BAB II
LANDASAN TEORI

Pada bab ini dijelaskan landasan teori dan metode yang digunakan pada tugas akhir ini dalam pengklasifikasian dokumen teks. Pembahasan dimulai dengan penjelasan mengenai penelitian sebelumnya yang sudah dilakukan yang berkaitan dengan klasifikasi dokumen teks dan metode untuk klasfikasi dokumen teks. Pada subbab berikutnya dijelaskan metode- metode yang digunakan dalam melakukan klasifikasi dokumen teks.

2.1 Penelitian Sebelumnya

Tinjauan pustaka membahas tentang beberapa penelitian sebelumnya untuk mendukung penelitian ini.

a. Penelitian terkait penggunaan teknik *Semi-Supervised Learning*

Pada penelitian yang berjudul berjudul Implementasi *Semi-Supervised Learning* Pada *Personalized Asthma Management System*, dikembangkan aplikasi manajemen penyakit asma yang bersifat personal untuk masing-masing penderita, dengan mengadopsi teknik *Semi Supervised Learning*. Data dan infromasi aktivitas harian penderita akan direkam oleh system, kemudian system akan mencari pola terkait faktor-faktor pemicu dan pemacu asma, serta mengklasifikasikannya berdasarkan pada derajat serangan asma. Pada penelitian ini dibangun web-based sistem manajemen penyakit asma. Sistem dibuat dengan menggunakan bahasa pemrograman PHP. Sistem ini dapat melakukan pencatatan, pencatatan ini dapat memberikan informasi mengenai manajemen asma yang baik yaitu melalui pencatatan setiap hari faktor pemicu dan faktor pemacu yang dialami oleh penderita, ketika terjadi serangan dan penanganan yang dilakukan. Teknik semi-supervised learning terbukti tepat diterapkan dalam kasus ini, terbukti dengan hasil pengujian sistem dengan tingkat akurasi 80% (Fiarni, Sipayung, Moningka, & Informasi, 2017).

b. Penelitian terkait penggunaan metode *Naïve Bayes Classification*

Penelitian yang berjudul Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan *Naïve Bayes* bertujuan untuk mengklasifikasikan dokumen teks berbahasa Indonesia dengan menggunakan metode Naive Bayes. Uji coba dilakukan dengan menggunakan sampel dokumen teks yang diambil dari sebuah media massa elektronik berbasis web. Manfaat dari penelitian ini adalah untuk mengelola informasi dari kumpulan dokumen

teks yang jumlahnya sangat besar. Data yang digunakan sebagai sampel dalam penelitian ini diambil dari www.tempointeraktif.com edisi tanggal 1 April 2002 sampai dengan 30 Juni 2002. Jumlah dokumen sampel yang digunakan adalah 2.400 dokumen, yang terbagi menjadi 4 kategori yaitu: Nasional, Metro, Nusantara, dan Ekonomi Bisnis. Untuk setiap jenis dokumen, dilakukan uji coba sebanyak 9 kali dengan proporsi dokumen training sebesar 10% sampai dengan 90%. Dari hasil percobaan, terlihat bahwa metode *Naïve Bayes* memiliki nilai akurasi yang tinggi, dan semakin meningkat sesuai dengan peningkatan porsi dokumen training Bahkan akurasi dari metode ini masih tetap tinggi yaitu sebesar 78,21%, walaupun menggunakan porsi dokumen training yang sangat kecil, yaitu sebesar 10% dari keseluruhan dokumen yang diproses (Samodra, Sumpeno, & Hariadi, 2009).

c. Penelitian terkait penggunan *Pseudo Labeling*

Penelitian dengan judul ***Breast Cancer Survivability Prediction Using Labeled, Unlabeled, and Pseudo-Labeled Patient Data*** menggunakan *labeled, unlabeled, dan pseudo label* dari data pasien untuk melakukan prediksi terhadap kelangsungan hidup pasien kanker payudara. Peneliti memaparkan bahwa kesulitan dalam pengumpulan data pasien berlabel menyebabkan peneliti untuk mempertimbangkan *semi-supervised learning* (SSL), mesin terbaru algoritma pembelajaran, karena juga mampu memanfaatkan data pasien yang tidak berlabel, yang relatif lebih mudah mengumpulkan. Untuk dapat menggunakan data pasien berlabel yang sedikit, peneliti mempertimbangkan konsep menandai label *virtual* ke data pasien yang tidak berlabel, yaitu dengan menggunakan teknik *Pseudo-label* dan perlakukan label-label itu seolah-olah benar berlabel. Hasil dari penelitian menggunakan basis data pasien kanker payudara ini mendapatkan akurasi sebesar 76% (Kim & Shin, 2013).

d. Penelitian terkait teknik *Semi-Supervised Learning* menggunakan *Pseudo Labeling*

Penelitian yang mengusulkan *semi-supervised learning* yang sederhana dan efisien untuk *deep neural network* memiliki judul *Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks*. Pada dasarnya, *neural network* yang diusulkan di *training* secara *supervised learning* dengan data berlabel dan tidak berlabel secara bersamaan. Untuk data yang tidak berlabel digunakan teknik *Pseudo-Labeling*. *Pseudo Labeling* hanya mengambil kelas yang memiliki probabilitas prediksi maksimum, digunakan seolah-olah prediksi tersebut adalah label yang benar. Peneliti menggunakan jaringan saraf dengan 1 lapisan tersembunyi. *Rectified Linear Unit* digunakan untuk unit tersembunyi, unit

sigmoid digunakan untuk unit *output*. Jumlah unit tersembunyi adalah 5000. Hasil dari penelitian ini adalah teknik *semi-supervised learning* dengan *pseudo labeling* mengungguli metode konvensional untuk data berlabel kecil meskipun sederhana. Namun, skema pelatihan kurang kompleks dibandingkan *Manifold Tangent Classifier* dan tidak menggunakan matriks kemiripan secara komputasional antara sampel yang digunakan dalam *Semi-Supervised Learning* (Lee, 2013).

Berdasarkan tinjauan pustaka yang telah dijabarkan sebagai referensi penelitian ini, bahwa belum ada sistem yang melakukan penelitian untuk klasifikasi dokumen medis menggunakan teknik semi-supervised learning dengan metode naïve bayes dan expectation maximization secara bersamaan.

Tabel 2. 1 Kesimpulan dari penelitian sebelumnya

No.	Judul Penelitian	Metode	Hasil Penelitian
1.	Implementasi *Semi-Supervised Learning* Pada *Personalized Asthma Management System* (Fiarni et al., 2017)	*Semi-Supervised Learning* untuk mencari hubungan antara terjadinya serangan asma dengan aktivitas dan kondisi pasien.	Penelitian ini bertujuan agar sistem ini dapat melakukan pencatatan, pencatatan ini dapat memberikan informasi mengenai manajemen asma yang baik yaitu melalui pencatatan setiap hari faktor pemicu dan faktor pemacu yang dialami oleh penderita, ketika terjadi serangan dan penanganan yang dilakukan.

2.	Klasifikasi Dokumen Teks Berbahasa Indonesia dengan Menggunakan *Naïve Bayes* (Samodra et al., 2009)	*Naïve Bayes* untuk klasifikasi dokumen teks berita dengan membagi kategori Nasional, Metro, Nusantara, dan Ekonomi Bisnis.	Penelitian ini bertujuan untuk mengklasifikasikan dokumen teks berbahasa Indonesia dengan menggunakan metode Naive Bayes. Uji coba dilakukan dengan menggunakan sampel dokumen teks yang diambil dari sebuah media massa elektronik berbasis web.
3.	*Breast Cancer Survivability Prediction Using Labeled, Unlabeled, And Pseudo-Labeled Patient Data* (Kim & Shin, 2013)	*Pseudo-labeling* untuk melakukan prediksi penyakit kanker payudara berdasarkan data pasien berlabel dan tidak berlabel.	Penelitian ini menggunakan *labeled, unlabeled, dan pseudo label* dari data pasien untuk melakukan prediksi terhadap kelangsungan hidup pasien kanker payudara. Untuk dapat memanfaatkan *labeled data* dan *unlabeled data* peneliti menggunakan teknik *semi-supervised learning*. Agar dapat

			menggunakan data pasien berlabel yang sedikit, peneliti mempertimbangkan konsep menandai label *virtual* ke data pasien yang tidak berlabel, yaitu dengan menggunakan teknik *Pseudo-label*.
4.	*Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks* (Lee, 2013)	*Semi-Supervised Learning*, *Pseudo-labeling* diterapkan untuk *deep neural network* yang menggunakan data berlabel dan tidak berlabel secara bersamaan.	Penelitian ini mengusulkan *semi-supervised learning* yang sederhana dan efisien untuk *deep neural network*. Untuk data yang tidak berlabel digunakan teknik *Pseudo-Labeling*. *Pseudo Labeling* hanya mengambil kelas yang memiliki probabilitas prediksi maksimum, digunakan seolah-olah prediksi tersebut adalah label yang benar. Hasil dari penelitian ini adalah teknik *semi-supervised learning*

			dengan *pseudo labeling* mengungguli metode konvensional untuk data berlabel kecil meskipun sederhana.

Penelitian yang akan dilakukan adalah menerapkan metode yang digunakan pada penelitian di atas yaitu teknik *Semi-Supervised Learning* dengan memanfaatkan metode *Multionomial Naïve Bayes Classification* dan *Pseudo Labeling* untuk melakukan klasifikasi dokumen medis. Hasil akhir dari penelitian ini adalah mengetahui apakah dengan menggunakan teknik *semi-supervised learning* proses klasifikasi dokumen medis yang dilakukan efektif atau tidak, dari perbandingan jumlah dokumen training akan diketahui berapa akurasi yang diperoleh. Adapun pada hasil akhir nanti akan diketahui masih efektif atau tidak jumlah data training yang hanya menggunakan data yang berlabel sedikit dengan perbandingan jumlah data *training* yang tidak berlabel lebih banyak.

2.2 Tinjauan Pustaka

2.2.1 Klasifikasi Dokumen Teks

Klasifikasi atau kategorisasi teks merupakan suatu proses penempatan suatu dokumen ke suatu kategori atau kelas sesuai dengan karakteristik dari dokumen tersebut. Dalam text mining, klasifikasi mengacu kepada aktifitas menganalisis atau mempelajari himpunan dokumen teks *pre-classified* untuk memperoleh suatu model atau fungsi yang dapat digunakan untuk mengelompokkan dokumen teks lain yang belum diketahui kelasnya ke dalam satu atau lebih kelas kelas *pre-defined* tersebut (Darujati et al., 2012).

Klasifikasi dokumen teks yang dilakukan pada penelitian ini adalah klasifikasi yang diimplementasikan pada dokumen medis. Klasifikasi dokumen medis adalah masalah sederhana namun sangat penting karena manfaatnya cukup besar mengingat jumlah dokumen yang ada setiap hari semakin bertambah. Manfaat dari klasifikasi dokumen adalah untuk pengorganisasian dokumen. Dengan jumlah dokumen yang sangat besar, untuk mencari sebuah dokumen akan lebih mudah apabila kumpulan dokumen yang dimiliki terorganisir dan telah dikelompokkan sesuai kategorinya masing-masing.

Contoh aplikasi penggunaan klasifikasi dokumen teks yang banyak digunakan adalah *e-mail spam filtering*. Pada aplikasi *spam filtering* sebuah *e-mail* diklasifikasikan apakah *e-mail* tersebut termasuk *spam* atau tidak dengan memperhatikan kata-kata yang terdapat di dalam *e-mail* tersebut. Sebuah dokumen dapat dikelompokkan ke dalam kategori tertentu 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. Perhatikan beberapa kalimat berikut ini:

1. Setiap minggu pertama bulan Agustus selalui diperingati Pekan ASI Sedunia. Pemberian ASI eksklusif selama 6 bulan dan ASI lanjutan secara optimal hingga 2 tahun atau lebih merupakan hal mutlak untuk meningkatan kesehatan bayi.
2. Penyakit jantung koroner terdiri dari penyakit jantung koroner stabil tanpa gejala, angina pektoris stabil, dan sindrom koroner akut. Penyakit jantung koroner stabil tanpa gejala biasanya diketahui dari skrining, sedangkan angina pektoris stabil didapatkan gejala nyeri dada bila melakukan aktivitas yang melebihi aktivitas sehari-hari.
3. Perawatan berlebih terhadap diabetes, terutama mereka yang menderita diabetes tipe 1, mungkin memiliki risiko hipoglikemia (gula darah rendah) yang meningkat jika mereka menerima terlalu banyak terapi penurun glukosa. Penelitian baru sekarang memperingatkan bahwa banyak orang dengan diabetes menghadapi risiko itu.

Pada kalimat (1) terdapat kata ASI, eksklusif, dan bayi. Kata-kata tersebut memiliki keterkaitan erat dengan masalah bayi, sehingga dapat disimpulkan bahwa kalimat (1) membahas tentang bayi. Kalimat (2) memiliki kata jantung, koroner, dan dada. Dari kata-kata tersebut akan muncul dugaan bahwa kalimat (2) sedang membahas masalah jantung. Terakhir, pada kalimat (3) terdapat kata diabetes, glukosa, dan hipoglikemia yang menunjukkan bahwa kalimat tersebut membahas tentang diabetes.

Kata ASI yang terdapat pada dokumen lain belum dapat dijadikan sebagai acuan bahwa dokumen lain tersebut membahas mengenai bayi. Apabila dokumen lain tersebut memiliki kata-kata lain yang mengarahkan kepada pembahasan bayi secara bersamaan, maka dapat disimpulkan bahwa dokumen tersebut membahas mengenai bayi. Untuk dapat menentukan kategori dari sebuah dokumen haruslah dilihat semua kata-kata yang terkait pada dokumen tersebut.

2.2.2 Machine Learning untuk Klasifikasi Dokumen Teks

Teknik *machine learning* mulai banyak digunakan untuk melakukan klasifikasi dokumen teks pada awal tahun 1990-an. Teknik ini 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. Teknik *clustering* umumnya merupakan teknik yang iteratif. Kategori-kategori yang ada untuk setiap dokumen biasanya belum diketahui secara eksplisit. Hal ini berbeda dengan teknik klasifikasi dimana kategori yang ada telah ditentukan sebelumnya.

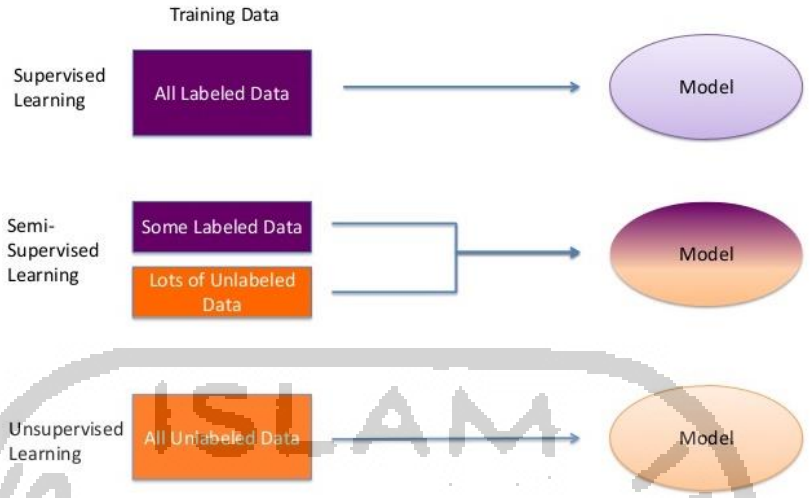
Pendekatan kedua adalah *supervised learning*. Pendekatan ini dilakukan dengan membangun sebuah *classifier* dari proses pembelajaran mengenai ciri dari tiap-tiap kategori yang ada. Pendekatan ini biasa disebut dengan teknik klasifikasi. Teknik ini membagi kumpulan dokumen yang dimiliki menjadi dokumen *training* dan dokumen *testing*. *Classifier* dibangun dengan mempelajari ciri tiap kategori berdasarkan dokumen *training* yang dimiliki. 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. Pada penelitian ini digunakan metode *Multinomial Naïve Bayes* untuk melakukan klasifikasi dokumen medis dan dalam proses klasifikasi untuk *unlabeled data* digunakan teknik *Pseduo Labeling*.

Metode *machine learning* dapat dipergunakan untuk klasifikasi dokumen teks. Hal ini ditunjukkan dengan penelitian yang telah dilakukan sebelumnya oleh Kim & Shin (2013). Penelitian tesebut melakukan prediksi terhadap *labeled data* (data berlabel) dan *unlabeled data* (data tidak berlabel) untuk mengetahui kelangsungan hidup pasien kanker payudara. Untuk melakukan prediksi terhadap dua jenis data yang berbeda maka penelitian ini menggunakan teknik semi-supervised learning agar dapat melakukan prediksi pada setiap data dengan jenis berbeda. Semi-supervissed learning mampu menggunakan data pasien yang tidak berlabel, tetapi akurasi prediksi tenik semi-supervised learning meningkat dengan jumlah data pasien yang berlabel, seperti kebanyakan algoritma dalam machine learning. Untuk mengatasi kesulitan dalam mendapatkan data pasien berlabel, dimungkinkan untuk mendapatkan lebih

banyak data berlabel dengan cara menghasilkan label data dari data yang tidak berlabel dan menganggapnya seolah-olah data tidak berlabel tersebut memiliki label. Hal ini dapat dilakukan dengan menerapkan teknik pseudo labeling. Perhatikan bahwa data pasien yang berlabel dan tidak berlabel diperoleh langsung dari dataset yang diberikan, sedangkan data pseudo-labeling dihasilkan secara buatan oleh model yang diusulkan dalam penelitian ini. Dari hasil yang diperoleh menggunakan pseudo-label pada data pasien, baik
data pasien berlabel dan tidak berlabel, akan meningkatkan teknis kualitas prognosis ketahanan kanker yang diharapkan dapat mengarah pada pengobatan yang lebih baik untuk pasien kanker. Penelitian ini membandingan penggunaan data pasien dengan melakukan prediksi menggunakan metode yang berbeda. Akurasi yang diperoleh dari proses pseudo-label dari teknik semi-supervised learning mencapai nilai sebesar 76%. Nilai akurasi yang diperoleh paling besar jika dibandingkan dengan metode lainnya.

2.2.3 Teknik *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 et al., 2019). Teknik *semi-supervised learning* menggunakan data yang tidak berlabel untuk mendapatkan lebih banyak pemahaman tentang struktur populasi secara umum. Namun bagaimana cara teknik *semi-supervised learning* dapat melakukan pembelajaran dengan hanya menggunakan data *training* berlabel dalam jumlah yang sedikit sedangkan data *training* tidak berlabel dalam jumlah yang besar. Jadi, untuk menyelesaikan masalah diatas, *semi-supervised learning* dapat memanfaatkan kedua jenis data sebagai data training baik data berlabel dan data tidak berlabel seperti yang ditunjukan pada Gambar 2.1.



Gambar 2. 1 Teknik *semi-supervised learning*

2.2.4 *Term Frequency-Inverse Document Frequency* (TF-IDF)

Term weighting atau pembobotan kata bertujuan untuk memberikan bobot nilai pada setiap kata. Perhitungan bobot ini memerlukan dua hal, yaitu *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF).

Term Frequency merupakan banyaknya jumlah kata atau *term* tertentu yang ada dalam suatu dokumen. Sementara *Inverse Document Frequency* adalah frekuensi kemunculan kata atau *term* pada seluruh dokumen. Nilai IDF berbanding terbalik dengan jumlah dokumen yang mengandung *term* tertentu. *Term* yang jarang muncul pada seluruh dokumen memiliki nilai IDF yang lebih besar dari nilai IDF *term* yang sering muncul. Jika pada setiap dokumen mengandung *term* tertentu, maka nilai IDF term tersebut bernilai 0. Hal ini menunjukan bahwa term yang muncul pada seluruh dokumen merupakan term yang tidak berguna untuk membedakan dokumen berdasarkan topik tertentu (Rahman, 2017). Rumus TF-IDF adalah seperti pada Persamaan (2.1).

$$W_{d,t} = tf_{d,t}\, x\, idf_t = tf_{d,t}\, x\, \log(\frac{N}{df_t}) \qquad (2.1)$$

Keterangan:

$W_{d,t}$ = Bobot term ke-t terhadap dokumen d

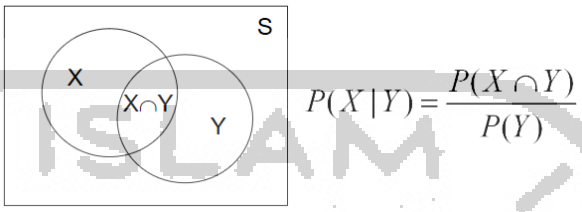
$tf_{d,t}$ = Jumlah kemunculan term t dalam dokumen d

N = Jumlah dokumen secara keseluruhan

df_t = Jumlah dokumen yang mengandung term t

2.2.5 *Multinomial Naïve Bayes*

Multinomial Naive Bayes adalah pengembangan dari teorama bayes yang merupakan teorema peluang bersyarat yaitu perhitungan bepeluang X bila diketahui adanya kejadian Y atau P(X|Y) (Basuki, 2006) dan ditunjukan pada Gambar 2.2.



Gambar 2. 2 Probabilitas bersyarat (Basuki & Ahmad 2006)

Metode *multinomial Naive Bayes* adalah salah satu metode yang banyak digunakan untuk klasifikasi. Dalam hal ini, metode *Multinomial Naive Bayes* digunakan untuk klasifikasi teks. Metode klasifikasi ini merupakan metode dengan pembelajaran terbimbing *(supervised learning)*. Pada metode ini dibutuhkan data training sebagai basis pengetahuan untuk model yang dibuat.

Multinomial Naive Bayes Classifier merupakan metode klasifikasi turunan dari teorema Bayes yang mana kelas dokumen tidak hanya ditentukan oleh kata tapi juga jumlah kemunculannya (I. H. Witten & Hall, 2011). Untuk memperhitungkan kelas dari dokumen maka dapat dilihat pada Persamaan (2.2) yang merupakan rumus untuk menghitung probabilitas suatu dokumen masuk ke dalam suatu kelas.

$$P(c|term\ dokumen\ d) = P(c)x\ P(t_1|c)x\ P(t_2|c)x\ P(t_3|c)x\ \dots\ x\ P(t_n|c) \quad (2.2)$$

Keterangan :

P(c)	= Probabilitas *prior* dari kelas c
t_n	= Kata dokumen d ke-n
$P(c\|term\ dokumen\ d)$	= Probabilitas suatu dokumen masuk ke kelas *c*
$P(t_n\|c)$	= Probabilitas kata ke-n dengan diketahui kelas *c*

Menghitung data prior masing-masing kelas dengan menggunakan rumus pada Persamaan (2.3) :

$$P(c) = \frac{N_c}{N} \qquad (2.3)$$

Keterangan :

N_c = Jumlah kelas c pada seluruh dokumen

N = Jumlah seluruh dokumen

Sementara rumus Multinomial yang digunakan dengan pembobotan TF-IDF dapat dilihat pada Persamaan (2.4). Persamaan (2.4) merupakan rumus untuk menghitung probabilitas kata ke-n data dokumen medis.

$$P(t_n \mid c) = \frac{W_{ct}+1}{(\Sigma w' \epsilon v w' ct)+B'} \qquad (2.4)$$

Keterangan :

$W_{ct} + 1$ = Nilai pembobotan tf-idf atau W dari *term* t di kategori c

$\Sigma w' \epsilon v w' ct$ = Jumlah total W dari keseluruhan *term* yang berada di kategori c.

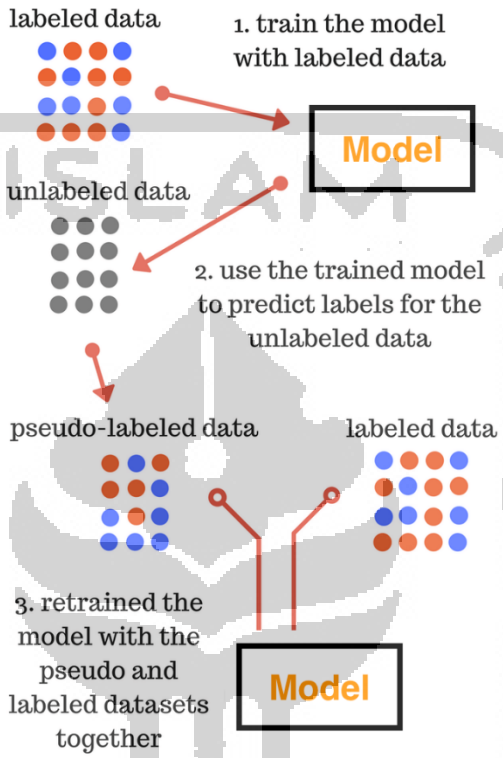
B' = Jumlah W kata unik (nilai idf tidak dikali dengan tf) pada seluruh dokumen.

Nilai probabilitas inilah yang digunakan untuk mneghitung kemungkinan label *category* dalam dokumen medis.

2.2.6 *Pseudo Labeling*

Untuk melakukan *training* sebuah *machine learning model* dengan menggunakan *supervised learning*, data harus memiliki label. Karna itu penggunaan data tidak berlabel tidak digunakan untuk *supervised learning* seperti klasifikasi dan regresi. Padahal data tidak berlabel bisa dikatakan memiliki jumlah yang sangat besar jika dibandingkan dengan data berlabel. Namun saat ini, pembelajaran *machine learning* dapat kita gunakan untuk melakukan klasifikasi menggunakan data berlabel dan data tidak berlabel secara bersamaan. Pembelajaran yang digunakan adalah *semi-supervised learning*. Selain menggunakan data tambahan untuk tujuan analitik, teknik *semi-supervised learning* bahkan dapat digunakan untuk membantu melatih model dengan menggabungkan data yang tidak berlabel dan berlabel untuk *training*

model. Mengimplementasikan *teknik semi-supervised learning* dapat dilakukan dengan menggunakan metode yang dikenal dengan *pseudo-labeling*. Langkah-langkah untuk menerapkan teknik *pseudo-label* ditunjukan pada gambar 2.3.



Gambar 2. 3 Langkah *pseudo labeling*

Secara sederhana, langkah awal *pseudo-label* adalah memanfaatkan data berlabel untuk melakukan *training data*. Kemudian, model yang dibangun menggunakan data berlabel dimanfaatkan untuk memprediksi data tidak berlabel sehingga menciptakan *pseudo-label data*. Langkah selanjutnya adalah menggabungkan data berlabel dan *pseudo-label data* yang dihasilkan sebagai satu *dataset* baru. *Dataset* baru tersebut kemudian digunakan untuk melakukan *training* dan menciptakan model baru yang siap digunakan untuk melakukan klasifikasi dokumen medis.