BAB I
PENDAHULUAN

1.1 Latar Belakang

Penyebaran informasi dalam bentuk dokumen digital telah berkembang dengan pesat dan setiap waktu terus mengalami pertumbuhan dan jumlahnya semakin besar. Media massa versi elektronik dan situs web di internet merupakan dua contoh media yang menggunakan dan menyebarkan informasi berbentuk dokumen digital. Mengelola informasi dari kumpulan dokumen teks yang jumlahnya sangat besar tentunya bukan pekerjaan yang mudah. Oleh karena itu diperlukan sebuah metode yang dapat mengorganisir dan mengklasifikasi dokumen secara otomatis, sehingga dapat mempermudah dalam pencarian informasi yang relevan dengan kebutuhan (Samodra, Sumpeno, & Hariadi, 2009).

Bidang yang mempelajari teknik-teknik untuk pengorganisasian dokumen teks secara umum dibagi menjadi dua kelompok, yaitu *classification* dan *clustering. Clustering* teks 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*) (Darujati et al., 2012).

Dalam penelitiannya, Trisedya (2009) menjelaskan bahwa teknik klasifikasi dapat dilakukan dengan dua cara yaitu dengan pendekatan *supervised learning* dan pendekatan *unsupervised learning*. Teknik yang banyak digunakan dalam *unsupervised learning* adalah teknik *clustering*. *Clustering* merupakan teknik mengelompokkan dokumen-dokumen, sehingga dokumen yang memiliki kemiripan dikumpulkan dalam sebuah *cluster* tertentu. Pendekatan kedua adalah *supervised learning*. Pendekatan ini dilakukan dengan membangun sebuah *classifier* dari proses pembelajaran mengenai ciri dari tiap-tiap kategori yang ada. Pendekatan *supervised learning* dapat dibagi menjadi *fully supervised learning* dan *semi supervised learning*. *Fully supervised learning* adalah teknik klasifikasi dimana semua dokumen *training* telah diketahui kategorinya. Naïve Bayes adalah contoh dari teknik *fully supervised learning*, sedangkan *semi supervised learning* adalah teknik klasifikasi dimana pembelajaran dilakukan dari dokumen *training* yang telah diketahui kategorinya dan dokumen *training* yang belum diketahui kategorinya.

Berdasarkan penelitian di atas, untuk mempermudah pencarian informasi yang sesuai dengan yang kita inginkan dan sesuai dengan kategorinya, maka pengklasifikasifikasian dokumen akan membantu bagaimana mendapatkan informasi, sehingga mempermudahan pengolahan dan penggunaannya sesuai kebutuhan dan tujuan yang ingin dicapai. Selain itu, hal yang harus diperhatikan adalah bagaimana cara melakukan klasifikasi dokumen medis saat data yang ada terdiri dari dua jenis dokumen yang berbeda yaitu dokumen berlabel dan dokumen tidak berlabel. Selain itu, *labeled documents* (dokumen berlabel) hanya tersedia dalam jumlah yang kecil. Permasalahan dokumen pembelajaran untuk melakukan klasifikasi dokumen ini dapat diatasi dengan pendekatan baru yang dapat mempelajari *labeled data* maupun *unlabeled data* walaupun *labeled data* hanya tersedia dalam jumlah yang kecil. Pendekatan ini dikenal dengan nama pendekatan *semi supervised learning*.

Teknik *semi-supervised learning* adalah metode yang efisien untuk menambah data *training* secara otomatis dari data yang tidak berlabel (*unlabeled data*). Selain itu, perkembangan dari banyak aplikasi pengolahan bahasa (*natural language app*) menganggap masalah ini adalah sebuah tantangan dimana data yang tidak berlabel (*unlabeled data*) relatif dalam jumlah yang berlimpah sedangkan data berlabel (*labeled data*) jumlahnya agak terbatas (Qiu, Cho, Ma, & Campbell, 2019). Berbeda dengan pendekatan *supervised learning*, teknik *semi-supervised learning* dapat meningkatkan kinerjanya dengan meningkatkan informasi dalam data yang tidak berlabel. Beberapa hasil terbaru dari Laine & Aila (2017); Miyato et al (2019); Tarvainen & Valpola (2017) menunjukkan bahwa teknik *semi-supervised learning* dapat mencapai kinerja dari teknik *supervised learning* dalam skenario tertentu.

Pada penelitiannya, (Andini, 2013) menjelaskan bahwa saat ini sulit untuk mengetahui dokumen berdasarkan kebutuhan. Oleh karena itu, untuk mengetahui dokumen berdasarkan kebutuhan perlu dibantu oleh klasifikasi dokumen teks, yaitu suatu proses pengelompokan dokumen ke kategori yang dapat digunakan untuk melakukan analisis.

Klasifikasi dokumen medis adalah hal yang sangat penting mengingat penyebaran informasi yang berkaitan dengan kesehatan tersebar luas secara digital. Untuk itu penting juga untuk kita mengetahui kategori-kategori yang ada dalam bidang kesehatan agar informasi yang diberikan dapat teroganisir dengan baik, selain itu informasi yang disampaikan juga dapat digunakan untuk *information retrieval*. Hal tersebut membuat penulis menganggap bahwa manfaat dari mengklasifikasikan dokumen medis sangat penting agar sebuah dokumen dapat dikelompokkan ke dalam kategori tertentu didalam dunia kesehatan berdasarkan kata- kata dan kalimat-kalimat yang ada di dalam dokumen tersebut karena pada dasarnya kata atau kalimat

yang terdapat di dalam sebuah dokumen memiliki makna tertentu dan dapat digunakan sebagai dasar untuk menentukan kategori sesuai topik dari dokumen tersebut.

Oleh karena itu, perancangan sistem menggunakan teknik *Semi Supervised Learning* menggunakan metode *Multinomial Naïve Bayes* dan *Pseudo-labeling* dilakukan agar dapat melakukan klasifikasi dokumen medis dengan baik. Penggunaan metode *Multinomial Naïve Bayes* pada penelitian ini diharapkan mampu melakukan *multiclass classification* sehingga menghasilkan data akurat agar dapat dijadikan bahan penelitian lebih lanjut. Selain itu, untuk melakukan *multiclass classification* penggunaan metode *Multinomial Naïve Bayes* memiliki kemampuan untuk mengklasifikasi dokumen dengan kesederhanaan dan kecepatan komputasinya namun memiliki komputasi tinggi. Penggunaan *Pseudo-labeling* di manfaatkan untuk proses klasifikasi dari *unlabeled data* sehingga teknik *Semi Supervised Learning* dapat digunakan dengan baik pada penelitian ini. Dengan demikian, proses klasifikasi yang telah dilakukan dapat mempermudah pencarian informasi berdasarkan kategori tertentu yang dibutuhkan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, maka diperoleh sebuah rumusan masalah, yaitu :

a. Bagaimana mengklasfikasi dokumen medis dengan memanfaatkan teknik *semi supervised learning*?
b. Bagaimana menganalisis tingkat perbandingan akurasi teknik *semi supervised learning* untuk klasifikasi dokumen medis dengan porsi dokumen yang berbeda?

1.3 Batasan Masalah

Agar tidak menyimpang dari perumusan masalah yang ada, maka ditentukan batasan-batasan masalah. Berikut batasan masalah pada penelitian ini:

a. Menggunakan 700 *train data* yang terdiri dari *labeled data* dan *unlabeled data*.
b. Menggunakan 100 *test data* yang seluruh data terdiri dari *labeled data*.
c. Data dokumen berupa artikel kesehatan diperoleh dari situs kesehatan Indonesia (terdiri dari: alodokter.com, halodoc.com, sehatq.com, klikdokter.com, hellosehat.com, doktersehat.com).
d. Kategori ditentukan terdiri dari 10 (kategori: bayi, diabetes, diet, jantung, kecantikan, kehamilan, kesehatan gigi dan mulut, kolesterol, kulit, mata).

1.4 Tujuan Penelitian

Tujuan penelitian ini sebagai berikut yaitu:

a. Sistem dapat mengklasifikasi dokumen medis dengan memanfaatkan teknik *semi-supervised learning*.
b. Sistem dapat menganalisis tingkat perbandingan akurasi teknik *semi supervised learning* untuk klasifikikasi dokumen medis dengan porsi dokumen yang berbeda.

1.5 Manfaat Penelitian

Adapun manfaat penelitian ini sebagai berikut:

a. Dapat membantu mempermudah *user* dalam memilih dan mengkategorikan dokumen.
b. Dapat meminimalkan waktu dan sumber daya manusia dalam pengklasifikasian dan pencarian dokumen dalam jumlah yang besar.
c. Untuk mengatasi gap, dimana saat kita mencari suatu informasi tertentu, banyak hal yang penting justru terlewatkan, malah yang tidak penting banyak terserap.

1.6 Metodologi Penelitian

Metode yang digunakan untuk menyelesaikan masalah pada penelitian ini akan diselesaikan seperti tahapan dibawah ini, berikut tahapannya:

a. Identifikasi Masalah

   Pada tugas akhir ini klasifikasi dokumen teks dilakukan dengn menggunakan metode *Multinomial Naïve Bayes* untuk melakukan klasifikasi dokumen medis dengan teknik *semi-supervised learning*.

b. Studi Literatur

   Penelitian ini melihat beberapa literatur yang berkaitan dengan pemanfaatan teknik *semi-supervised learning* dengan merujuk beberapa literatur yang membahas topik yang sama. Seperti dalam beberapa jurnal, buku, dan blog.

c. Pengumpulan Data

   Data yang digunakan pada percobaan tugas akhir ini terdiri dari 700 data *training* dan 100 data *testing*. Data yang digunakan terbagi menjadi 10 kategori yang berkaitan dengan medis yaitu kesehatan bayi, diabetes, diet, jantung, kecantikan, kehamilan, kesehatan gigi dan mulut, kolesterol, kulit, mata.

d. *Pre-processing*

Sebelum proses klasifikasi dilakukan, data yang sudah berhasil dikumpulkan harus melalui proses *cleaning* agar data yang dhasilkan bisa di normalisasi menjadi bentuk baku, normalisasi symbol dan tanda baca, mengubah semua huruf menjadi huruf kecil (*casefolding*), *stemming* yaitu mengubah kata berimbuhan menjadi kata dasar, serta menghilangkan *Stopword.* Selanjutnya, agar dokumen-dokumen akan lebih mudah melalui proses klasifikasi, seluruh data bersih hasil *preprocessing* akan disimpan ke dalam satu dokumen baru.

e. *Feature Extraction*

Penelitian ini menggunakan fitur pembobotan *tf-idf*. Pada pembobotan *tf-idf* nilai fitur akan dihitung berdasarkan kemunculan fitur pada sebuah dokumen dibagi dengan jumlah dokumen yang memiliki fitur tersebut.

f. *Training*

Metode *machine learning* yang akan digunakan pada percobaan tugas akhir ini adalah *Multinomial Naïve Bayes*. *Semi-supervised learning* memanfaatkan *labeled document* dan *unlabeled document* dalam proses *training*, agar dapat memanfaatkan dua jenis dokumen yang berbeda tersebut, penulis menggunakan teknik *psedo labeling* untuk melakukan klasifikasi dokumen medis. Dalam penerapannya, teknik *pseudo labeling* akan menggunakan metode *multinomial naïve bayes* untuk melakukan training data baik untuk *labeled data* dan *unlabeled data*.

1. *Multinomial Naïve Bayes*

   Multinomial Naïve Bayes merupakan metode *fully supervised learning* yang memerlukan tahap pembelajaran untuk membangun model probabilistik. Model probabilistik tersebut nantinya akan digunakan untuk melakukan perhitungan *prior* dan *conditional probability* dokumen *testing* dalam menentukan kategori dari dokumen *testing* tersebut. Pada penelitian ini, metode yang digunakan adalah *Multinomial Naïve Bayes*.

2. *Pseudo Labeling*

   Teknik *pseudo labeling*, tidak memberi label secara manual pada data yang tidak berlabel (*unlabeled data*). Namun, yang dilakukan pada teknik *pseudo labeling* ini adalah dengan memberikan perkiraan label berdasarkan data yang diberi label (*labeled data*).

g. Analisis dan Evaluasi

Pada penelitian ini dilakukan uji validitas terhadap data *testing*. Uji validitas menggunakan data *testing* sebagai proses validasi yang diperoleh dari training berupa model *classifier*. Model *classifier* tersebut adalah model *classifier* dari *labeled data*, model *classifier* dari *unlabeled data*, dan model *classifier* dari kombinasi *labeled data* dan *unlabeled data*. Dari proses uji validitas yang dilakukan akan diperoleh nilai akurasi dari masing-masing pengujian menggunakan model *classifier*. Nilai akurasi tersebut akan di analisis apakah menghasilkan akurasi yang baik, cukup baik, atau bahkan buruk.

h. Implementasi Aplikasi

Model *classifier* yang dibangun dari kombinasi *labeled data* dan *unlabeled data* akan di implementasikan ke dalam sebuah *website*. *Website* tersebut dapat melakukan prediksi terhadap data dokumen teks *input*an user menggunakan model *classifier* yang dapat dipilih user sesuai dengan model *classifier* yang dibangun menggunakan porsi data *training* yang berbeda.

1.7 Sistematika Penulisan

Sistematika penulisan bertujuan untuk memahami lebih jelas mengenai tugas akhir ini. Penulisan materi pada tugas akhir ini dikelompokkan menjadi beberapa bab yang secara garis besar adalah sebagai berikut :

BAB I PENDAHULUAN

Bab ini menguraikan tentang latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan tugas akhir ini.

BAB II LANDASAN TEORI

Bab ini menguraikan secara singkat tentang penelitian-penelitian yang telah dilakukan sebelumnya yang memiliki keterkaitan dengan tugas akhir ini. Pada bab ini juga dijelaskan tentang beberapa teori yang digunakan antara lain: Klasifikasi dokumen, *Machine Learning*, *Multiomial Naive Bayes*, *Psedo Labeling*.

BAB III METODOLOGI

Bab ini berisi tentang langkah-langkah identifikasi masalah berupa tujuan penyelesaian masalah, prosedur yang digunakan sumber dan teknik pengumpulan data. Kemudian akan dijelaskan mengenai model yang dibuat.

BAB IV HASIL DAN PEMBAHASAN

Pada bab ini berisi tentang tahapan dan hasil dari penelitian tentang hasil klasifikasi dokumen medis menggunakan teknik *semi-supervised learning* melalui data dokumen medis.

BAB V KESIMPULAN DAN SARAN

Pada bab ini, diuraikan tentang kesimpulan yang penulis dapatkan dari pengerjaan tugas akhir, dan saran pengembangan untuk perbaikan pada penelitian selanjutnya.