# *PEMANFAATAN TEKNIK SEMI-SUPERVISED LEARNING UNTUK KLASIFIKASI DOKUMEN MEDIS*

Dhomas Hatta Fudholi
Jurusan Teknik Informatika, Fakultas Teknologi Industri
Univeristas Islam Indonesia
Yogyakarta
e-mail : hatta.fudholi@fti.uii.ac.id

Kiki Purnama Juwairi
Jurusan Teknik Informatika, Fakultas Teknologi Indsutri
Universitas Islam Indonesia
Yogyakarta
e-mail : 15523234@students.uii.ac.id

***Abstract*— Penyebaran informasi dalam bentuk dokumen digital telah mengalami pertumbuhan yang sangat pesat. Dengan menggunakan metode klasifikasi teks, maka kumpulan dokumen yang jumlahnya sangat besar tersebut dapat diorganisir sedemikian rupa sehingga dapat mempermudah dan mempercepat pencarian informasi yang dibutuhkan. Maka dari itu perlu dibuat sistem terkait dokumen medis yang mana dokumen medis tersebut dapat diklasifikasi dengan tepat. Penelitian ini bertujuan untuk menerapkan metode klasifikasi *Naïve Bayes* dan algoritma *Pseudo Labeling* dalam mengelompokkan dokumen medis yang sesuai dengan dokumen yang diinput sehingga menghasilkan kelompok-kelompok yang sesuai. Uji coba dilakukan dengan menggunakan sampel dokumen medis yang diambil dari sebuah media kesehatan elektronik berbasis web. Hasil eksperimen menunjukkan bahwa metode *Naïve Bayes* menggunakan teknik *Semi-Supervised Learning* dapat digunakan secara efektif untuk mengklasifikasikan dokumen medis. Hal ini terlihat dari hasil eksperimen, yaitu dengan porsi *labeled documents* dan *unlabeled documents* yaitu 400:300 mampu mengklasifikasi dokumen dengan tepat mencapai 82.00%. Namun, pada hasil eksperimen dengan porsi *labeled documents* dan *unlabeled documents* yaitu 300:400 hanya mengklasifikasi dokumen dengan tepat mencapai 78.00%. Hasil penelitian diatas menunjukan bahwa jumlah *labeled documents* dan *unlabeled documents* sangat menentukan performa untuk melakukan klasifikasi.**

***Keywords—Multionomial Naïve Bayes; Pseudo Labeling; Semi-Supervised Learning; Klasifikasi; Dokumen medis***

## I. PENDAHULUAN

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, Sampeno & Hariadi, 2009).

Bidang yang mempelajari teknik-teknik untuk pengorganisasian dokumen teks secara umum dibagi menjadi dua kelompok, yaitu classification dan clustering. Menurut Darujati & Gumelar (2012) 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).

Berdasarkan penelitian diatas, 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*.

## II. TINJAUAN PUSTAKA

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, Sampeno & Hariadi, 2009).

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.

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 karna 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. 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. 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. Sangat 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.

Untuk itu, penulis berencana untuk merancang sistem yang dapat mengklasifikasi dokumen medis menggunakan teknik semi supervised learning. Dengan demikian, proses klasifikasi yang telah dilakukan dapat mempermudah pencarian informasi berdasarkan kategori tertentu yang dibutuhkan.

## III. METODE PENELITIAN

### A. Pengumpulan Data

Langkah selanjutnya adalah melakukan 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. Data training dari 10 kategori medis tersebut juga sudah dibagi menjadi data ber-label (labeled data) dan yang tidak ber-label (unlabeled data). Kemudian, proses pengujian dilakukan dengan membagi proses pengujian ke dalam empat tahap. Pengujian pertama menggunakan 400 data berlabel dan 300 data tidak berlabel (model 400:300), pengujian kedua menggunakan 300 data berlabel dan 400 data tidak berlabel (model 300:400), pengujian ketiga menggunakan 200 data berlabel dan 500 data tidak berlabel (model 200:500), dan yang terkahir dilakukan adalah pengujian keempat yaitu menggunakan 100 data berlabel dan 600 data tidak berlabel (model 100:600). Keempat pengujian tersebut akan di training dan model yang dihasilkan akan digunakan untuk melakukan klasifikasi dokumen medis.

### B. *Pre-processing*

Preprocessing adalah langkah yang harus dilakukan untuk membersihkan data. Data yang sudah dikumpulkan seringkali mengandung karakter-karakter yang tidak perlu yang harus dihilangkan atau biasa disebut dengan noise. Tidak hanya dihilangkan beberapa kata yang masih belum baku secara penulisan, slang word, dan bahasa asing pun juga perlu dinormalisasi dengan harapan dengan semakin baiknya data yang dimiliki hasil yang akan diperoleh dari proses selanjutnya dalam proses pengolahan bahasa ini akan menjadi lebih baik (Hidayatullah & Ma'arif, 2017).

- Menghilangkan non-ASCII.
- Menghilangkan Punctuation, Digit, URL, White space. Menghapus tanda baca seperti tanda baca titik, koma, tanda seru, tanda tanya, dan lain-lain.
- Mengubah semua huruf menjadi huruf kecil.
- Menghapus stopword. Stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah "yang", "dan", "di", "dari" dan seterusnya.
- Mengubah kata berimbuhan menjadi kata dasar. Mengubah kata berimbuhan menjadi kata dasar atau yang sering disebut dengan stemming.

### C. *Feature Extraction*

Feature Extraction adalah proses dari penggunaan informasi dari sekumpulan data untuk menciptakan fitur yang membuat algoritma machine learning bekerja. Penelitian ini menggunakan satu jenis fitur yaitu pembobotan tf-idf. Selanjutnya adalah menghitung tf-idf dengan menggunakan rumus logaritma seperti yang dapat dilihat pada persamaan (1)

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i}) \quad (1)$$

Berdasarkan persamaan (1) dapat kita lihat bahwa nilai tertinggi adalah jika sebuah kata jarang muncul dalam banyak dokumen, artinya kata ini memiliki kekuatan diskriminatif yang tinggi dalam sebuah dokumen. Untuk nilai yang terendah setelahnya menunjukkan kata ini muncul beberapa kali dalam dokumen yang berbeda sehingga menjadikan kata tersebut memiliki relevansi yang tidak jelas. Dan nilai yang paling rendah adalah kata yang selalu muncul di dalam banyak dokumen atau kata yang sama sekali tidak ada dalam dokumen.

D. *Training*

Training atau pelatihan adalah salah satu langkah penyelesaian untuk mengerjakan tugas akhir. Training dilakukan untuk melatih data yang sudah dipersiapkan agar dapat diklasifikasi. Metode machine learning yang akan digunakan pada percobaan tugas akhir ini adalah Multinomial Naïve Bayes dan untuk menerapkan metode Multinomial Naïve Bayes ke dalam teknik semi-supervised learning yang menggunakan labeled data dan unlabeled data secara bersamaan, Pseudo Labeling adalah teknik yang mendukung. Dua metode tersebut merupakan metode yang bisa digunakan pada teknik semi-supervised learning untuk klasifikasi dokumen medis.

- *Multinomial Naïve Bayes*

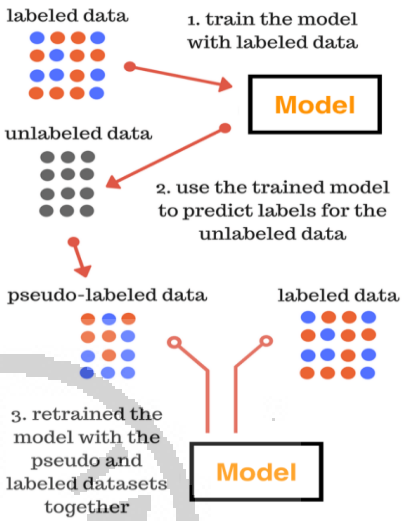
  Naïve Bayes merupakan metode *fully supervised learning* yang memerlukan tahap pembelajaran untuk membangun model probabilistik. Model probabilistik tersebut nantinya akan digunakan untuk melakukan perhitungan *prior* dan *conditional probability* dokumen *testing* dalam menentukan kategori dari dokumen *testing* tersebut. Naïve Bayes membangun model probabilistik dari *feature extraction* yang digunakan dari data *labeled*. Pembuatan model probabilistik selesai dilakukan, langkah terakhir yang dilakukan adalah penentuan kategori menggunakan persamaan.

$$P(c|term\ dokumen\ d) = P(c)x\ P(t_1|c)x\ P(t_2|c)x\ P(t_3|c)x\ \dots\ x\ P(t_n|c)$$

- *Pseudo Labeling*

  Proses klasifikasi dokumen medis menggunakan teknik Pseudo-Labeling juga melibatkan Multinomial Naïve bayes sebagai algoritma untuk melakukan klasifikasi pada labeled data. Cara kerja teknik pseudo-labeling ini adalah dengan cara menggunakan model yang dibangun dari proses klasifikasi menggunakan labeled data untuk melakukan prediksi terhadap unlabeled data. Hasil prediksi dari unlabeled data ini lah yang dmaksudkan sebagai pseudo-label data yang mana label hasil prediksi dianggap sebagai label sebenarnya dari data. Dan dengan memanfaatkan labeled data dan unlabeled data yang ada, maka dibangun model classifier baru. Model classifier tersebut yang nanti akan digunakan sebagai model classifier dari teknik semi-supervised learning.

$$\alpha + \beta = \chi. \quad (1) \quad (1)$$




Gambar 1 Proses *pesudo-labeling*

E. Analisis dan Evaluasi

Pada percobaan untuk tugas akhir ini digunakan 100 data untuk melakukan uji validitas untuk mengetahui efektifitas model yang sudah dibangun sebelumnya. 100 data tersebut merupakan data testing yang akan digunakan untuk menguji validitas dari model classifier yang dibangun.

Pada penelitian ini, analisis yang dilakukan adalah dengan menggunakan model 400:300, model 300:400, model 200:500, dan model 100:00. Empat model tersebut akan digunakan untuk melakukan klasifikasi pada seluruh data testing yang ada. Pada keempat model tersebut terdapat tiga model classifier yang dibangun. Model tersebut adalah model yang dibangun dengan data berlabel, data tidak berlabel, dan data kombinasi dari data berlabel dan tidak.

F. Implementasi Aplikasi

Langkah ini digunakan untuk membangun aplikasi berbasis website yang akan digunakan untuk melakukan klasifikasi dokumen medis. Dalam pengembangannya, website menggunakan Flask sebagai framework dalam Bahasa pemrograman python. Website yang dibangun akan digunakan sebagai predictor. Model yang digunakan untuk melakukan prediksi pada teks dokumen medis adalah model kombinasi data yang dibangun berdasarkan dua jenis data yaitu data berlabel dan data tidak berlabel. Model yang dibangun dibagi menjadi 4 porsi dokumen. Porsi pertama adalah 400 data berlabel dan 300 data tidak berlabel (model 400:300), porsi kedua adalah 300 data berlabel dan 400 data tidak berlabel (model 300:400), porsi ketiga adalah 200 data berlabel dan 500 data tidak berlabel (model 200:500), dan yang terakhir adalah porsi keempat berupa 100 data berlabel dan 600 data tidak berlabel (model 100:600).

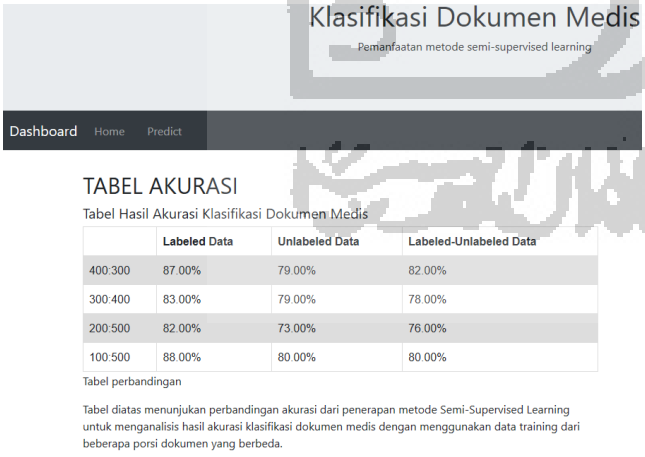
## IV. HASIL DAN PEMBAHASAN

Klasifikasi adalah proses untuk melakukan pembagian atau kategorisasi sesuatu kedalam kelas-kelas tertentu. Pada penelitian ini yang menjadi objek klasifikasi adalah dokumen medis yang terdapat pada website kesehatan yang ada di Indonesia dan berbahasa Indonesia. Penelitian ini menggunaan teknik Semi-Supervised Learning dengan menerapkan metode Pseudo Labeling dan menggunakan algoritma Multinomial Naive Bayes untuk melakukan klasifikasi pada dokumen medis. Proses klasifikasi yang dilakukan menggunakan labeled data (data berlabel) dan unlabeled data (data tidak berlabel) secara bersamaan sebagai data training. Pada data berlabel pemberian label dilakukan dengan memberi label pada data training dengan cara self-labeling pada dokumen medis. Sedangkan data tidak berlabel nantinya akan di klasifikasi dan menghasilkan pseudo-label.

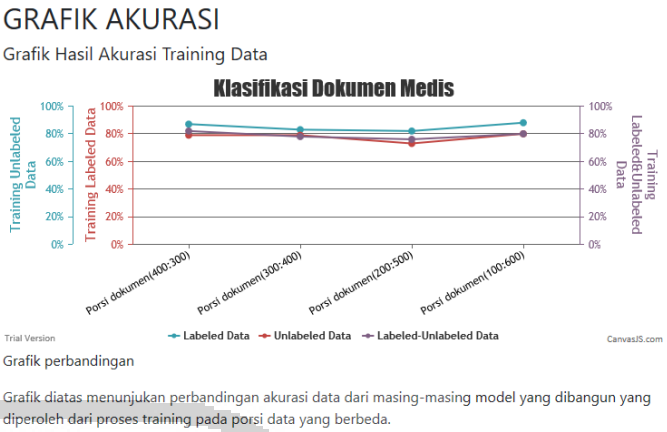
Pada proses training pengujian dilakukan dengan membagi porsi yang berbeda pada masing-masing jenis data. Pengujian pertama, menggunakan 400 data berlabel dan 300 data tidak berlabel (model 400:300). Pengujian kedua, menggunakan 300 data berlabel dan 400 data tidak berlabel (model 300:400). Pengujian ketiga, menggunakan 200 data berlabel dan 500 data tidak berlabel (model 200:500). Terakhir adalah pengujian keempat, menggunakan 100 data berlabel dan 600 data tidak berlabel (model 100:600). Proses klasifikasi yang dilakukan pada data training dengan masing-masing porsi dokumen berlabel dan dokumen tidak berlabel ini terdapat 10 kategori yang berbeda pada bidang kesehatan yaitu kesehatan bayi, diabetes, diet, jantung, kecantikan, kehamilan, kesehatan gigi dan mulut, kolesterol, kulit, mata.

### A. Analisis

Analisis dilakukan untuk mengetahui apakah model classifier yang sudah dibangun dengan menerapkan metode Pseudo Labeling menggunakan algoritma Multinomial Naïve Bayes mampu menghasilkan akurasi yang cukup baik dalam melakukan klasifikasi dokumen medis. Analisis dilakukan pada data testing berjumlah 100 data berlabel dengan 10 kategori berbeda dibidang medis yang akan dilakukan uji validitas.



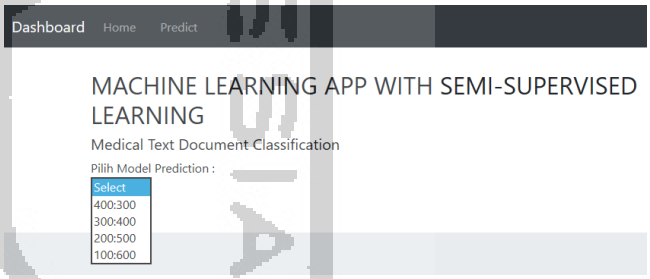

Gambar 2 Tabel akurasi hasil klasifikasi




Gambar 3 Grafik akurasi hasil klasifikasi

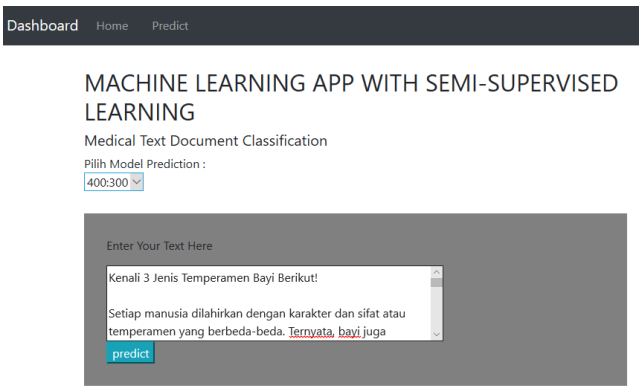
### B. Klasifikasi

Untuk melakukan klasifikasi pada website dibuah halaman prediksi yang berguna untuk melakukan prediksi data dokumen medis inputan user. Untuk melakukan klasifikasi, user harus memilih satu dari empat model classifier yang ada pada website. Model classifier pilihan digunakan untuk melakukan prediksi pada text input. Model classifier pada website adalah model classifier yang dibangun pada percobaan 1, 2, 3, dan 4 menggunakan data kombinasi. Jadi, proses klasifikasi teks dokumen medis ini melibatkan empat model classifier yaitu model 400:300, 300:400, 200:500, dan 100:600. Berdasarkan empat pilihan model classifier tersebut selanjutnya dibuat menu dropdown untuk memilih model classifier yang tersedia pada proses klasifikasi.



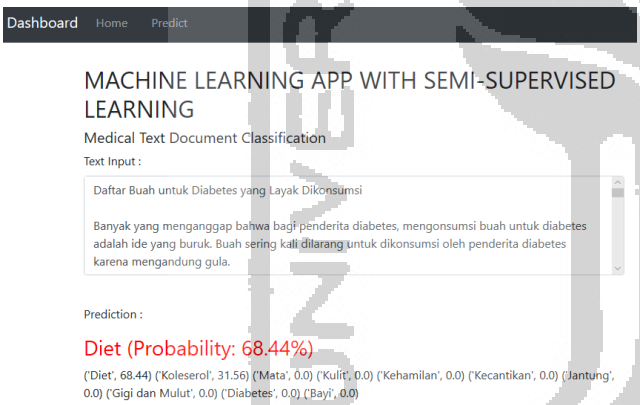

Gambar 4 Model *predictor*

Setelah memilih model *classifier* yang akan digunakan untuk melakukan klasifikasi seperti yang ditunjukan oleh Gambar 4 akan muncul *text box*. *Text box* tersebut digunakan untuk menginputkan teks dokumen medis yang akan di klasifikasi menggunakan model *classifier* yang dipilih.




Gambar 5 *Text input user*

Setelah memasukan teks berupa dokumen medis ke dalam *text box* dan memilih model *classifier* yang akan digunakan untuk klasifikasi pada *website*, langkah selanjutnya adalah melakukan prediksi teks yang sudah diinputkan. Prediksi dilakukan untuk mengetahui kategori dari teks yang diinputkan. Selain mengetahui kategori/kelas dari teks dokumen tersebut, *user* juga akan mengetahui nilai probabilitas dari setiap kategori dokumen yang ada pada sistem. Untuk memulai proses klasifikasi dan mengetahui kategori dari teks dokumen yang diinputkan, maka *user* harus menekan tombol *predict* yang berwarna biru. Setelah menekan tombol *predict* maka prediksi akan dilakukan oleh sistem dan hasil dari proses prediksi tersebut akan ditampilkan pada *website*.




Gambar 6 Hasil klasifikasi dokumen medis

Berdasarkan Gambar 6 menunjukan hasil klasifikasi yang diperoleh dari teks dokumen medis. Untuk setiap proses klasifikasi menggunakan setiap model akan mengeluarkan *output* hasil prediksi yang dilakukan model *classifier* yang dipilih. *Output* yang ditampilkan oleh *website* berupa informasi *text input* yaitu teks dokumen medis yang di klasifikasi. Kemudian, ada hasil prediksi dan nilai probabilitas dari perhitungan klasifikasi yang dilakukan. Terakhir, adalah informasi perhitungan nilai prbabilitas setiap kategori yang ditampilkan berdasarkan nilai tertinggi hingga terendah dari probabilitas yang diperoleh. Urutan nilai probabilitas yang tertinggi hingga terendah ditampilkan dari kiri ke kanan.

C. Hasil Pengujian Fungsionalitas

Berdasarkan data *training* yang diperoleh, model *classifier* dibangun berdasarkan dua jenis data yaitu data berlabel dan data tidak berlabel. Proses *training* dilakukan pada setiap percobaan. Model pada setiap data kombinasi yang dibangun pada proses training akan digunakan untuk melakukan klasifikasi pada website. Sehingga, *website* dapat menampilkan menu prediksi yang mana menu tersebut digunakan untuk melakukan klasifikasi pada teks dokumen medis. Berdasarkan website yang dibangun maka dilakukan uji fungsionalitas sistem.

| Skenario Pengujian | Hasil |
|---|---|
| Menampilkan halaman home | Sukses |
| Menampilkan tabel hasil pengujian validitas | Sukses |
| Menampilkan grafik hasil pengujian validitas | Sukses |
| Menampilkan halaman predict | Sukses |
| Dapat memilih model classifier pada menu dropdown | Sukses |
| Menambahkan teks dokumen ke dalam text box | Sukses |
| Melakukan prediksi menggunakan model 400:300 | Sukses |
| Melakukan prediksi menggunakan model 300:400 | Sukses |
| Melakukan prediksi menggunakan model 200:500 | Sukses |
| Melakukan prediksi menggunakan model 100:600 | Sukses |
| Menampilkan hasil prediksi | Sukses |
| Menampilkan probabilitas hasil prediksi | Sukses |
| Menampilkan probabilitas masing-masing kategori | Sukses |

## KESIMPULAN

Penelitian ini menggunakan dua jenis data yang berbeda yaitu data berlabel dan data tidak berlabel, cara yang dilakukan untuk memanfaatkan dua jenis data tersebut adalah teknik Semi-Supervised Learning. Pada data berlabel klasifikasi dilakukan dengan menggunakan algoritma Multinomial Naïve Bayes, dan memanfaatkan metode Pseudo Labeling pada data tidak berlabel. Pemanfaatan dua metode tersebut untuk melakukan klasifikasi menggunakan teknik Semi-Supervised

Learning berhasil digunakan pada penelitian ini. Ketersediaan data berlabel dan tidak berlabel dengan jumlah yang terbatas dapat mempengaruhi nilai akurasi, untuk itu penelitian ini dilakukan dengan membagi percobaan menggunakan porsi data yang berbeda karena dapat menunjukan nilai akurasi terbaik disetiap porsinya. Setelah model kombinasi dibangun, hasil akurasi yang diperoleh model 400:300 adalah 82%, model 300:400 adalah 78%, model 200:500 adalah 76%, dan model 100:600 adalah 80%.

## REFERENCES


[1] Hidayatullah, A. F., & Ma'arif, M. R. (2017). Pre-processing Tasks in Indonesian Twitter Messages Pre-processing Tasks in Indonesian Twitter Messages. https://doi.org/10.1088/1742-6596/755/1/011001

[2] Kim, J., & Shin, H. (2013). Breast cancer survivability prediction using labeled , unlabeled , and pseudo-labeled patient data. 613–618. https://doi.org/10.1136/amiajnl-2012-001570

[3] Lee, D. (2013). Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks.

[4] Qiu, Z., Cho, E., Ma, X., & Campbell, W. M. (2019). Graph-Based Semi-Supervised Learning for Natural Language Understanding. 151–158.

[5] Samodra, J., Sumpeno, S., & Hariadi, M. (2009). Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan Naïve Bayes. Seminar, 1–4.