrulhy@gmail.com

Abstract

In STMIK Indonesia Banjarmasin each student who works on their final projects are given three lecturers determined by the outcome of the meeting of the concerned departments. Various obstacles may occur including the meeting time interfering with the course time, the delay in submitting students' final project proposal draft to be discussed in the meeting and the results of the meeting not being suitable because the material covered in the final project is not mastered by the supervisor who has been specified. For that purpose, this research proposes to make a system that enables easy and appropriate classification of final assignment proposal draft of students to lecturers who are experienced and have the appropriate specialization and expertise.

Support Vector Machine (SVM) is one of the best ways to solve the problem of classification. Since SVM can only classify into two classes, multiclass SVM method is used. From the results of classification using SVM, sorting was done using Cosine Similarity for the final result.

In this research, 184 students' final project proposal drafts were used for the test. Seventy or as much as 129 documents were used for the training process, and the remainder, namely 30% or a total of 55 documents were used for the testing process. From the test results using SVM classification and sorting by Cosine Similarity, the success achieved was 85%.

Keyword : SVM, Cosine Similarity, Mullticlass Classifications, Matlab.

Latar Belakang

Tugas akhir adalah salah satu syarat yang wajib dikerjakan oleh setiap mahasiswa untuk menyelesaikan pendidikan tinggi mereka. Dalam menyelesaikan tugas akhir seorang mahasiswa memerlukan pembimbing, yaitu satu atau beberapa dosen berpengalaman yang dapat membimbing mereka mengerjakan penelitian yang akan dilakukan. Setiap program studi memiliki berbagai topik berbeda yang dapat dibahas menjadi sebuah penelitian untuk dijadikan sebagai tugas akhir. Penguasaan dari setiap materi oleh para calon pembimbing yang akan membimbing memiliki perbedaan, oleh karena itu dipilih pembimbing yang menguasai dan memiliki peminatan bidang keahlian sesuai dengan topik penelitian mahasiswa yang akan dibimbing.

Pada Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Indonesia Banjarmasin mekanisme pemilihan dosen pembimbing dimulai setelah semua mahasiswa yang akan mengerjakan tugas akhir mengumpulkan *draft* (konsep) proposal. Para dosen struktural pada jurusan yang bersangkutan yang berwenang akan melakukan rapat untuk membagi dosen pembimbing sesuai bidang keahlian mereka berdasarkan topik dan materi

penelitian yang diajukan mahasiswa melalui draft proposal tugas akhir yang mereka kumpulkan.

Pembagian dosen pembimbing secara *manual* memiliki kendala seperti waktu pelaksanaan rapat yang biasa dilakukan saat padatnya jadwal perkuliahan. Masalah akan bertambah ketika banyaknya mahasiswa yang telah melakukan pendaftaran tugas akhir terlambat mengumpulkan draft proposal dan tidak tepatnya pembagian dosen pembimbing karena tidak sesuai dengan peminatan dan keahlian dosen tersebut. Dari beberapa masalah tersebut proses pembagian akan lebih mudah dan cepat jika dilakukan secara otomatis dengan memanfaatkan pengetahuan seperti menggunakan teknik klasifikasi dokumen.

Tujuan

Tujuan yang diharapkan dari penelitian pada penelitian ini adalah seperti yang dituliskan sebagai berikut:

1. Memudahkan pembagian tugas akhir pada pembimbing yang tepat dan berpengalaman.
2. Memudahkan dan mempercepat mahasiswa dan mahasiswi mengetahui dosen pembimbing dari tugas akhir yang akan mereka kerjakan.
3. Menghemat waktu yang akan digunakan untuk melakukan pembagian dosen pembimbing.

Ruang Lingkup

Dalam ruang lingkup ini dijelaskan mengenai obyek penelitian, data yang diteliti, permasalahan dan batasan masalah yang dibahas pada penelitian ini, algoritma yang digunakan dan blok diagram proses penelitian.

Obyek Penelitian

Obyek penelitian pada penelitian ini adalah draft proposal Tugas Akhir dari mahasiswa dan mahasiswi STMIK Indonesia Banjarmasin jurusan Teknik Informatika.

Data Input dan Output

Data input yang digunakan adalah berupa *draft* (konsep) proposal tugas akhir yang di *upload* oleh mahasiswa/i berbentuk file dokumen *word* (*.doc / *.docx) dan diubah menjadi file *text* (teks). *Data output*-nya berupa nama para calon dosen pembimbing yang didapat setelah proses lebih lanjut dari hasil klasifikasi. Data yang digunakan sejumlah 184 draft proposal untuk proses pelatihan digunakan 70% yaitu sejumlah 129 draft proposal dan untuk proses klasifikasi digunakan 30% yaitu 55 draft proposal.

Rumusan Masalah

Masalah yang diangkat pada penelitian ini adalah bagaimana cara mengklasifikasikan proposal tugas akhir berdasarkan bidang keahlian. Selanjutnya menentukan 3 dosen pembimbing untuk proposal tugas akhir tersebut, berdasarkan peminatan dosen pembimbing pada bidang kehlian hasil klasifikasi.

Batasan Masalah

Agar penelitian sesuai dengan yang diharapkan, berikut adalah batasan masalah dari penelitian ini yang meliputi:

1. Draft proposal tugas akhir yang dimasukkan hanya dari mahasiswa jurusan Teknik Informatika pada STMIK Indonesia Banjarmasin.
2. Menggunakan Support Vector Machine untuk mengklasifikasi draft proposal.
3. Hasil dari klasifikasi diurutkan menggunakan Cosine Similarity, 3 dosen teratas menjadi dosen pembimbing, dan yang tertinggi menjadi koordinator pembimbing.

Support Vector Machine

Support Vector Machine merupakan algoritma yang awalnya diciptakan oleh Vladimir N. Vapnik dan Alexey Ya. Chervonenkis pada tahun 1963, untuk standar terbaru diusulkan pada tahun 1993 dan dipublikasikan pada tahun 1995 oleh Corinna Cortes dan Vladimir N. Vapnik, algoritma ini mampu menganalisis data dan mengenali pola yang dapat digunakan untuk klasifikasi dan analisa regresi. Ide dasarnya adalah mengklasifikasi data atau pola menjadi dua bagian.

Stemming Indonesian

Pada penelitian yang dibahas oleh Jelita Asian bersama Hugh E. Williams dan S.M.M. Tahaghoghi ini diuraikan cara meningkatkan algoritma stemming yang telah diajukan oleh Bobby Nazief dan Mirna Adriani pada tahun 1996.

Pada umumnya kata pada bahasa indonesia tersusun dari satu *root word* (kata dasar), dan beberapa awalan (*prefix*/prefiks) yaitu *Derivation Prefixes* dan akhiran (*suffix*/sufiks) yaitu *Derivation Suffixes* dan *Inflection Suffixes*.

Data Mining

Data mining didefinisikan sebagai proses komputasi untuk menganalisis data dalam jumlah besar dengan mengekstrak pola dan informasi yang berguna (Gullo, 2015). Dalam beberapa dekade terakhir, data mining telah banyak mendapat sebutan lain seperti *knowledge discovery, business intelligence, predictive modeling, predictive analytics*, dan beberapa lainnya. Tetapi, tidak sedikit orang yang mendefinisikan data mining sebagai sinonim dari istilah populer lainnya yaitu *knowledge discovery from data* (KDD) dan yang lain melihat data mining hanya sebagai salah satu tahapan dari knowledge discovery.

Pada proses knowledge discovery seperti ditunjukkan pada Gambar 2.4, terdapat beberapa tahapan proses yang dilakukan yaitu:

- *Cleaning data*, yaitu proses untuk mengeliminasi *noise* (pengganggu) dan data yang tidak konsisten).
- *Integrasi data*, yaitu proses penggabungan data jika data diperoleh dari berbagai sumber.
- Seleksi data, yaitu proses pemilihan data yang benar-benar berguna untuk dianalisis.
- Transformasi data, yaitu proses transformasi data menjadi bentuk yang sesuai untuk dilakukan proses data mining.
- Data mining, yaitu proses dimana metote-metode khusus diaplikasikan untk mengekstrak inforrmasi dan pola data.
- *Pattern evaluation*, yaitu proses untuk mengidentifikasi pola-pola dan informasi menarik yang didapatkan dari data.

Pentingnya data mining saat ini terutama didorong oleh banyaknya data yang dikumpulkan dan disimpan dengan berbagai aplikasi terkemuka terkini, seperti data web, data *e-commerce*, data pembelian, transaksi bank, dan sebagainya. Data yang dihasilkan oleh aplikasi-aplikasi tersebut umumnya merupakan jenis *Big Data* dimana data tersebut sulit diolah atau dimengerti secara sederhana. *Big Data* merupakan data yang mempunyai tiga karakteristik yaitu jumlah (*volume*) dan variasi (*variety*) besar, serta bergerak cepat (*velocity*), sehingga melampaui kapasitas pengolahan database konvensional. Hingga saat ini, data mining telah banyak diakui sebagai suatu alat analisis data serbaguna yang bisa diaplikasikan untuk menganalisis big data dalam berbagai bidang, tidak hanya dalam bidang teknologi informasi tetapi juga dalam dunia pengobatan klinis, sosiologi, fisika, dan banyak lainnya.

Penggunaan data mining dibedakan menjadi dua jenis fungsi yaitu prediktif dan deskriptif. Penggalian prediktif mengacu pada pembangunan model yang berguna untuk memprediksi perilaku atau nilai-nilai di masa depan. Tugas deskriptif meliputi klasifikasi dan prediksi, tugas yang dilakukan seperti membangun beberapa model (atau fungsi) yang menggambarkan kelas atau konsep data oleh satu set objek data yang label kelasnya diketahui (training set), sehingga dapat memprediksi kelas yang labelnya tidak diketahui; deteksi penyimpangan, yaitu berurusan dengan penyimpangan data, yang didefinisikan sebagai perbedaan antara nilai yang terukur dan nilai referensi; analisis evolusi, yaitu, mendeteksi dan menggambarkan pola yang teratur dalam data yang perilakunya berubah dari waktu ke waktu. Sedangkan tujuan penggalian deskriptif yaitu membangun model untuk mendeskripsikan data menjadi bentuk yang mudah dimengerti, efektif, dan efisien. Contoh dari tugas deskriptif diantaranya karakterisasi data, yang tujuan utamanya adalah untuk meringkas karakteristik umum atau fitur dari kelas target data; association rule, yaitu, menemukan aturan yang menunjukkan kondisi atribut-nilai yang sering muncul bersama-sama dalam himpunan data; dan clustering, yang bertujuan untuk membentuk kelompok yang memiliki kohesif tinggi dan terpisahkan dengan baik dari satu set objek data.

Text Mining

Text mining atau text analytics adalah istilah yang mendeskripsikan sebuah teknologi yang mampu menganalisis data teks semi-terstruktur maupun tidak terstruktur, hal inilah yang membedakannya dengan data mining dimana data mining mengolah data yang sifatnya terstruktur. Pada dasarnya, *text mining* merupakan bidang interdisiplin yang mengacu pada perolehan informasi (*information retrieval*), *data mining*, pembelajaran mesin (*machine learning*), statistik, dan komputasi linguistik. Secara umum konsep pekerjaan *text mining* mirip dengan data mining, yaitu penggalian prediktif dan penggalian deskriptif. *Text mining* mengekstrak indeks numerik yang bermakna dari teks dan kemudian informasi yang terkandung dalam teks akan diakses dengan menggunakan berbagai algoritma *data mining* (statistik dan *machine learning*). Beberapa tahun terakhir, penggunaan dan penelitian mengenai text mining telah banyak mendapat perhatian dan aktif dilakukan seiring dengan semakin banyaknya data teks yang diperoleh dari berbagai jaringan sosial, web, dan aplikasi lainnya. Sebagian besar informasi teks yang disimpan tersebut seperti misalnya artikel berita, makalah, buku, perpustakaan digital, pesan email, blog, dan halaman web.

Text mining dapat menganalisis dokumen, mengelompokkan dokumen berdasarkan kata-kata yang terkandung di dalamnya, serta menentukan kesamaan di antara dokumen untuk mengetahui bagaimana mereka berhubungan dengan variabel lainnya. Aplikasi yang paling umum dilakukan text mining saat ini misalnya penyaringan spam, analisis sentimen, mengukur preferensi pelanggan, meringkas dokumen, pengelompokan topik penelitian, dan banyak lainnya. Menurut Miner et al (2012), pekerjaan text mining dikelompokkan menjadi 7 daerah praktek yaitu:

- Pencarian dan perolehan informasi (*search and information retrieval*), yaitu penyimpanan dan penggalian dokumen teks misalnya dalam mesin pencarian (*search engine*) dan pencarian kata kunci (*keywords*).
- Pengelompokan dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan

menggunakan metode klaster (*clustering*) data mining.

- Klasifikasi dokumen, yaitu pengelompokan dan pengkategorian kata, istilah, paragraf, atau dokumen dengan menggunkan metode klasifikasi (*classification*) data mining berdasarkan model terlatih yang sudah memiliki label.
- *Web mining*, yaitu penggalian informasi dari internet dengan skala fokus yang spesifik.
- Ekstraksi informasi (*information extraction*), yaitu mengidentifikasi dan mengekstraksi informasi dari data yang sifatnya semi-terstruktur atau tidak terstruktur dan mengubahnya menjadi data yang terstruktur.
- *Natural language processing* (NLP), yaitu pembuatan program yang memiliki kemampuan untuk memahami bahasa manusia.
- Ekstraksi konsep, yaitu pengelompokan kata atau frase ke dalam kelompok yang mirip secara semantik.

Untuk memperoleh tujuan akhir dari text mining, diperlukan beberapa tahapan proses yang harus dilakukan seperti ditunjukkan pada Gambar 2.6. Data terpilih yang akan dianalisis pertama akan melewati tahap Pra-proses dan representasi teks, hingga akhirnya dapat dilakukan *knowledge discovery*

Pra-proses (Pre-processing Task)

Data yang diinput perlu melewati fase pra-proses terlebih dahulu agar dapat dimengerti oleh sistem pengolahan text mining dengan baik. Fase pra-proses merupakan fase yang penting untuk menentukan kualitas proses selanjutnya (proses klasifikasi dan pengelompokan). Tujuan utama fase pra-proses adalah untuk mendapatkan bentuk data siap olah untuk diproses oleh data mining dari data awal yang berupa data tekstual. Fitur-fitur fase pra-proses terdiri dari beberapa tahap, dimulai dari pemilihan dokumen yang digunakan (dokumen yang mengandung ancaman, caci maki, SARA, dan pornografi dihilangkan). Berikutnya adalah *Tokenization*, yaitu proses pemisahan teks menjadi potongan kalimat dan kata yang disebut token.

Penyusunan Vektor (Representation)

Proses operasi algoritma belajar (*learning algorithms*) tidak bisa langsung memproses dokumen teks dalam bentuk aslinya. Oleh karena itu, setelah tahap *pre-processing*, dokumen diubah menjadi representasi yang lebih mudah dikelola. Biasanya, dokumen akan diwakili oleh vektor. Model vektor dibangun dari dokumen dengan mengubah token-token dalam dokumen menjadi vektor numerik yang akan dioperasikan berdasarkan operasi aljabar linear. Dalam rangka membangun model vektor, perlu dilakukan proses pembobotan. Skema pembobotan yang paling banyak digunakan adalah skema *term frequency-inverse document frequency* (TF-IDF). *Term frequency* (TF) didefinisikan sebagai jumlah kemunculan suatu kata/istilah dalam suatu dokumen. Misalnya TF pada dokumen pertama untuk kata/istilah "jalan" adalah 2, karena kata/istilah tersebut muncul 2 kali dalam dokumen pertama. Pada asusmsi pembobotan dibalik TF-IDF, kata-kata dengan nilai TF yang tinggi akan mendapat bobot yang tinggi kecuali jika jumlah dokumen yang mengandung kata tersebut juga tinggi (*inverse document frequency* (IDF)).

Support Vector Machine

Support Vector Machines (SVMs) adalah seperangkat metode pembelajaran terbimbing yang menganalisis data dan mengenali pola, digunakan untuk klasifikasi dan analisis regresi. Algoritma SVM asli diciptakan oleh Vladimir Vapnik dan turunan standar saat ini (margin lunak) diusulkan oleh Corinna Cortes dan Vapnik

Vladimir. SVM standar mengambil himpunan data input, dan memprediksi, untuk setiap masukan yang diberikan, kemungkinan masukan adalah anggota dari salah satu kelas dari dua kelas yang ada, yang membuat sebuah SVM sebagai penggolong non-probabilistik linier biner. Karena sebuah SVM adalah sebuah pengklasifikasi, kemudian diberi suatu himpunan pelatihan, masing-masing ditandai sebagai milik salah satu dari dua kategori, suatu algoritma pelatihan SVM membangun sebuah model yang memprediksi apakah data yang baru jatuh ke dalam suatu kategori atau yang lain.

SVM pada awalnya digunakan untuk klasifikasi data numerik, tetapi ternyata SVM juga sangat efektif dan cepat untuk menyelesaikan masalah- masalah data teks. Data teks cocok untuk dilakukan klasifikasi dengan algoritma SVM karena sifat dasar teks yang cenderung mempunyai dimensi yang tinggi, dimana terdapat beberapa fitur yang tidak relevan, tetapi akan cenderung berkorelasi satu sama lain dan umumnya akan disusun dalam kategori yang terpisah secara linear.

Secara intuitif, model SVM merupakan representasi dari data sebagai titik dalam ruang, dipetakan sehingga kategori contoh terpisah dibagi oleh celah jelas yang selebar mungkin. Data baru kemudian dipetakan ke dalam ruang yang sama dan diperkirakan termasuk kategori berdasarkan sisi mana dari celah data tersebut berada.

Lebih formal, Support Vector Machine membangun *hyperplane* atau himpunan hyperplane dalam ruang dimensi tinggi atau tak terbatas, yang dapat digunakan untuk klasifikasi, regresi atau tugas-tugas lainnya. Secara intuitif, suatu pemisahan yang baik dicapai oleh hyperplane yang memiliki jarak terbesar ke titik data training terdekat dari setiap kelas (margin fungsional disebut), karena pada umumnya semakin besar margin semakin rendah *error* generalisasi dari pemilah.

Ketika masalah asal mungkin dinyatakan dalam dimensi ruang terbatas, sering terjadi bahwa dalam ruang, himpunan tidak dipisahkan secara linear. Untuk alasan ini diusulkan bahwa ruang dimensi terbatas dipetakan ke dalam sebuah ruang dimensi yang jauh lebih tinggi yang mungkin membuat pemisahan lebih mudah dalam ruang itu. Skema SVM menggunakan pemetaan ke dalam ruang yang lebih besar sehingga cross product dapat dihitung dengan mudah dalam hal variabel dalam ruang asal membuat beban komputasi yang wajar. *Cross product* di ruang yang lebih besar didefinisikan dalam hal fungsi kernel K (x, y) yang dapat dipilih sesuai dengan masalah. Sekumpulan hyperplane dalam ruang besar yang didefinisikan sebagai himpunan titik-titik yang *cross product* dengan vektor dalam ruang yang konstan. Vektor mendefinisikan hyperplanes dapat dipilih untuk menjadi kombinasi linear dengan parameter αi dari gambar vektor fitur yang terjadi pada database. Dengan pilihan ini sebuah *hyperplane* di titik x di ruang fitur yang dipetakan ke *hyperplane* ini ditentukan oleh relasi:

$$\sum_i \alpha i K\,(xi,x) = constant$$

Perhatikan bahwa jika K (x, y) menjadi kecil ketika y tumbuh lebih lanjut dari x, setiap elemen dalam pengukuran penjumlahan dari tingkat kedekatan titik uji x ke titik xi pada database yang sesuai. Dengan cara ini jumlah kernel di atas dapat digunakan untuk mengukur kedekatan relatif masing-masing titik uji dengan titik data yang berasal dalam satu atau yang lain dari himpunan yang akan dikelompokkan. Perhatikan fakta bahwa himpunan titik x dipetakan ke hyperplane yang manapun , dapat cukup rumit sebagai akibat mengijinkan pemisahan yang lebih kompleks antara himpunan yang jauh dari convex di ruang asli.

Karena SVM hanya bisa mengklasifikasi menjadi dua bagian maka digunakan cara kelas jamak (*multiclass*). Terdapat dua macam cara Multiclass SVM yang terkenal yaitu *One Against All/One Versus All/One Against Rest* dan *One Against One/One Versus One.*

Metode One Versus All menggunakan cara 1 kelas melawan sisa kelas yang ada. Jadi jika jumlah kelas adalah k maka dibangun model sebanyak k. Setiap model klasifikasi ke-i dilatih dengan menggunakan keseluruhan data, untuk mencari solusi permasalahan.

Sedangkan pada metode One Versus One dibangun model berdasarkan jumlah kelas (k) sebanyak $\frac{k(k-1)}{2}$. Setiap model klasifikasi dilatih pada data dari dua kelas.

Cosine Similarity

Cosine Similarity adalah ukuran kemiripan antar dua vektor, yang mana tidak ada vektor yang kosong. Berikut ini adalah persamaan menghitung Cosine Similarity:

$$\cos(\theta_{ij}) = \frac{\sum_k A_{ik} B_{jk}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k B_{jk}^2}}$$

Implementasi

Berdasarkan batasan masalah, tinjauan pustaka dan teori penunjang maka pada bab ini membahas mengenai implementasi dan penerapan algoritma dan membahas segmen program mengenai program klasifikasi dokumen proposal tugas akhir.

Analisa Sistem

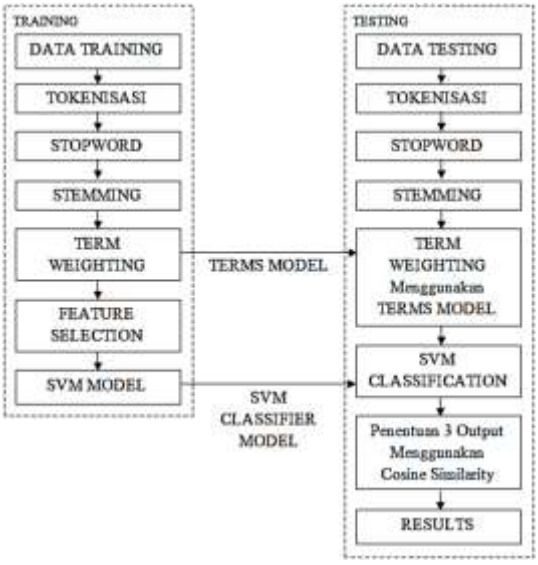
Dari sistem yang masih berjalan, setiap awal semester pada STMIK Indonesia Banjarmasin, mengadakan pendaftaran ujian tugas akhir atau skripsi, syarat mendaftar mahasiswa dan mahasiswi diwajibkan menyelesaikan administrasi dan menyerahkan draft fisik proposal kepada bagian jurusan.

Pada bagian jurusan akan mengadakan rapat dan memilihkan 3 (tiga) dosen untuk membimbing setiap mahasiswa atau mahasiswi.

Pada penelitian ini sistem yang berjalan akan diubah, pada sistem yang dirancang ini mahasiswa atau mahasiswi diminta menyerahkan draft proposal berbentuk file dokumen Word berupa *.doc atau *.docx. Mahasiswa atau mahasiswi akan mendapatkan jawaban 3 nama dosen pembimbing setelah menunggu 1 hari.

Dokumen yang telah dikumpulkan diubah menjadi file teks untuk dapat di proses lebih lanjut pada program, file teks tersebut akan melalui beberapa pra-proses teks agar dimengerti oleh sistem pengolahan teks.

Blok Diagram



Gambar 1 Blok Diagram

Data Training

Data training yang digunakan adalah berkas tugas akhir mahasiswa jurusan Teknik Informatika dari tahun 2012 sampai tahun 2015. Setelah diseleksi didapatkan sejumlah 184 dokumen yang dapat digunakan. Dari hasil analisa 184 dokumen dapat diklasifikasi menjadi 8 kelas, dan terdapat 13 dosen yang membimbing. Untuk data training digunakan 70% dari

184 dokumen, yaitu sejumlah 129 dokumen.

Preprocessing

Proses preprocessing dibagi menjadi beberapa tahap yaitu:

1. Tokenisasi
2. Stopword Filtering
3. Stemming

Pembobotan

Diberikan pembobotan dengan menggunakan TF-IDF.

$$tfidf(w) = tf \times log\frac{N}{df(w)}$$

SVM

Pada penelitian ini pelatihan menggunakan library dari Matlab, untuk bagian pelatihan digunakan fungsi svmtrain, fungsi ini mengggunakan metode optimisasi dengan persamaan sebagai berikut:

$$c = \sum_i \alpha_i k(s_i x) + b$$

Dimana k adalah fungsi kernel, s_i adalah support vector, α_i adalah bobot, b adalah bias dan x adalah vektor. Nilai c digunakan untuk mengklasifikasi x jika lebih dari atau sama dengan 0 maka termasuk kelas pertama, dan jika tidak maka masuk kelas kedua.

Perankingan dengan Cosine Similarity

Pada penelitian ini di setiap kelas target terdapat beberapa dosen, oleh karena itu diberikan perangkingan pada para dosen tersebut berdasarkan kemiripan tugas akhir yang pernah dibimbing dan draft proposal testing. Berikut persamaan cosine similarity yang digunakan untuk melakukan proses perangkingan.

$$\cos(\theta_{ij}) = \frac{\sum_k A_{ik} B_{jk}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k B_{jk}^2}}$$

Pada bagian pembobotan dihitung masing-masing term pada sebuah dokumen. Dari beberapa dokumen yang memiliki pembimbing yang sama dijumlahkan menjadi vector baru berlabelkan bidang keahlian dan nama pembimbingnya. Vektor-vektor tersebut digunakan untuk menghitung similarity pada bagian ini(perangkingan).

Hasil

Dari uji coba sebanyak 184 draft proposal tugas akhir mahasiswa yang digunakan, sebesar sebanyak 129 dokumen digunakan untuk proses training atau sebesar 70% dari jumlah draft proposal yang digunakan, dan sisanya yaitu sebanyak 55 dokumen digunakan untuk proses testing atau sekitar 30% dari jumlah draft proposal. Dari hasil uji coba klasifikasi menggunakan SVM dan perangkingan menggunakan Cosine Similarity keberhasilan yang dicapai adalah 85%.

Kesimpulan

Berikut ini adalah kesimpulan yang diambil dari hasil penelitian dan implementasi program klasifikasi:

- Uji coba menggunakaan draft proposal tugas akhir mahasiswa sebanyak 184. Sebesar 70% atau sebanyak 129 draft proposal digunakan untuk proses training, dan sisanya yaitu sebesar 30% atau sebanyak 55 dokumen digunakan untuk proses testing. Dari hasil uji coba klasifikasi menggunakan SVM.
- Untuk menghitung ranking dihitung nilai cosinus antara fitur data testing dan fitur dosen pembimbing. Fitur dosen yang digunakan berasal dari rata-rata nilai TFIDF dari proposal yang pernah dibimbing dosen tersebut pada kelas hasil klasifikasi SVM.
- Dilakukan perbandingan menggunakan fitur dosen yang berasal dari rata-rata seluruh fitur TFIDF dokumen yang dibimbing oleh dosen terkait. Hasil uji coba

menunjukkan fitur dosen yang berasal dari fitur TFIDF pada kelas hasil klasifikasi SVM lebih baik.

Daftar Pustaka:

1. Asian, Jelita., Williams, Hugh E., Tahaghoghi, S.M.M., 2007. "*Stemming Indonesian"*, School of Computer Science and Information Technology RMIT University, GPO Box 2476V, Melbourne 3001, Australia.

2. Dinu, Liviu P., Ionescu, Radu-Tudor., 2012. Department of Computer Science, University of Bucharest, *"A Rank-based Approach of Cosine Similarity with Applications in Automatic Classification"*. 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing. PP 260-264.

3. Joachims, Thorsten., 1998. *"Text Categorization with Support Vector Machine: Learning with Many Relevant Features"*. Proceedings of the 10th European Conference on Machine Learning, Chemnitz. PP 137-142.

4. Liang, Jiu-Zhen., 2004. *"SVM Multi-Classifier and Web Document Classification"*. Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai. PP 1347-1351.

5. Wang, Zi-Qiang., Sun, Xia., Zhang ,De-Xian., Li, Xin., 2006. *"An Optimal SVM-based Text Classification Algorithm"*. Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian. PP 1378 – 1381.

6. Zhang, Jian-Pei., Li, Zhong-Wei., Yang, Jing., 2005. *"A Parallel SVM Training Algorithm on Large-Scale Classification Problems"*. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou. PP 1637-1641.

.