Implementasi Text Mining untuk Pengelompokkan Dokumen Skripsi Menggunakan Metode k-Means Clustering
Studi Kasus : Library Management System (LMS) Perbanas Institute

Miracle Mu'ammar Veryo
Departemen Teknologi Informasi
Institut Perbanas
Jakarta, Indonesia
1714000023@perbanas.id

Mitsal Shafiq Sulasno
Departemen Teknologi Informasi
Universitas Indonesia
Depok, Indonesia
mitsal.shafiq01@ui.ac.id

Valentinus Paramarta
Departemen Teknologi Informasi
Institut Perbanas
Jakarta, Indonesia
valentinus13@perbanas.id

***Abstrak*— Semakin bertambahnya topik penelitian skripsi menumbuhkan peluang semakin banyaknya mahasiswa yang mengambil topik penelitian yang sama atau hampir serupa. Sehingga diharapkan dengan adanya proses pengelompokan dokumen secara otomatis, maka tidak harus membuka halaman terlalu banyak, karena dokumen hasil pencarian telah dikelompokkan berdasarkan kategori yang dapat menggambarkan isi dari suatu dokumen. Penelitian ini bertujuan untuk mengetahui bagaimana performa metode pembelajaran mesin dapat diterapkan untuk pengelompokkan dokumen skripsi mahasiswa Perbanas Institute pada website digital library. Proses pengelompokkan dokumen skripsi ini dilakukan dengan menggunakan metode k-means clustering dengan mengambil judul dan program studi sebagai informasi penting yang dapat mewakili isi dari dokumen. Selanjutnya dokumen akan dilakukan pre-processing terlebih dahulu menggunakan metode text mining. Untuk tahap pre-processing dibagi menjadi beberapa bagian, yakni case folding, stemming, stopwords remove, bag of words, weighting, feature selection dan normalization. Setelah dokumen melewati tahap pre-processing, maka dokumen dapat dikelompokkan menggunakan metode dari k-means clustering. Hasil dari penelitian ini dapat disimpulkan bahwa pengelompokkan dokoumen skripsi menggunakan metode k-means clustering sangat baik diimplementasikan pada kasus library management systems Perbanas Institute.**

Kata Kunci—machine learning, clastering, skripsi, library management system

I. PENDAHULUAN

Perkembangan teknologi saat ini sudah berkembang sangat pesat, perkembangan teknologi ini menyebabkan ledakan jumlah dokumen elektronik yang tersimpan didalam repository perpustakaan. Sehingga menimbulkan penumpukkan dokumen dan akan menyita *resource* atau memori untuk menyimpan data tersebut. Banyaknya jumlah data dokumen elektronik yang tersimpan didalam website digital library Perbanas, seperti karya ilmiah dari sivitas akademika mulai dari skripsi, laporan penelitian, laporan kerja praktek dan lain sebagainya telah tersedia dalam versi digital. Pada umumnya dokumen-dokumen elektronik tersebut tidak disertai dengan jumlah informasi atau pengetahuan yang disarikan.

Laporan penelitian mahasiswa atau biasa disebut dengan skripsi dapat dikelompokkan berdasarkan program studi, tema penelitian, objek penelitian maupun metode dari penelitian tersebut. Hasil dari pengelompokkan skripsi akan memperlihatkan bagaimana pola kemiripan dan keterkaitan antar penelitian dari waktu ke waktu. Hasil pengelompokkan pun dapat memperlihatkan materi yang banyak diminati oleh mahasiswa dan yang kurang diminati oleh mahasiswa pada waktu tertentu. Seperti yang kita ketauhi, setiap tahun jumlah dokumen skripsi selalu bertambah seiring dengan jumlah mahasiswa yang sedang menempuh semester akhir. Selaih hal tesebut, semakin bertambahnya topik penelitian skripsi ini menumbuhkan peluang semakin banyaknya mahasiswa yang mengambil penelitian dengan tema, objek dan metode penelitian yang mirip atau hampir sama.

Machine Learning (ML) atau pembelajaran mesin merupakan pendekatan dalam *Artificial Intelligence* (AI) yang banyak digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan masalah atau melakukan otomatisasi. Sesuai namanya, ML mencoba menirukan bagaimana proses manusia atau makhluk cerdas belajar dan menggeneralisasikannya (Ahmad, 2017). Terdapat 2 tipe algoritma pada Machine Learning (ML) yaitu *supervised learning* dan *unsupervised learning*. Algoritma *supervised learning* adalah algoritma *Machine Learning* (ML) menggunakan data berlabel, contoh dari *supervised learning* yaitu : *classification* dan *regression*. Sedangkan *unsupervised learning* adalah algoritma *Machine Learning* (ML) yang hanya diberikan sejumlah sampel masukan tanpa label, contoh dari *unsupervised learning* yaitu : *clustering* dan *association* [1].

Ada banyak algoritma yang bisa digunakan untuk melakukan pengelompokkan dokumen, salah satunya adalah algoritma K-Means Clustering. Peneliti menggunakan algoritma ini berdasarkan penelitian-penelitian terdahulu yang telah dilakukan oleh Siti Munifah, Abdul Syukur & Catur Supriyanto (2015); Very Kurnia Bakti & Jatmiko Indriyanto (2017); Rahmah Widya Astuti, Badi'ah & Bagus Satrio (2019); dan Muhammad Rafi Muttaqin & Meriska Defriani (2020) [2-5]. Dimana dalam beberapa penelitian tersebut didapatkan bahwa metode K-Means Clustering dapat cukup baik melaksanakan tugas untuk mengklasifikasikan dokumen berdasarkan tema, topik, bidang keahlian, maupun program studi. Sehingga pada penelitian ini bertujuan untuk mengetahui performa metode pembelajaran mesin dengan menggunakan metode K-Means Clustering untuk pengelompokkan dokumen skripsi mahasiswa Perbanas Institute pada *website digital library*.

Berdasarkan permasalahan yang dipaparkan di atas maka rumusan masalah pada penelitian ini adalah:

RQ: Bagaimana performa metode pembelajaran mesin dengan menggunakan metode K-Means Culstering untuk klasterisasi dokumen skripsi mahasiswa program studi manajemen, sistem informasi, dan teknik informatika Perbanas Institute?

II. LANDASAN TEORI

Pada bagian ini akan dijelaskan terkait penelitian terdahulu yang telah dilakukan. Penelitian yang dijelaskan pada bagian ini adalah penelitian yang memiliki topik terkait penggunaan berbagai algoritma *machine learning* dalam klasterisasi dokumen.

A. Penelitian Terdahulu Terkait Penggunaan Machine Learning Dalam Klasterisasi Dokumen

Penelitian dengan menggunakan algoritma K-Means Clustering yang dilakukan oleh Siti Munifah, Abdul Syukur & Catur Supriyanto (2015) penggunaan algoritma K-Means Clustering pada pengelompokkan arsip universitas dengan menggunakan fungsi similaritas Manhattan Distance. Berdasarkan hasil penelitian bahwa penggunaan algoritma K-Means Clustering pada analisis clustering untuk proses clustering dokumen surat, terlihat adanya kenaikan tingkat akurasi pada Manhattan Distance. Proses clustering menggunakan pembobotan TF-IDF melalui feature selection chi square, akan menghabiskan waktu lebih cepat dan tingkat akurasi lebih besar dibandingkan dengan sebelum penambahan feature selection [2].

Pada penelitian lain yang dilakukan oleh Very Kurnia Bakti & Jatmiko Indriyanto (2017) bahwa pengklasteran dokumen abstrak tugas akhir berbahasa Indonesia dengan menerapkan algoritma K-Means Clustering, menghasilkan nilai jarak antar klaster yang lebih baik. Dengan demikian, metode K-Means clustering sangat baik digunakan dalam penerapan aplikasi sistem temu kembali tugas akhir dikarenakan hasil dari klaster yang dibentuk K-Means Clustering sudah dapat mengelompokan dokumen TA berdasarkan tema tugas akhir masing-masing program studi [3].

Beberapa penelitian mengenai pengelompokkan dokumen dengan menggunakan algoritma K-Means Clustering, yaitu penelitian yang dilakukan oleh Rahmah Widya Astuti, Badi'ah & Bagus Satrio (2019) bahwa algoritma K-Means untuk identifikasi topik pada dokumen Tugas Akhir informatika Unissula dapat mengelompokan dokumen pada topik yang sesuai, dan untuk mengetahui jumlah k menggunakan rumus SSE (Sum of Square Error) dihasilkan 3 cluster yang memiliki nilai maksimal atau terbaik. Dengan jumlah cluster 3 menjadi titik Elbow jumlah cluster terbaik karena mengubah penurunan hasil SSE yang awalnya berkurang 2 terus menerus menjadi berkurang 1 sehingga terbentuk cluster 1,2 dan 3 [4].

Penelitian terbaru juga dilakukan oleh Muhammad Rafi Muttaqin & Meriska Defriani (2020) dengan menggunakan metode K-Means clustering hasil pengelompokan dokumen topik skripsi dengan nilai cluster yang paling tinggi dapat menunjukan kemampuan mahasiswa pada tiap kelompok bidang keahlian. Jumlah matakuliah yang paling banyak pada suatu kelompok bidang keahlian menandakan bahwa mahasiswa memiliki kemampuan lebih baik pada bidang tersebut sehingga direkomendasikan topik skripsi yang sesuai dengan kelompok bidang keahlian tersebut [5,6].

B. Library Management Systems

Library Management System (LMS) merupakan sistem informasi perpustakaan berbasis website yang digunakan untuk membantu pihak perpustakaan dalam mengelola data perpustakaan, otomatisasi dan meningkatkan

pelayanan. Pada penelitian ini LMS yang akan digunakan yaitu website digital library Perbanas.

C. Machine Learning`

Machine Learning (ML) atau pembelajaran mesin adalah pendekatan dalam Artificial Intelligence (AI) yang banyak digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan masalah atau melakukan otomatisasi. Sesuai namanya, Machine Learning (ML) mencoba menirukan bagaimana proses manusia atau makhluk cerdas belajar dan menggeneralisasi (Ahmad, 2017) [1].

Machine Learning (ML) membantu menangani dan memprediksi data yang sangat besar dengan cara merepresentasikan data-data tersebut dengan algoritma pembelajaran. Machine learning dapat membantu komputer memprogram diri mereka sendiri. Jika pemrograman adalah pekerjaan untuk membuat otomatis, maka machine learning mengotomatisasi proses otomatis. Pada dasarnya machine learning membiarkan data melakukan pekerjaan.

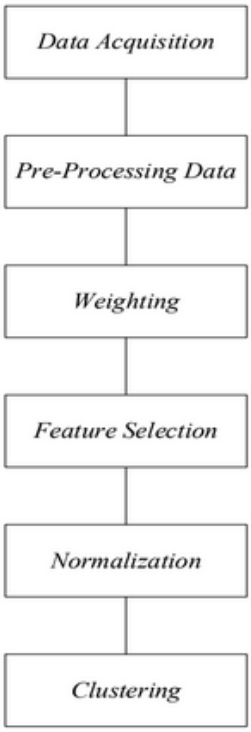
Klasterisasi adalah adalah proses pengelompokan dataset dokumen yang merujuk pada similarity (kemiripan) pola data dokumen ke dalam suatu cluster, sedangkan yang tidak memiliki kemiripan akan dikelompokkan ke dalam cluster yang lain. Ada empat jenis cara pembelajaran pada Machine Learning (ML), yaitu :

- *Supervised Learning*, data pembelajaran mencakup keluaran yang sudah ditentukan.
- *Unsupervised Learning*, data pembelajaran tidak mencakup keluaran yang ditentukan.
- *Semi-supervised Learning*, data pembelajaran mencakup beberapa keluaran yang ditentukan.
- *Reinforcement Learning*, pemberian hadiah dari setiap serangkaian tindakan yang dilakukan.

D. Algoritma K-Means Clustering

K-means adalah salah satu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa supervise (unsupervised). K-means clustering juga merupakan salah satu metode yang melakukan pengelompokkan data dengan sistem partisi, metode k-means berusaha mengelompokkan data yang ada ke dalam beberapa kelompok.

Kelompok yang terbentuk akan membentuk karakteristik yang sama satu sama lainnya dan akan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok lain. Metode ini berusaha meminimalkan variasi antar data yang ada didalam suatu cluster dan memaksimalkan variasi dengan data yang ada di cluster lainnya.



Gambar 1. Tahapan Penelitian

Adapun langkah-langkah pada algoritma K-Means Clustering adalah sebagai berikut:

1. Menentukan jumlah k sebagai cluster yang ingin dibentuk
2. Menentukan pusat cluster secara acak sebanyak k
3. Menentukan jarak setiap data terhadap pusat cluster (centroid)
4. Mengelompokkan setiap data yang bersangkutan berdasarkan kedekatannya dengan centroid (jarak terkecil)
5. Menentukan pusat cluster baru. Memperbaharui nilai centroid dari rata-rata cluster yang bersangkutan dengan menggunakan rumus :

$$y_j(t+1) = \frac{1}{N_{sj}} \sum j \in sj \; x_j$$

Dimana:

$y_j(t+1)$ merepresentasikan centroid baru pada iterasi ket+1 dan N_{sj} merepresentasikan banyaknya data pada cluster j.

6. Mengulangi langkah 3 hingga 5 sampai anggota yang ada pada tiap cluster tidak berubah

Tabel 1. Hasil Klasterisasi Dokumen

Program Studi	**Jumlah Dokumen**	**Jenis Cluster**	**Anggota Cluster Terklasifikasi Benar**	**Anggota Cluster Terklasifikasi Tidak Benar**
Teknik Informatika	100	*Cluster* 0	69 *items*	31 *items*
Manajemen	100	*Cluster* 1	100 *items*	0 *items*
Sistem Informasi	100	*Cluster* 2	70 *items*	30 *items*

7. Jika langkah 6 sudah terpenuhi, maka nilai pusat cluster pada perulangan terakhir akan digunakan sebagai parameter untuk kelompok dokumen skripsi.

III. METODOLOGI

Pada bagian ini akan dijelaskan data yang digunakan dalam penelitian, metode dan alat bantu yang digunakan, dan tahapan penelitian yang dilakukan.

A. Data Penelitian

Data dokumen skripsi yang digunakan pada penelitian ini berasal dari website digital library Perbanas Institute. Untuk percobaan sistem peneliti menggunakan 300 data dokumen skripsi, setiap seratus dokumen skripsi yang digunakan berasal dari program studi yang berbeda. Untuk program studi yang digunakan yaitu Program Studi manajemen, Program Studi Sistem Informasi, dan Program Studi Teknik Informatika.

B. Tahapan Penelitian

Penelitian ini dilakukan dengan melaksanakan beberapa tahapan penelitian dimulai dari pengumpulan data (*data acquistion*), melakukan pengolahan data mentah sehingg dapat diolah lebih lanjut (pre-processing data), pembobotan kata (weighting), pemilihan serta penyaringan fitur (fiture selection), normalisasi kata (normalization), proses klasifikasi menggunakan algoritma K-Means Clustering (clustering), dan analisis seperti yang ditunjukkan pada Gambar 1.

IV. HASIL DAN PEMBAHASAN

Pada penelitian ini telah didapatkan hasil clustering 300 dokumen skripsi dari tiga program studi di Perbanas Institute dengan menggunakan metode k-means Clustering. Hasil yang didapatkan berdasarkan proses klasterisasi dokumen dapat dilihat pada Tabel 1.

Dengan mengacu pada Tabel 1 dapat dilihat bahwa variasi hasil dari anggota setiap cluster. Hasil anggota cluster 1 lebih baik dibandingkan anggota cluster 0 dan anggota cluster 2 dilihat dari jumlah anggota cluster yang tertebak benar. Pada anggota cluster 1 atau cluster program studi Manajemen dapat dilihat bahwa metode K-Means Clustering dapat dengan baik mengklasterikasikan dokumen skirpsi yang digunakan dalam data set. Hal ini mungkin dikarenakan oleh keyword yang digunakan pada judul skripsi program studi Manajemen termasuk unik dan tidak digunakan pada judul skripsi Teknik Informatika dan Sistem Informasi, sehingga dapat terklasterisasi lebih baik dibandingkan program studi lainnya.

Sedangkan hasil anggota cluster 0 dan cluster 2 yang diperoleh hampir rata. Dimana untuk cluster 0 yang merepresentasikan program studi Teknik Informatika didapatkan bahwa dari 100 dokumen yang digunakan dalam data set, 69 dokumen skripsi dapat terklasifikasi dengan benar sebagai program studi Teknik Informatika sedangkan 31 dokumen skripsi tidak terklasifikasi dengan benar dan terklasifikasi sebagai dokumen skripsi program studi Sistem Informasi. Sedangkan untuk cluster 2 yang merepresentasikan program studi Sistem Informasi didapatkan bahwa dari 100 dokumen yang dijadikan sebagai data set, 70 dokumen skripsi dapat terklasifikasikan dengan benar sebagai dokumen skripsi program studi Sistem Informasi sedangkan 30 dokumen terklasifikasikan tidak benar dan terklasifikasi sebagai dokuemen skripsi program studi Teknik Informatika. Hal ini dikarenakan antara keyword pada judul skripsi Teknik Informatika dan Sistem Informasi saling beririsan atau hampir sama antara satu dengan yang lainnya.

V. KESIMPULAN

Dari hasil penelitian dan pembahasan yang telah dilakukan dapat diambil kesimpulan bahwa clustering dokumen menggunakan K-Means Clustering dapat dilakukan pada dokumen skripsi LMS Perbanas Institute. Dari hasil clustering dengan memasukkan 300 dokumen skripsi dari program studi manajemen, sistem informasi, dan teknologi informasi didapatkan bahwa 100 dokumen untuk program studi manajemen dapat terklasterisasi dengan benar, 70 dokumen untuk program studi sistem informasi dapat terklasifikasi dengan benar dan 30 dokumen terklasifikasi tidak benar, dan 69 dokumen untuk program studi teknik informatika dapat terklasifikasi dengan benar dan 31 dokumen terklasifikasi tidak benar dengan menggunakan metode K-Means Clustering. Sehingga dapat disimpulkan bahwa metode K-Means Clustering dapat cukup baik diimplementasikan pada kasus library management systems Perbanas Institute.

Dari hasil penelitian yang didapatkan bahwa pemilihan fitur dipilih berdasarkan teknik univariate statistical test. Sehingga disarankan untuk penelitian selanjutnya untuk

mencoba metode pemilihan fitur yang dilakukan secara manual sesuai dari konteks dokumennya. Kemudian untuk saran berikutnya yaitu menambahkan metode clustering dan mengoptimasi metode K-Means Clustering dengan metode lainnya.

REFERENCES

[1] A. Ahmad, "Mengenal Artificial Intelligence, Machine Learning, Neural Network, dan Deep Learning," Jurnal Teknologi Indonesia, 2017

[2] S. Munifah, A. Syukur, dan C. Supriyanto, "Pengelompokkan Arsip Universitas menggunakan Algoritma K-Means dengan Fetature Selection Chi Square," Jurnal Teknologi Informasi Vol. 11 No. 2. Semarang : Universitas Dian Nusantoro, 2015

[3] V. K. Bakti, dan J. Indiyanto, "Klasterisasi Dokumen Tugas Akhir Menggunakan K-Means Clustering sebagai Analisa Penerapan Sistem Temu Kembali," Jurnal Ilmiah Manajemen Informatika dan Komputer, vol. 01, 2017

[4] R. W. Astuti, Badi'ah, dan B. Satrio, "Rancang Bangun Sistem Klasterisasi Dokumen Menggunakan Metode K-Means Untuk Identifikasi Topik Dokumen tugas Akhir Program Studi Teknik Informatika Universitas Islam Sultan Agung," Konferensi Ilmiah Mahasiswa Unissula. Semarang : Universitas Islam Sultan Agung, 2019

[5] M. R. Muttaqin, dan M. Defriani, "Algoritma K-Means untuk Pengelompokkan Topik Skripsi Mahasiswa," Jurnal Ilmiah ILKOM Vol. 12 No. 2. Purwakarta : STT Wastukancana, 2020

[6] W. T. H. Putri dan R. Hendrowati, "Penggalian Teks dengan Model Bag Of Words terhadap Data Twitter," Jurnal Muara Sains, Teknologi, Kedokteran, dan Ilmu Kesehatan Vol. 02 No. 01 Hal. 129-138. Jakarta : Universitas Paramadina Jakarta, 2018