

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi memiliki dampak yang sangat signifikan dalam kehidupan sehari-hari, mulai dari kegiatan yang sederhana hingga kegiatan yang membutuhkan tingkat ketelitian yang tinggi. Kegiatan yang umum dilakukan oleh sebuah instansi adalah kegiatan pengarsipan dokumen, baik dokumen dalam bentuk fisik maupun elektronik. Umumnya kegiatan pengarsipan melibatkan dokumen dengan jumlah yang cukup besar, sehingga diperlukan suatu metode yang praktis dan efisien dalam pengelolaannya. Salah satu metode yang digunakan dalam pengelolaan dokumen adalah pengklasteran atau pengklasifikasian dokumen.

Pengklasifikasian dokumen elektronik dengan jumlah yang banyak diperlukan agar data yang terkumpul dapat diproses menjadi informasi yang tepat. Pengklasifikasian dokumen dilakukan dalam upaya memisahkan atau mengelompokkan dokumen berdasarkan ciri-ciri atau kategori tertentu. Dengan banyaknya dokumen proses pengklasifikasian tidak mungkin dilakukan secara manual karena memerlukan banyak waktu dan tenaga. Salah satu metode yang dapat digunakan adalah dengan pengklasifikasian secara otomatis dengan *text mining*. Banyak metode *text mining* yang digunakan dalam mengklasifikasikan dokumen atau teks, salah satunya adalah algoritma *cosine similarity*. Permasalahan yang dihadapi pada *text mining* adalah jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah, serta data noise.

Sehingga sumber data yang digunakan pada *text mining* adalah kumpulan teks yang memiliki bentuk yang tidak terstruktur atau setidaknya semi terstruktur [1]. Klustering biasa digunakan pada banyak bidang, seperti text mining, pengenalan pola (pattern recognition), pengklasifikasian gambar (image classification), ilmu biologi, pemasaran, perencanaan kota, pencarian dokumen, dan lain sebagainya. Tujuan dari klustering adalah untuk menentukan pengelompokan dari suatu set data. Akan tetapi tidak ada "ukuran terbaik" untuk pengelompokan data. Untuk pengelompokan data tergantung tujuan akhir dari klustering, maka diperlukan suatu kriteria sehingga hasil klustering seperti yang diinginkan [2].

Pengklasifikasian teks sangat dibutuhkan dalam berbagai macam aplikasi, terutama aplikasi yang jumlah dokumennya bertambah dengan cepat. Ada dua cara dalam penggolongan teks, yaitu text clustering dan klasifikasi teks. *Text clustering* berhubungan dengan menemukan sebuah struktur kelompok yang belum kelihatan (tak terpandu atau *unsupervised*) dari sekumpulan dokumen. Sedangkan pengklasifikasian teks dapat dianggap sebagai proses untuk membentuk golongan golongan (kelas-kelas) dari dokumen berdasarkan pada kelas kelompok yang sudah diketahui sebelumnya (terpandu atau *supervised*) [2]. Metode *cosine*, *jaccard*, dan *k-nearest neighbor* (K-NN) yang digunakan pada proses klasifikasi dokumen teks dengan hasil akhir dari percobaan 33 kali dengan key yang berbeda dan total 6326 dokumen didapat metode cosine yang nilai kemiripannya tertinggi yaitu 41% dari metode jaccard 19% dan k-nearest neighbor (K-NN) 40%, karena metode cosine similarity mempunyai konsep

normalisasi panjang vektor data dengan membandingkan Ngram yang sejajar satu sama lain dari 2 pembandingan [3].

Objek penelitian ini adalah data Judul Buku dalam bentuk elektronik dengan menspesifikasikan genre buku keagamaan. Data Judul Buku akan diklasifikasikan ke dalam beberapa kategori secara otomatis seperti Akhlaq, Al-Qur'an & Hadits, Dialog Agama, Fiqih & Muamalat, Hikmah, Kalam, Sejarah Islam, Spiritualitas, Tasawuf, Wacana Islam, Zikir & Do'a. Untuk itu, judul penelitian yang dilakukan untuk tugas akhir ini adalah adalah **“IMPLEMENTASI COSINE SIMILARITY UNTUK PENENTUAN GENRE BUKU BERKONTEN ISLAMI DI PT MIZAN MEDIA UTAMA”**

1.2 Rumusan Masalah

Berdasarkan dari latar belakang, maka dapat dirumuskan menjadi beberapa masalah diantaranya:

1. Bagaimana cara penerapan *text preprosesing* pada buku di PT. Mizan media utama?
2. Bagaimana akurasi *performance* algoritma *cosine similarity* pada dokumen judul buku di PT. Mizan?

1.3 Tujuan

Adapun tujuan dari penelitian ini untuk:

- a. Mengetahui cara penerapan *text preprosesing* pada buku di PT. Mizan media utama.

- b. Mengetahui tingkat akurasi *performance* algoritma *cosine similarity* pada dokumen judul buku di PT. Mizan.

1.4 Batasan Masalah

Berdasarkan latar belakang yang telah dipaparkan, maka rumusan masalah dalam penelitian ini adalah:

- a. Bahasa pemrograman yang digunakan adalah bahasa pemrograman PHP
- b. Algoritma yang digunakan adalah *cosine similarity* dan TFIDF.
- c. Genre yang dipilih tentang genre buku keagamaan dengan 11 kategori (Akhlak, al-qur'an & hadits, dialog agama, fiqh & muamalat, hikmah, kalam, sejarah islam, spiritualitas, tasawuf, wacana islam, zikir & do'a)
- d. Fitur dalam aplikasi yaitu Upload judul buku, pengelolaan akun, pengelolaan buku, klasifikasi jenis buku dan hasil akurasi.
- e. Aplikasi yang dibangun hanya untuk menganalisis penerapan *text preprossing* dan tingkat akurasi algoritma *cosine similarity* pada judul buku di PT Mizan Media Utama.
- f. Aplikasi yang dibangun menggunakan metode pengembangan perangkat lunak *prototype*.

1.5 Metodologi Penelitian

Untuk dapat memecahkan masalah pada penelitian ini maka dapat digunakan metodologi penyelesaian sebagai berikut:

a. Studi Pustaka

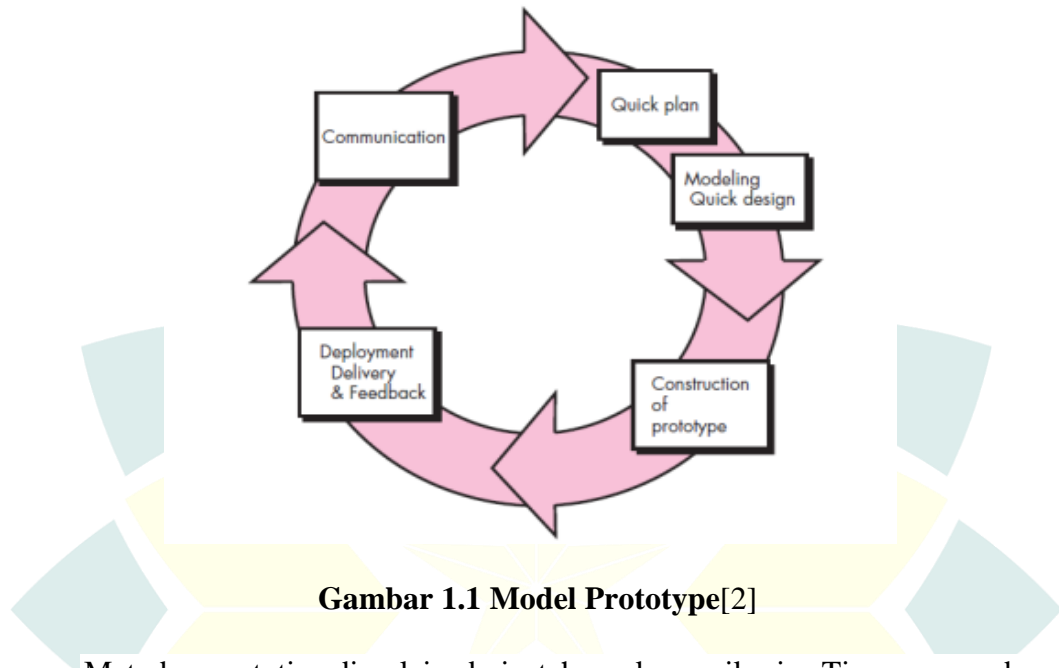
Pencarian informasi dan pemahaman literatur melalui berbagai media, referensi dari buku, jurnal ilmiah, internet dan forum yang berkaitan dengan ilmu pengetahuan tentang Pengolahan data text, *Cosine Silmilarity* dan TF-IDF untuk proses pengklasifikasian. Pada tahap ini, menggunakan penerapan tahap pertama pada metode pengembangan *prototype* yaitu *listen to costumer* yaitu dengan wawancara kepada para ahli yang sudah memahami tentang ilmu yang berkaitan dengan pembuatan aplikasi ini, untuk mengumpulkan materi yang dibutuhkan dalam membangun Sistem Klasifikasi Gendre Buku di Pt. Mizana Media Utama menggunakan *Cosine Similarity* dan dan TF-IDF.

b. Pemodelan Sistem

Pada teknik ini dilakukan perancangan aplikasi menggunakan metode *Unified Modelling Language* (UML), kemudian diterapkan pada pembuatan aplikasi menggunakan bahasa pemograman PHP dan HTML.

c. Metode pengembangan Sistem

Adapun metode pengembangan perangkat lunak yang akan dibuat yaitu menggunakan metodologi *Prototype*. Gambar 1.1 merupakan alur dari metode *prototype*.



Gambar 1.1 Model Prototype[2]

Metode prototipe dimulai dari tahap komunikasi. Tim pengembang perangkat lunak melakukan pertemuan dengan para *stakeholder* untuk menentukan kebutuhan perangkat lunak yang saat itu diketahui dan untuk menggambarkan area-area dimana definisi lebih jauh untuk iterasi selanjutnya.

Perencanaan iterasi pembuatan prototipe dilakukan secara cepat. Setelah itu dilakukan pemodelan dalam bentuk “rancangan cepat”. Pembuatan rancangan cepat berdasarkan pada representasi aspek-aspek perangkat lunak yang akan terlihat oleh para *end user* (misalnya rancangan antarmuka pengguna atau format tampilan). Rancangan cepat merupakan dasar untuk memulai konstruksi pembuatan prototipe.

Prototipe kemudian diserahkan kepada para *stakeholder* untuk mengevaluasi *prototipe* yang telah dibuat sebelumnya dan memberikan umpan balik yang akan digunakan untuk mempeluas spesifikasi kebutuhan.

Iterasi akan terjadi pada saat prototipe diperbaiki untuk memenuhi kebutuhan dari para *stakeholder*.

Tahapan-tahapan dalam *Prototyping* adalah sebagai berikut:

- a. *Communication* (komunikasi), yaitu analisis terhadap kebutuhan pengguna.
- b. *Quick Design* (desain cepat), yaitu pembuatan desain menggunakan UML secara cepat untuk selanjutnya dikembangkan kembali.
- c. *Contruccion of Prototype* (Pembentukan *Prototype*), yaitu pembuatan perangkat prototype termasuk pengujian dan penyempurnaan.
- d. Evaluasi terhadap prototype, yaitu mengevaluasi *prototype* dan memperhalus analisis terhadap kebutuhan pengguna.
- e. Perbaiki *prototype*, yaitu pembuatan type yang sebenarnya berdasarkan hasil dari evaluasi *prototype*.
- f. Produksi akhir, yaitu memproduksi perangkat secara benar sehingga dapat digunakan oleh pengguna.

Salah satu hal terpenting mengenai metodologi ini, cepat atau lambat akan disingkirkan dan hanya digunakan untuk tujuan dokumentasi. Kelemahannya adalah metode ini tidak memiliki analisa dan rancangan yang mendalam yang merupakan hal penting bagi sistem yang sudah kokoh, terpercaya dan bisa dikelola.

1.6 State of The Art

Tabel 1.1 adalah hasil penelitian yang telah dilakukan dan memiliki hubungan dengan penelitian yang akan dilakukan antara lain:

Tabel 1.1 State of The Art

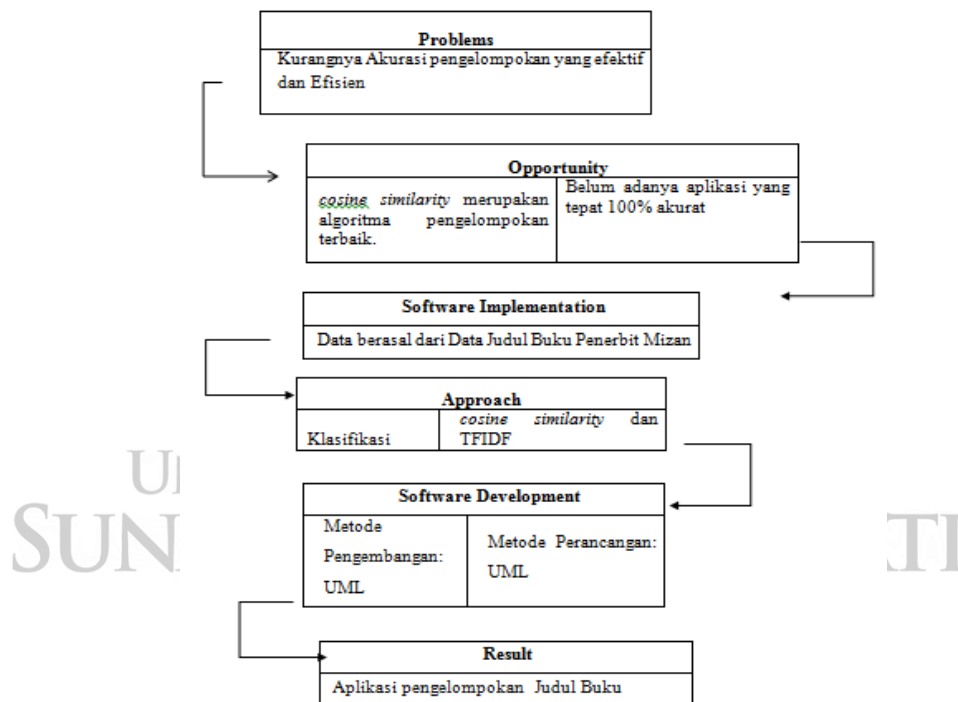
No	Peneliti	Data	Metode	Tujuan	Hasil
1	Sugiyamta	Dokumen Elektronik	Cosine Similarity Dan Single Pass Clustering	membuat sistem deteksi kemiripan dokumen menggunakan Algoritma Cosine Similarity dan teknik mengelompokkan dokumen dengan Algoritma Single Pass Clustering	Hasil pengukuran kemiripan dengan cosine similarity untuk dokumen abstrak tersebut setelah dibandingkan antara pengukuran manual dan sistem, mempunyai tingkat akuransi 99%
2	Indri Sudanawati Rozas	Dokumen Elektronik	Metode Bayesian Network dan Cosine Similarity	Pemilihan kontrol berdasarkan nilai similarity	Pemilihan kontrol keamanan informasi menggunakan cosine similarity dan term standar pada data threat memiliki nilai akurasi rata-rata sebesar 70% sedangkan menggunakan term yang diperluas mencapai 90%.
3.	Rizki Tri Wahyuni , Dhidik Prastiyanto , dan Eko Suprptono	Dokumen Elektronik	Algoritma Cosine Similarity dan Pembobotan TF-IDF	Tujuan penelitian ini adalah membuat sebuah sistem yang dapat mengklasifikasi dokumen secara otomatis menggunakan algoritma cosine similarity dalam proses pengklasterannya dan menggunakan metode	Hasil pengujian tingkat kelayakan sistem menunjukkan hasil bahwa sistem sangat layak dengan persentase rata-rata hasil pengujian sebesar 88,3%. hasil pengujian tingkat kelayakan dengan beberapa pakar kearsipan menunjukkan bahwa sistem juga

No	Peneliti	Data	Metode	Tujuan	Hasil
				pembobotan TFIDF.	sangat layak untuk digunakan dengan persentase rata-rata hasil pengujian sebesar 87,5%.
4	Verry Prima Anindya Rhosady	Dokumen Elektronik	Cosine Similarity	Membangun sebuah aplikasi pengklasifikasi halaman web berdasarkan <i>content</i>	Dari hasil penelitian tampak bahwa algoritma <i>Cosine Similarity</i> cocok untuk pengklasifikasian dokumen namun tidak tepat 100%, dikarenakan kurang banyaknya data latih yang ditanam pada aplikasi.
5	Danang Setyo Nugroho, Yanuar Firdaus A.w., Warih Maharani	Dokumen Elektronik	Metode Cosine Similarity Dan Correlation Based Similarity	Membangun aplikasi recommender system berbasis item-based collaborative filtering yang mengimplementasikan metode Cosine based Similarity dan correlation based similarity	metode cosine similarity menunjukkan nilai yang lebih baik karena nilai similarity dikatakan baik apabila nilainya berada di interval antara 0.0 sampai 1.0 dan hasilnya metode cosine selalu mendekati 1 . sedangkan pada metode correlation similarity terdapat interval antara -1 sampai 1.0
6.	Ogie Nurdiana, Jumadi, Dian Nursantika	Dokumen Elektronik	cosine similarity, algoritma jaccard similarity, knearest neighbor	Mengimplementasikan text mining menggunakan perbandingan algoritma cosine similarity dengan algoritma	metode cosine similarity menjadi usulan alternatif untuk mencari kemiripan dari teks mining

No	Peneliti	Data	Metode	Tujuan	Hasil
			(K-NN)	jaccard	
7	Moh.iqbal	Dokumen Elektronik	Cosine similarity, TF-IDF	Mengimplemtna sikan text mining menggunakan algoritma cosine similarity dengan pembobotan TF-IDF	metode cosine similarity menunjukkan hasil akurasi 75,92% dikarenakan tiap dokumen. Hal ini desabkan faktor dari setiap dokumen memiliki imputan data dan jumlah kata yang sedikitdari setiap judul atau dokumen

1.7 Kerangka Pemikiran

Adapun kerangkan pemikiran dari Aplikasi yaitu pada Gambar 1.1



Gambar1.2 Kerangka Pemikiran

1.8 Sistematika Penulisan

Sistematika penulisan dari tiap bab dalam laporan tugas akhir ini bertujuan untuk mendapatkan keterarahan dan sistemasi dalam penulisan sehingga mudah dipahami, adapun sistematika secara umum dari penulisan laporan ini adalah:

BAB I PENDAHULUAN

Bab I merupakan pengantar yang memberikan gambaran mengenai permasalahan- permasalahan yang kemudian akan dibahas pada bab-bab selanjutnya. Terdapat beberapa pokok bahasan dalam bab ini, yaitu latar belakang, rumusan masalah, batasan masalah, tujuan, metodologi penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Pada bab II akan dijelaskan tentang teori-teori yang digunakan dalam analisa permasalahan yang ada, teori-teori yang digunakan dalam perancangan dan implementasi serta hal-hal yang berguna dalam proses penyelesaian penulisan tugas akhir ini.

BAB III ANALISIS DAN PERANCANGAN

Bab III membahas mengenai analisis dari permasalahan yang ada saat ini dan analisis kebutuhan yang diperlukan untuk mengatasi permasalahan tersebut. Pembuatan desain dari aplikasi dengan mengacu pada analisis yang dibahas. Desain aplikasi yang akan dijelaskan terbagi menjadi tiga bagian, meliputi desain *user interface*, desain data dan desain proses. Pada BAB III, akan menggunakan tahap kedua pada *prototype* yaitu *build / revise mockup*. *Prototype* yang dibuat disesuaikan dengan kebutuhan sistem yang telah didefinisikan sebelumnya dari keluhan pengguna.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Pada bab IV akan dijelaskan tentang spesifikasi aplikasi, kebutuhan aplikasi, implementasi aplikasi, dan pengujian yang dilakukan terhadap aplikasi yang dibangun. Pengujian yang dilakukan oleh *user* akan mengetahui apakah aplikasi yang telah dibangun telah dapat menyelesaikan permasalahan yang ada. Pada BAB IV akan diterapkan pengembangan *prototype* tahap ketiga yaitu *customer test drives mockup* yang akan diterapkan pada pembuatan aplikasi ini. Pada tahap ini, *prototype* dari sistem diuji coba oleh pelanggan atau pengguna kemudian dilakukan evaluasi terhadap kekurangan-kekurangan dari kebutuhan pelanggan. *Developer* kemudian kembali mendengarkan keluhan pengguna dan melakukan perbaikan terhadap *prototype* yang ada.

BAB V PENUTUP

Bab V berisikan tentang kesimpulan yaitu jawaban dari rumusan masalah yang ada dan juga saran yang diperlukan yaitu penelitian selanjutnya yang dapat melebihi apa yang sudah ada pada batasan masalah penelitian ini.