KLASIFIKASI INFORMASI WISATA KULINER INDONESIA DARI MEDIA SOSIAL *TWITTER* MENGGUNAKAN *NAIVE BAYES CLASSIFIER*

SKRIPSI

Oleh :

Dwi Hasifah

NIM 162410102022

PROGRAM STUDI TEKNOLOGI INFORMASI

FAKULTAS ILMU KOMPUTER

UNIVERSITAS JEMBER

2020

KLASIFIKASI INFORMASI WISATA KULINER INDONESIA DARI MEDIA SOSIAL *TWITTER* MENGGUNAKAN *NAIVE BAYES CLASSIFIER*

SKRIPSI

Diajukan guna melengkapi tugas akhir dan memenuhi salah satu syarat untuk menyelesaikan pendidikan Sarjana (S1) Program Studi Teknologi Informasi Fakultas Ilmu Komputer Universitas Jember dan mencapai gelar Sarjana Komputer

Oleh :

Dwi Hasifah

NIM 162410102022

PROGRAM STUDI TEKNOLOGI INFORMASI

FAKULTAS ILMU KOMPUTER

UNIVERSITAS JEMBER

2020

PERSEMBAHAN

Skripsi ini saya persembahkan untuk :

1. Allah SWT yang senantiasa memberikan rahmad dan hidayah-Nya untuk mempermudah dan memperlancar dalam mengerjakan skripsi;
2. Ibunda Siti Aminah tercinta dan Ayahanda Suroto tercinta;
3. Saudara kandung Indah Mutmainah yang selalu memberi semangat;
4. Nenek Hj. Siti Khodija;
5. Teman-teman seperjuangan Program Studi Teknologi Informasi Fakultas Ilmu Komputer Universitas Jember angkatan 2016;
6. Guru-guru dan tenaga pengajar saya sejak taman kanak-kanak hingga perguruan tinggi;
7. Almamater Program Studi Teknologi Informasi Fakultas Ilmu Komputer Universitas Jember.

MOTTO

"Jangan menanti hingga esok apa yang mampu kita kerjakan sekarang ! Sekarang berjuang besok raih kemenangan, Yakinlah ! karena selama ada keyakinan semua akan menjadi mungkin".

-Dwi Hasifah-

"Ciptakan konsekuensi jika tidak bisa menyelesaikan tantanganmu !"

-Dee Lestari-

PERNYATAAN

Saya yang bertanda tangan di bawah ini:

Nama : Dwi Hasifah

NIM : 162410102022

Menyatakan dengan sesungguhnya bahwa karya ilmiah yang berjudul "Klasifikasi Informasi Wisata Kuliner Indonesia dari Media Sosial *Twitter* menggunakan *Naive Bayes Classifier",* adalah benar-benar hasil karya saya sendiri, kecuali jika dalam pengutipan substansi disebutkan sumbernya, belum pernah diajukan pada instansi manapun, dan bukti karya jiplakan. Saya bertanggung jawab atas keabsahan dan kebenaran isinya sesuai dengan sikap ilmiah yang harus dijunjung tinggi.

Demikian pernyataan ini saya buat dengan sebenarnya, tanpa adanya tekanan dan paksaan dari pihak manapun serta bersedia mendapat sanksi akademik jika dikemudian hari pernyataan ini tidak benar.

Jember, 24 Januari 2020

Yang menyatakan,

Dwi Hasifah

NIM 162410102022

SKRIPSI

KLASIFIKASI INFORMASI WISATA KULINER INDONESIA DARI MEDIA SOSIAL *TWITTER* MENGGUNAKAN *NAIVE BAYES CLASSIFIER*

Oleh :

Dwi Hasifah

NIM 162410102022

Pembimbing :

Dosen Pembimbing Utama : Achmad Maududie,ST.,M.Sc.
NIP 197004221995121001

Dosen Pembimbing Pendamping : Tio Dharmawan,S.Kom.,M.Kom.
NIP 760016851

PENGESAHAN PEMBIMBING

Skripsi berjudul "Klasifikasi Informasi Wisata Kuliner Indonesia dari Media Sosial *Twitter* menggunakan *Naive Bayes Classifier"* telah diuji dan disahkan pada:

hari, tanggal : Jumat, 24 Januari 2020
tempat : Program Studi Teknologi Informasi Fakultas Ilmu Komputer Universitas Jember

Disetujui oleh:

Pembimbing I,	Pembimbing II,
Achmad Maududie,ST.,M.Sc.	Tio Dharmawan,S.Kom.,M.Kom.
NIP 199700422199512100 1	NIP 760016851

PENGESAHAN PENGUJI

Skripsi berjudul "Klasifikasi Informasi Wisata Kuliner Indonesia Dari Media Sosial *Twitter* Menggunakan *Naive Bayes Classifier"*, telah diuji dan disahkan pada:

hari, tanggal : Jumat, 24 Januari 2020
tempat : Program Studi Teknologi Informasi Fakultas Ilmu Komputer Universitas Jember

Disetujui oleh :

Penguji I,	Penguji II,
Prof.Dr.Saiful Bukhori,ST.,M.Kom.	Gama Wisnu Fajarianto,S.Kom.,M.Kom.
NIP 196811131994121001	NIP 760015717

Mengesahkan
Dekan Fakultas Ilmu Komputer,

Prof. Dr. Saiful Bukhori, ST., M.Kom.
NIP 196811131994121001

RINGKASAN

Klasifikasi Informasi Lokasi Kuliner Indonesia Dari Media Sosial *Twitter* Menggunakan *Naive Bayes Classifier*; Dwi Hasifah, 162410102022, 2020; 78 halaman, Program Studi Teknologi Informasi Fakultas Ilmu Komputer Universitas Jember.

Klasifikasi informasi daftar wisata kuliner Indonesia merupakan proses memahami, mengklasifikasi, dan mengolah data tekstual secara otomatis untuk mendapat informasi daftar wisata kuliner Indonesia. Konten *tweet* sangat beragam dan salah satu pembahasan yang cukup populer saat ini yaitu mengenai wisata kuliner. Keberagaman konten *tweet* mengenai kuliner tidak hanya *tweet* yang terkait dengan wisata kuliner, namun juga *tweet* tentang cara pembuatan, rasa, bahan, harga hingga bentuk makanan yang sebenarnya tidak terkait dengan informasi wisata kuliner.

Pendekatan *text mining* menjadi alternatif terbaik untuk mengartikan makna dari setiap *tweet. Text mining* merupakan proses mengeksplorasi dan menganalisis sejumlah besar data teks tidak terstruktur yang dapat mengidentifikasi konsep, pola, topik, kata kunci, dan atribut lainnya dalam sebuah data. Salah satu metode yang sering digunakan dalam pengelompokan informasi berbasis teks adalah *Naive Bayes Classifier*. *Naive Bayes Classifier* merupakan salah satu metode yang banyak digunakan berdasarkan probabilitas *P* atribut *x* dari setiap kelas *y* data yang didasarkan pada asumsi *naif* atau *independen* yang kuat. Pengembangan pendekatan klasifikasi daftar wisata kuliner berbasis metode *Naïve Bayes Classifier* terhadap 5000 dataset yaitu 80% data *training* untuk membangun model dan 20% data *testing* untuk menguji model terhadap 10 kelas klasifikasi yaitu: soto, gudeg, mie, sate, rujak, pempek, rendang, pecel, kuliner lain, dan bukan kuliner menghasilkan nilai uji akurasi sebesar 86.5%.

PRAKATA

Puji syukur kehadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan skripsi dengan judul "Klasifikasi Informasi Lokasi Kuliner Indonesia Dari Media Sosial *Twitter* Menggunakan *Naive Bayes Classifier"*. Skripsi ini disusun untuk memenuhi salah satu syarat menyelesaikan pendidikan Strata Satu (S1) pada Program Studi Teknologi Informasi Fakultas Ilmu Komputer Universitas Jember.

Penyusunan skripsi ini tidak lepas dari bantuan berbagai pihak. Oleh karena itu, penulis menyampaikan terima kasih kepada:

1. Allah SWT yangs senantiasa memberikan rahmad dan hidayah-Nya untuk mempermudah dan melancarkan dalam mengerjakan skripasi;
2. Muhammad Arief Hidayat,S.Kom.,M.Kom selaku Dosen Pembimbing Akademik yang telah membimbing selama penulis menjadi mahasiswa;
3. Achmad Maududie,ST,M.Sc. selaku Dosen Pembimbing Utama dan Tio Dharmawan,S.Kom.,M.Kom. selaku Dosen Pembimbing Pendamping yang telah meluangkan waktu, pikiran, dan perhatian dalam penulisan skripsi;
4. Prof.Dr.Saiful Bukhori,ST.,M.Kom. selaku Dosen Pembahas I dan Gama Wisnu Fajarianto,S.Kom.,M.Kom selaku Dosen Pembahas II yang telah berkenan untuk menguji skripsi ini dan memberikan masukan serta saran untuk pengembangan diri penulis dan skripsi ini;
5. Seluruh Bapak dan Ibu dosen beserta staff karyawan di Program Studi Teknologi Informasi Fakultas Ilmu Komputer Universitas Jember;
6. Guru-guru dan tenaga pengajar pendidikan formal maupun informal sejak taman kanak-kanak hingga perguruan tinggi;
7. Ibunda tercinta Siti Aminah dan Ayahanda tercinta Suroto yang selalu mendukung dan mendoakan;
8. Saudara kandung tersayang Indah Mutmainah, saudara ipar Fathul Imami, dan saudara sepupu Agustin Maulidatus Soleha yang selalu memberi semangat;

9. Sahabat terbaik Ega Nur Tantiana, Alfina Apriliani dan Andry Dermawan, Noni Namida Oliviani, Nuril Ilmi Al Islami, Intan Berliana Safitri, Rizky Berlia Oktaviandi, Ratna Syavira Maulida, Muhammad Sukron, Riski Septia Nuhaida, Ely Rahmawati, Rosidatul Hotimah, dan Diyah Ika Pratiwi yang selalu menemani, membantu dan memberi semangat;
10. Keluarga Angkatan 2016 Program Studi Teknologi Informasi Fakultas Ilmu Komputer Universitas Jember (FIGORA);
11. Kepengurusan organisasi Himpunan Mahasiswa Teknologi Informasi periode 2018/2019;
12. Kepengurusan asisten Laboratorium Basis Data periode 2017/2018
13. Keluarga KKN-200 Gunung Putri gelombang II 2018/2019
14. Teman-teman Kontrakan Danau Toba VI
15. Dan seluruh pihak yang membantu penulis dalam mensukseskan skripsi ini, yang tidak dapat disebutkan secara rinci.

Dengan harapan penelitian ini nantinya terus berlanjut dan berkembang. Penulis menyadari bahwa skripsi ini masih jauh dari sempurna, oleh sebab itu penulis mengharapkan adanya masukan yang bersifat membangun dari semua pihak. Penulis berharap skripsi ini dapat bermanfaat bagi semua pihak

Jember, 24 Januari 2020

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1. PENDAHULUAN

1.1 Latar Belakang

Keanekaragaman kuliner di Indonesia menyimpan potensi yang besar untuk dikembangkan sebagai jasa penunjang dalam pengembangan potensi wisata kuliner. Saat ini, wisata kuliner menjadi semakin populer serta sudah menjadi bagian dari gaya hidup masyarakat. Keberadaan media sosial juga mengantarkan wisata kuliner Indonesia semakin populer dan mendorong masyarakat untuk lebih mengenalnya (Hanifah and Nurhasanah 2018). Salah satu media sosial yang popular saat ini adalah *Twitter*.

Twiter merupakan media sosial berbasis minat untuk berinteraksi, mengikuti, serta mencari informasi (Indraloka et al. 2017). Walaupun *Twitter* memiliki fitur pencarian sesuai dengan kata kunci yang dientrikan, namun fitur tersebut tidak cukup efektif untuk mendapatkan daftar informasi wisata kuliner yang ada. Apabila dientrikan kata kunci berupa nama makanan, maka tidak hanya *tweet* yang terkait dengan wisata kuliner yang didapatkan, namun juga *tweet* tentang cara pembuatan, rasa, bahan, harga, hingga bentuk makanan yang sebenarnya tidak terkait dengan informasi wisata kuliner. Hal ini membutuhkan *text mining* dalam mengolah *tweet* tersebut.

Text mining merupakan proses mengeksplorasi dan menganalisis sejumlah besar data teks tidak terstruktur yang dapat mengidentifikasi konsep, pola, topik, kata kunci, dan atribut lainnya dalam sebuah data (Indraloka et al. 2017). *Text mining* efektif digunakan untuk mencari informasi ataupun mengelompokkan informasi berbasis teks (Ni Luh Ratniasih, Made Sudarma 2017). *Importance of a Term in a Document* (ITD) *and Importance of a Term for expressing Sentiment* (ITS) merupakan suatu algoritma untuk mengevaluasi kemampuan sebuah kata dalam menggambarkan suatu makna terhadap sebuah dokumen (Deng, Luo, and Yu 2014). Algoritma ini menggabungkan dua konsep untuk perhitungan bobot yaitu normalisasi frekuensi kata yang muncul dalam sebuah dokumen dan logaritma frekuensi dokumen yang merepresentasikan terhadap sebuah kelas. Salah satu metode yang sering digunakan dalam pengelompokan informasi berbasis teks adalah *Naive Bayes Classifier*.

Naive Bayes Classifier merupakan salah satu metode yang banyak digunakan berdasarkan probabilitas *P* atribut *x* dari setiap kelas *y* data yang didasarkan pada asumsi *naif* atau *independen* yang kuat (Agus Hermanto 2016). Metode *Naïve Bayes Classifier* menghasilkan nilai yang pasti dan akurasi yang baik karena metode tersebut memperkecil kemungkinan kesalahan pada pengklasifikasian. Hasil dari klasifikasi dokumen menggunakan *Naïve Bayes Classifier* pada penelitian (Pandhu and Agus 2016) dengan data *training* sebanyak 260 dokumen politik dan 222 dokumen ekonomi menggunakan 40 data *testing* menunjukkan nilai akurasi yang baik pada keseluruhan klasifikasi, dengan akurasi keseluruhan klasifikasi sebesar 85%. Penelitian lain mengenai klasifikasi berdasarkan analisis positif, negative, dan netral dalam bahasa Indonesia, Inggris, dan Vietnam (Le et al. 2019) menyatakan metode *Naïve Bayes Classifier* terbukti keakuratannya dalam pengambilan data secara umum dengan menghasilkan akurasi 98,2%.

Mengacu pada ketidaksesuaian fitur pencarian pada *Twitter* untuk mendapatkan daftar wisata kuliner, dan kemampuan algoritma *Importance of a Term in a Document* (ITD) *and Importance of a Term for expressing Sentiment* (ITS) dalam mengevaluasi sebuah kata dalam menggambarkan suatu makna terhadap sebuah dokumen serta keuntungan metode *Naïve Bayes Classifier* untuk mengelompokkan dokumen teks, maka penulis mengembangkan pendekatan klasifikasi daftar wisata kuliner berbasis metode *Naïve Bayes Classifier* menggunakan algoritma *Importance of a Term in a Document* (ITD) *and Importance of a Term for expressing Sentiment* (ITS). Pendekatan ini didasarkan pada proses klasifikasi data *tweet* yang akan dijadikan sebagai model dalam klasifikasi daftar wisata kuliner Indonesia. Sehingga dengan adanya sistem ini diharapkan dapat mengklasifikasi daftar wisata kuliner dari *Twitter* dan memudahkan pencarian informasi wisata kuliner pada media sosial *Twitter*.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang, permasalahan yang harus diselesaikan dalam penelitian ini yaitu

a. bagaimana cara mengklasifikasi postingan *tweet* yang memiliki informasi wisata kuliner Indonesia

b. berapa tingkat akurasi hasil implementasi metode *Naïve Bayes Classifier* dalam mengklasifikasi informasi wisata kuliner Indonesia mengunakan algoritma *Importance of a Term in a Document* (ITD) *and Importance of a Term for expressing Sentiment* (ITS)

1.3 Tujuan dan Manfaat Penelitian

1.3.1 Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini yaitu

a. menemukan informasi daftar wisata kuliner Indonesia pada *Twitter*
b. mengetahui tingkat akurasi hasil implementasi metode *Naïve Bayes Classifier* dalam mengklasifikasi informasi wisata kuliner Indonesia mengunakan algoritma *Importance of a Term in a Document* (ITD) *and Importance of a Term for expressing Sentiment* (ITS)

1.3.2 Manfaat Penelitian

Manfaat dari tercapainya penelitian ini yaitu :

a. informasi pada postingan *tweet* mengenai daftar wisata kuliner Indonesia lebih tertata
b. pencarian informasi pada postingan *tweet* mengenai daftar wisata kuliner Indonesia lebih mudah
c. mengetahui tingkat akurasi hasil implementasi metode *Naive Bayes Classifier* dalam klasifikasi postingan *tweet* pada *Twitter*

1.4 Batasan Masalah

Penulis memberikan batasan masalah untuk objek dan tema yang dibahas supaya tidak terjadi penyimpangan dalam proses penelitian. Batasan masalah dalam penelitian ini antara lain :

a. Sumber data yang dimanfaatkan yaitu *Twitter API* dengan menggunakan *API key* penulis
b. Metode yang digunakan yaitu *Naive Bayes Classifier* tanpa membandingkan dengan metode lain
c. Sistem berfokus pada 10 kelas klasifikasi yaitu soto, gudeg, mie, sate, rujak, pempek, rendang, pecel, kuliner lain, dan bukan kuliner
d. Klasifikasi kuliner Indonesia berfokus pada postingan *tweet* yang

mengandung informasi wisata kuliner

e. Dataset yang digunakan yaitu 5.000 postingan *tweet* terakhir atau terbaru
f. Pengujian sistem harus sesuai dengan daftar kata pada data *training*, jika tidak sesuai harus melakukan penambahan data *training* terlebih dahulu.
g. Dalam menjalankan sistem apabila *API Twitter down* maka proses pengumpulan data juga terhenti.

BAB 2. TINJAUAN PUSTAKA

2.1 *Twitter Aplication Programming Interface (API)*

Twitter Aplication Programming Interface (API) merupakan akses programatik ke data *Twitter* kepada perusahaan, pengembang, dan pengguna. *API* merupakan sejumlah fungsi yang dapat digunakan pengembang perangkat lunak untuk mengolah data saat membangun perangkat lunak (JJrgens and Jungherr 2016). Saat menggunakan *Twitter API*, didapatkan beberapa kode berupa *consumer key, consumer secret, access token, dan access key*. Kode tersebut digunakan untuk proses autentikasi ke *Twitter* sehingga dapat mengakses informasi yang ada di *Twitter*. Tahapan untuk mendapatkan kode Twitter API yaitu :

a. Kunjungi laman *website* pengembang *Twitter* pada halaman *Twitter* yang terletak pada pojok kanan bawah
b. Masukkan *e-mail* dan nomor *handphone* yang telah terverifikasi untuk *login*
c. Buatlah aplikasi baru dan pilih alasan mengapa menggunakan *developer tools*
d. Isilah formulir dan setujui ketentuan yang diberikan oleh pihak *Twitter*
e. Verifikasi akun developer melalui *e-mail*
f. Buatlah aplikasi dan isi deskripsi aplikasi
g. Pilih tab *Keys and token (Customers Key* dan *Customers Secret Key* telah didapatkan)
h. Buat akses token dengan memilih tombol *create* (*Access Token Key* dan *Access Token Secret Key* telah didapatkan)

Pemanggilan *Twitter API* dilakukan dengan menggunakan salah satu *library Twitter* dengan menggunakan bahasa pemrograman *Python* yaitu *Tweepy*. *Tweepy* merupakan *library Python* yang dapat mengakses *API* milik *Twitter* sehingga perangkat lunak yang akan dibangun dapat berinteraksi dengan data dari *Twitter*. *Library* ini digunakan untuk menjembatani bahasa pemograman *Python* dengan *Twitter*. Dengan menggunakan *library* ini, data *tweet* dapat dikumpulkan dan diakses sebagai sumber data penelitian (Wisdom and Gupta 2016). Data yang

telah didapatkan akan melalui *preprocessing* dalam penerapan *text mining.*

2.2 Text Mining

Text mining merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi, dimana *text mining* merupakan variasi dari *data mining* yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual dengan jumlah besar (Benjamin Bengfort, Tony Ojeda 2018). *Text mining* memerlukan beberapa tahap awal untuk mempersiapkan agar teks dapat diubah menjadi lebih terstruktur. Salah satu implementasi dari *text mining* adalah tahap *Text Preprocessing* yang perlu dilakukan yaitu :

2.2.1 *Case Folding*

Case folding yaitu merubah semua karakter huruf pada sebuah kalimat menjadi huruf kecil dan menghilangkan karakter yang dianggap tidak valid seperti angka, tanda baca, hastag, karakter kosong, dan *Uniform Resources Locator (URL).* Pada proses ini penulis menggunakan fungsi *lower*() yang merupakan bawaan dari P*ython*. Setelah data melalui proses *case folding*, selanjutnya akan melalui proses *tokenizing*.

2.2.2 *Tokenizing*

Proses *tokenizing* yaitu memecah dokumen teks yang terdiri dari sekumpulan kalimat menjadi bagian-bagian kata yang disebut token (Indraloka et al. 2017). Setelah melalui proses *tokenizing* kita bisa mendapatkan jumlah kemunculan setiap token nya. Proses *tokenization* bisa menggunakan fungsi *split()* yang merupakan bawaan dari *Python*. Setelah data melalui proses *tokenizing*, selanjutnya akan melalui proses S*topword Removal*.

2.2.3 *Stopword Removal*

Stopword removal merupakan tahapan mengambil kata-kata yang dianggap penting dari hasil *tokenizing* atau membuang kata-kata yang dianggap tidak terlalu mempunyai arti penting dalam kalimat *tweet* (Sunada 2017). Dalam proses *text mining* proses ini digunakan untuk mengurangi jumlah kata yang harus diproses. Melakukan *Stopword Removal* Bahasa Indonesia bisa menggunakan *library* Sastrawi.

Sastrawi merupakan *library* yang dapat digunakan untuk mendapatkan kata dasar dari kata yang kita inputkan serta *library* ini juga mendukung proses *Stopword Removal* (Indraloka et al. 2017). Selain memanfaatkan *library* tersebut, dapat juga menambahkan kamus *stopword* terhadap kata-kata yang tidak ada dalam kamus Sastrawi seperti kata yang berupa singkatan, kata sinonim, dan kata yang belum ada pada kamus Sastrawi. Setelah data melalui proses *Stopword Removal*, selanjutnya akan melalui proses *Stemming*.

2.2.4 *Stemming*

Stemming bertujuan untuk mentransformasikan kata menjadi kata dasarnya (*root word*) dengan menghilangkan semua imbuhan baik awalan maupun akhiran. Pada proses ini penulis menggunakan *library* Sastrawi. Sastrawi *Python* merupakan *library* yang dapat digunakan untuk mendapatkan kata dasar dari kata yang kita inputkan. Algoritma yang digunakan oleh *library* ini adalah algoritma Nazief dan Andriani, dimana algoritma ini merupakan salah satu algoritma yang cukup populer untuk melakukan *stemming* kata dalam Bahasa Indonesia (Indraloka et al. 2017).

Data yang telah melalui tahap *preprocessing* akan diproses menggunakan algoritma *Importance of a Term in a Document (ITD) and Importance of a Term for expressing Sentiment (ITS)* untuk mendapatkan bobotan kata.

2.3 *Algoritma Importance of a Term in a Document (ITD) and Importance of a Term for expressing Sentiment (ITS)*

Algoritma *Importance of a Term in a Document* (ITD) *and Importance of a Term for expressing Sentiment* (ITS) digunakan dalam proses perhitungan bobot (W) terminologi kata. Algoritma ITD digunakan untuk mengevaluasi kemampuan sebuah kata dalam menggambarkan suatu makna terhadap sebuah dokumen (Deng, Luo, and Yu 2014). Perhitungan ITD dapat dituliskan sebagai berikut:

$$ITD(f_i, d_j) = \begin{cases} 1 & f_{ij} > 0 \\ 0 & otherwise \end{cases} \quad \text{.....(1)}$$

$$ITD(f_i, d_j) = tf_{ij} \quad \text{.....(2)}$$

$$ITD(f_i, d_j) = 0.5 + \frac{0.5\ x\ f_{ij}}{\max_k f_{kj}} \quad \text{.....(3)}$$

Keterangan :

f_{ij} : Frekuesi *term* dalam suatu dokumen.

f_i : Frekuensi *term*

d_j : Dokumen

$maxf_{kj}$: Nilai maksimum dari kumpulan frekuensi *term* yang ada pada suatu dokumen.

Pada persamaan 1 menggunakan bobot biner untuk mendefinisikan $ITD(f_i, d_j)$ dengan pemberian nilai 1 pada kata yang terdapat dalam sebuah dokumen dan nilai 0 pada kata yang tidak terdapat dalam sebuah dokumen. Pada persamaan 2 menghitung kata f_i (frekuensi *term*) yang terdapat pada dokumen d_j karena kata yang terdapat dalam dokumen menunjukkan makna dalam dokumen tersebut. Persamaan 3 menggunakan normalisasi frekuensi kata yang muncul dalam sebuah dokumen yaitu dengan memperhatikan jumlah kemunculan kata. Sehingga pada tahap ini penulis menggunakan persamaan 3 karena dapat menghasilkan nilai akurasi yang lebih tinggi daripada persamaan lainnya. Setelah mendapatkan bobot *ITD* maka akan diberlakukan pembobotan *Importance of a Term for expressing Sentiment (ITS)*. Algoritma *ITS* digunakan untuk menetapkan bobot kata pada semua dokumen berdasarkan fungsi statistik untuk mengekspresikan sentimen atau kelas klasifikasi (Deng, Luo, and Yu 2014). Beberapa fungsi perhitungan ITS yaitu:

a. *Document Frequency* (DF)

Document frequency merupakan jumlah kemunculan kata yang ada dalam sebuah dokumen yang merepresentasikan terhadap sebuah kelas. Pertama ditentukan jumlah kata yang menyusun sebuah dokumen kemudian ditentukan frekuensi kata yang menyusun dokumen tersebut. Unit pengukur yang umum digunakan untuk menghitung adalah bit sehingga menggunakan *logaritma* (log) . Persamaan ini dapat ditulis sebagai berikut:

$$DF(f_i) = \log\left(\frac{N}{DF(w)}\right) + 1 \quad \text{.....(4)}$$

$keterangan$:

$N = total\ dokumen$

$$DF(w) = jumlah\ dokumen\ yang\ memiliki\ kata\ w$$

b. *Based on mutual information*

Based on mutual information banyak digunakan dalam pemodelan bahasa statistik asosiasi kata. Mutual Information merupakan nilai ukur yang menyatakan keterikatan atau ketergantungan antar variabel. Fungsi ini dapat ditulis sebagai berikut:

$$MI(f_i) = max\left\{log\frac{P(f_i,D^1)}{P(f_i)x\,P(D^1)}, log\frac{P(f_i,D^2)}{P(f_i)x\,P(D^2)}\right\} \quad \text{.....(5)}$$

c. *Based on information gain*

Based on information gain digunakan untuk mengukur jumlah bit informasi yang diperoleh untuk mengetahui kelas prediksi yang akan ada atau tidaknya kata dalam sebuah data. penghitungan Gain Ratio adalah hasil dari penghitungan Mutual Information dibagi dengan hasil penghitungan Entropy. Fungsi ini dapat ditulis sebagai berikut :

$$IG(f_i) = \frac{MI(f_i,D^k)}{E(f_i)} = \frac{log\frac{P(f_i,D^k)}{P(f_i)P(D^k)}}{log\frac{1}{P(f_i)}} \quad \text{.....(6)}$$

Keterangan :

$P(f_i, D^k)$: Probabilitas dokumen mengandung term f_i dan merupakan kelas dokumen D^k

$P(f_i)$: Probabilitas term f_i pada suatu dokumen

$P(D^k)$: Probabilitas dokumen merupakan dokumen D^k

Pada tahap ini penulis menggunakan rumus persamaan 4 karena merupakan persamaan yang tidak menyatakan keterikatan variable sehingga sesuai dengan *Naïve Bayes Classifier*. Setelah mendapatkan bobot ITD dan bobot ITS maka diberlakukan fungsi berikut untuk mengetahui bobot dokumen :

$$W_{i,j} = ITD(f_i, d_j) \times ITS(f_i) \quad \text{.....(7)}$$

Setelah mendapatkan bobot masing-masing kata pada dokumen, selanjutnya dilakukan penyusunan model dengan menggunakan metode *Naive Bayes Classifier.*

2.4 *Naive Bayes Classifier* (NBC)

Naive Bayes Classifier menempuh dua tahap dalam proses klasifikasi teks,

yaitu tahap pelatihan dan tahap pengujian. Pada tahap pelatihan dilakukan proses analisis terhadap sampel data berupa pemilihan vocabulary, yaitu kata yang mungkin muncul dalam koleksi *tweet* sampel yang sedapat mungkin dapat menjadi representasi *tweet*. Selanjutnya adalah penentuan probabilitas prior bagi tiap kelas berdasarkan sampel *tweet*. Pada tahap pengujian ditentukan nilai kelas prediksi dari suatu *tweet* berdasarkan *term* yang muncul pada *tweet* yang diklasifikasi. Proses klasifikasi *Naive Bayes Classifier* terhadap data yaitu dengan mempresentasikan setiap kelas dengan atribut " $X_1, X_2, X_3, \ldots X_n$ " yang mempunyai makna bahwa X_1 untuk kata pertama, X_2 adalah kata kedua, dan seterusnya (Sunada 2017). Tahap pertama dalam permodelan Naive Bayes Classifier yaitu menghitung prior probability. Prior probability merupakan perhitungan probabilitas dari total data. Persamaan teorema *Bayes* yaitu :

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)} \quad \text{.....(8)}$$

$keterangan$:

x = Data dengan class yang belum diketahui

c = Hipotesis data merupakan suatu class spesifik

P(c|x) = Probabilitas hipotesis berdasar kondisi (posteriori probability)

P(c) = Probabilitas hipotesis (prior probability)

P(x|c) = Probabilitas berdasarkan kondisi pada hipotesis

P(x) = Probabilitas c

Rumus diatas menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (Posterior) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik karakteristik sampel pada kelas C (disebut likelihood), dibagi dengan peluang kemunculan karakteristik sampel secara global (disebut evidence). Sehingga rumus diatas dapat pula ditulis sebagai berikut :

$$P(V_j) = \frac{D_j}{D} \quad \text{.....(9)}$$

Keterangan :

$P(V_j)$: prior probability

D : jumlah dokumen

D_j : jumlah dokumen (D) terhadap sebuah kelas

Untuk mendapatkan bobot term terhadap semua kelas prediksi atau *conditional probability* menggunaan persamaan (10) yaitu :

$$P(X_1|V_j) = \frac{W_{i,j}+1}{N+N_j} \quad \text{.....(10)}$$

$keterangan$:

$W_{i,j}$ = $bobot\ term$

N = jumlah term pada sebuah dokumen

N_j = total term pada dokumen

Saat melakukan proses klasifikasi dokumen, *Naive Bayes Classifier* akan mencari nilai probabilitas tertinggi dari :

$$V_{map} = \underset{V_j eV}{argmax} \frac{P(X_1, X_2, X_3, \dots, X_n|V_j)P(V_j)}{P(X_1, X_2, X_3, \dots, X_n)} \quad \text{.....(11)}$$

Jika nilai dari $P(X_1, \dots, X_n)$ adalah konstan untuk semua kelas V_j maka persamaan (11) dapat ditulis :

$$V_{MAP} = \underset{V_j eV}{argmax}\, P(X_1, X_2, X_3, \dots, X_n|V_j)P(V_j) \quad \text{.....(12)}$$

Sehingga dari persamaan (11) dapat ditulis sebagai :

$$V_{MAP} = \prod_{i=1}^{n} P(X_1|V_j)P(V_j) \quad \text{.....(13)}$$

Keterangan :

V_{MAP} : bobot dokumen terhadap semua kata yang diujikan

V_j : kelas prediksi *tweet*, dengan :

J_0 : soto

J_1 : gudeg

J_2 : mie

J_3 : sate

J_4 : rujak

J_5 : pempek

J_6 : rendang

J_7 : pecel

J_8 : kuliner lain

J_9 : bukan kuliner

$P(X_1|V_j)$: probabilitas dari V_j

Hasil klasifikasi yang telah didapat akan ukur nilai akurasinya dengan menggunakan C*onfusion Matrix*.

2.5 *Confusion Matrix*

Confusion matrix merupakan alat ukur yang standar digunakan untuk mengetahui seberapa akurat hasil perkiraan dari sistem klasisfikasi. Istilah yang digunakan untuk nilai akurasi dalam confusion matrix adalah sebagai berikut:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \quad \text{.....(14)}$$

1. *True Positif* (TP) *True Positif* adalah merupakan data yang klasifikasi riilnya positif dan diprediksi positif.
2. *True Negative* (TN) *True Negative* adalah merupakan data yang klasifikasi riilnya negatif dan diprediksi negatif.
3. *False Positif* (FP) *False Positif* adalah merupakan data yang klasifikasi riilnya negatif dan diprediksi positif.
4. *False Negative* (FN) *False Negative* adalah merupakan data yang klasifikasi riilnya positif dan diprediksi negatif.

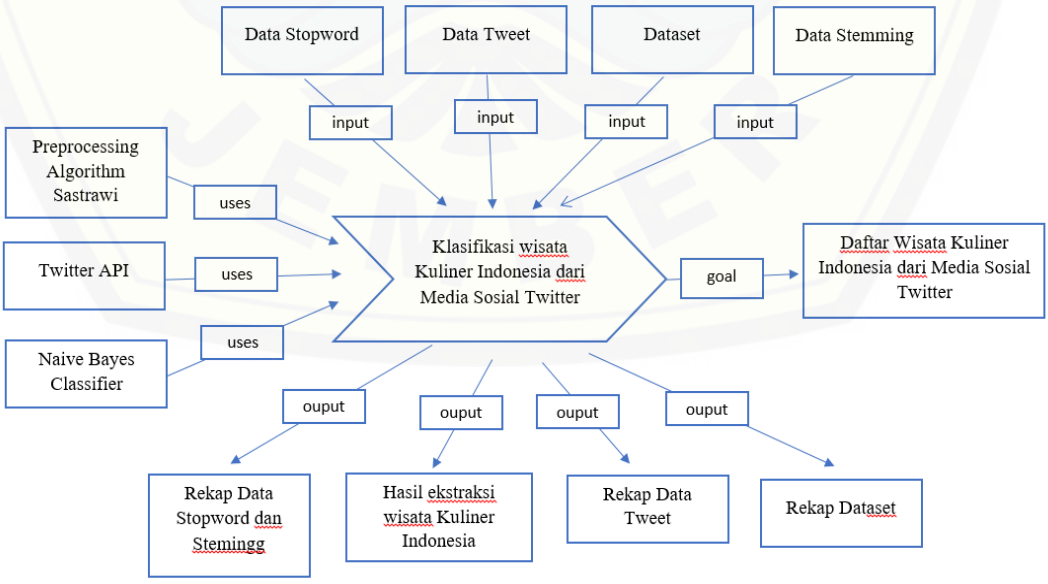
BAB 3. METODOLOGI PENELITIAN

3.1 Jenis Penelitian

Penelitian yang dilakukan merupakan jenis penelitian kuantitatif. Penelitian kuantitatif merupakan metode penelitian yang berlandaskan pada filsafat positivisme yang digunakan untuk meneliti populasi pada sampel tertentu. (Ni Luh Ratniasih, Made Sudarma 2017). Penelitian ini termasuk penelitian kuantitatif karena menggunakan sampel dalam pengumpulan data berupa postingan pada *Twitter*. Serta mengolah data numerik untuk melakukan pengujian terhadap hipotesis yang telah ditentukan.

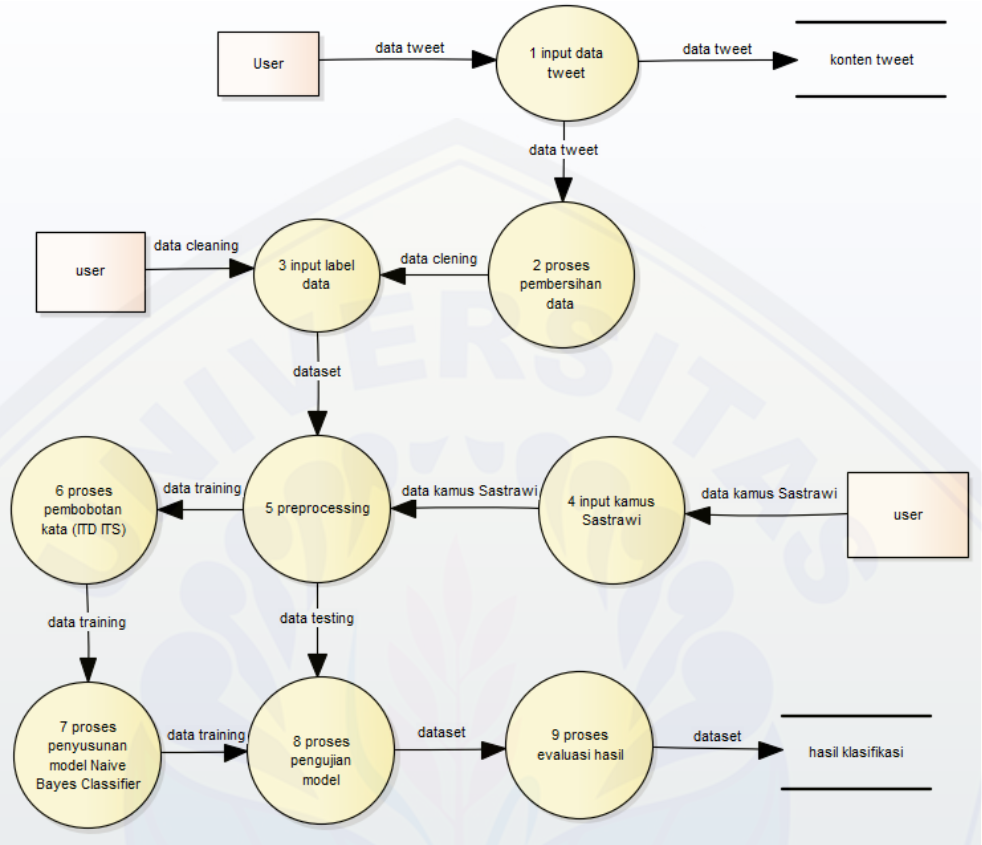
3.2 Objek Penelitian

Objek penelitian merupakan postingan (*tweet)* pada sosial media *Twitter*. Pada penelitian ini data didapat dari pengambilan teks secara bertahap untuk diolah menjadi daftar wisata kuliner Indonesia yang meliputi : soto, gudeg, mie, sate, pempek, rendang, pecel, kuliner lain, dan bukan kuliner. Terdapat 80% data *training* sebagai media pembelajaran untuk membangun model dan 20% data *testing* sebagai pengujian model. Dataset yang digunakan merupakan 5.000 *tweet* terakhir atau *tweet* terbaru yaitu 4000 *tweet* sebagai data *training* dan 1000 *tweet* untuk data *testing*. Proses input dan output data yang digunakan digambarkan dalam Gambar 3.1.



Gambar 3. 1 Data Flow Input Output

Data flow diagram sistem klasifikasi informasi wisata kuliner dari media sosial *Twitter* dapat dilihat pada Gambar 3.2.



Gambar 3. 2 Data Flow Diagram Sistem

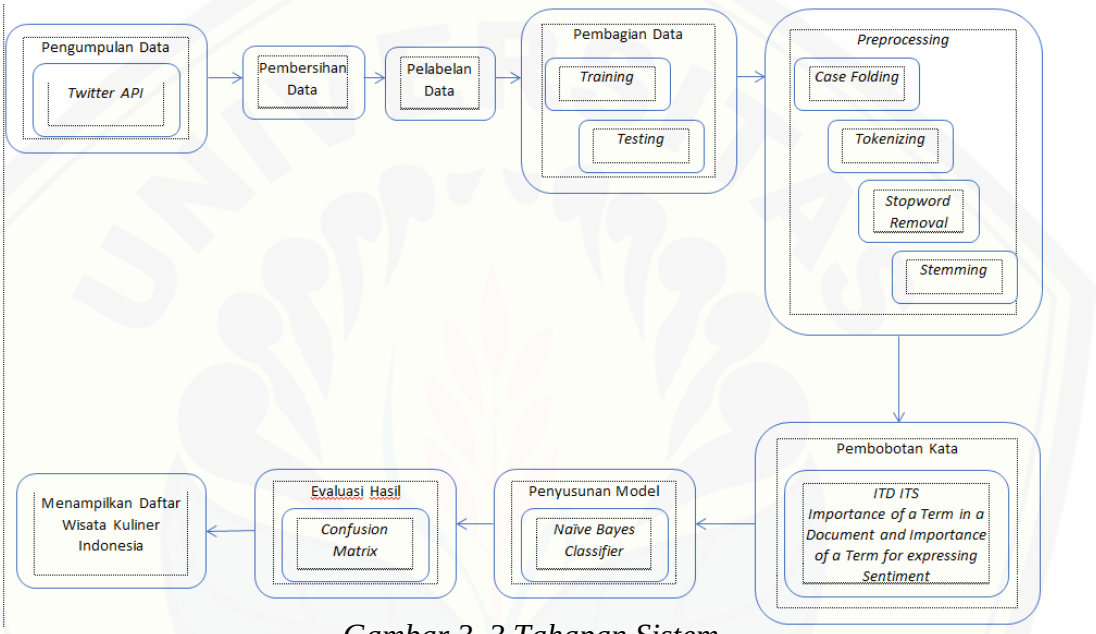
Pada data flow diagram diatas terdapat *terminator user* selaku pemberi *input* kepada sistem yaitu menginputkan data *tweet* ketika pengumpulan data, menginputkan data *cleaning* untuk diberi label, dan menginputkan data kamus Sastrawi untuk diproses pada tahap *preprocessing*. Terdapat data store untuk menyimpan data pada database yaitu konten *tweet* dan hasil klasifikasi. Serta terdapat sembilan proses aktifitas yang mengolah *input* menjadi *output* yaitu: *input* data *tweet*, proses pembersihan data, input label data, input kamus Sastrawi, *preprocessing*, proses pembobotan kata dengan ITD ITS, proses penyusunan model dengan *Naïve Bayes Classifier*, proses pegujian model, dan proses evaluasi hasil. Sehingga alur data flow diagram sistem dapat dilihat pada Gambar 3.2.

3.3 Tempat dan Waktu Penelitan

Tempat dilaksanakannya penelitian ini yaitu pada media sosial *Twitter* dengan memanfaatkan *Twitter API* sebagai sumber data. Waktu penelitian dilakukan selama tiga bulan, dimulai pada bulai Oktober 2019 sampai dengan Januari 2020.

3.4 Tahapan Penelitian

Alur dari tahapan sistem dapat dilihat pada Gambar 3.3 di bawah ini :



Gambar 3. 3 Tahapan Sistem

Pada alur tahapan sistem dimulai dengan pengumpulan data sampai tahap menampilkan kelas klasifikasi kuliner. Penjelasan setiap prosesnya dapat dilihat pada Tabel 3.1.

No	Tahap	Input	Proses	Output
1	Pengumpulan Data	*Twitter API*	Pengumpulan data dilakukan dengan menggunakan *Twitter API* dengan menggunakan *library tweepy* sebagai penghubung dengan	*Dataset*

			Python dan menggunakan adaptor *pscopg2* sebagai penguhubung dengan database *PostgreSQL*	
2	Pembersihan Data	Dataset	*Cleaning* data dengan penghapusan atribut yang tidak diperlukan seperti penghapusan kata *retweet*, simbol, angka, emoticon, url, dan karakter kosong	Dataset yang telah di*cleaning*
3	Pelabelan Data	Dataset yang telah di*cleaning*	Pelabelan data merupakan proses pemberian label kelas terhadap *tweet* sesuai dengan isi dan kelas kategori	Dataset yang telah diberi label
4	Pembagian Data	Dataset yang telah diberi label	Pembagian data 80% *training* untuk membangun model dan 20% data *testing* untuk menguji model	Data *training* dan data *testing*
5	*Text Preprocessing*	Data *training*	*Casefolding* : mengubah dataset menjadi *lowercase*.	*Dataset tweet* yang menjadi *lowercase*
			Tokenizing : melakukan pemotongan kalimat menjadi kata / token.	*Dataset tweet* yang telah menjadi kata / token

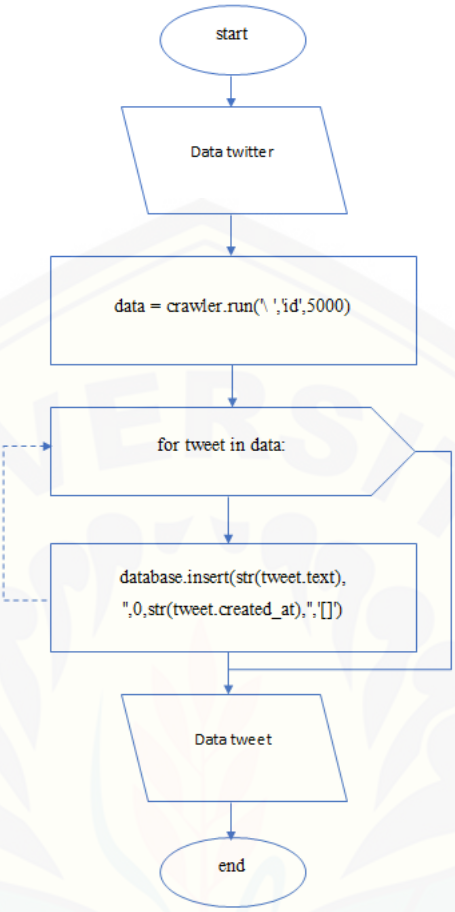
		Dataset tweet yang telah menjadi kata / token	*Stopword* : menghilangkan kata yang tidak memiliki makna penting	*Dataset tweet* yang telah ter*filter* oleh *stopword*
			Stemming : menghilangkan kata imbuhan baik awalan maupun akhiran	*Dataset tweet* yang telah menjadi kata dasar
6	*Importance of a Term in a Document (ITD) and Importance of a Term for expressing Sentiment (ITS)*	Dataset yang telah melewati *preprocessing*	Melakukan proses perhitungan normalisasi frekuensi jumlah kemunculan kata	Dataset dengan bobot ITD
			Melakukan proses perhitungan pembobotan kata berdasarkan fungsi statistik untuk mengekspresikan sentimen atau kelas klasifikasi	*Dataset* dengan bobot ITD ITS
7	*Naive Bayes Classifier*	Data *training* yang telah memiliki bobot ITD ITS	Melakukan proses perhitungan *prior probability* untuk mendapatkan bobot kelas	Kelas prediksi dengan bobot *prior probability*
			Melakukan perhitungan *conditional probability* / *training*	Dataset dengan bobot *training*

			yaitu pembobotan kata terhadap kelas untuk membangun model	
		Data *Testing*	Melakukan pengujian / *testing naive bayes* yaitu perhitungan bobot dokumen terhadap semua kelas untuk pengujian model	Dataset dengan bobot *testing*
			Melakukan proses perbandingan terhadap bobot dokumen untuk mendapatkan hasil bobot tertinggi	Kelas klasifikasi wisata kuliner
8	Evaluasi hasil akurasi menggunakan *confusion matrix*	Kelas prediksi dan kelas klasifikasi *testing*	Melakukan proses perhitungan persentasi akurasi sistem klasifikasi wisata kuliner	Hasil persentase akurasi sistem klasifikasi wisata kuliner

Tabel 3. 1 Penjelasan Algoritma Sistem

3.4.1 Pengumpulan Data

Pengumpulan data dilakukan dengan memanfaatkan *Twitter API* mengunakan *library Tweepy* yang tersedia pada bahasa pemrograman *Python*. Dataset yang digunakan merupakan 5.000 *tweet* terakhir atau *tweet* terbaru yang kemudian disimpan dalam database *PostgreSQL* menggunakan adaptor database *psycopg2*. *Flow chart* tahap *crawling* data dapat dilihat pada Gambar 3.4.



Gambar 3. 4 Flow Chart Crawling Data

3.4.2 Pembersihan Data

Pembersihan data meliputi proses pembuangan angka menggunakan fungsi python *re.sub (r"\d+","",var)*, membuang simbol menggunakan fungsi *translator(string.maketrans ("","),string.punctuation)*, membuang karakter kosong menggunakan fungsi *strip( )*, membuang url menggunakan fungsi *join (\w+:\/\/\S+)*, membuang kata *retweet* dengan menggunakan fungsi *re.compile('RT')*, membuang *emoticon* dengan menggunakan fungsi *join(re.sub("([@#][^\s]+)|([^0-9A-Za-z \t])*. Setelah data melalui tahap pembersihan data, data akan melalui tahap pelabelan data.

3.4.3 Pelabelan Data

Pada penelitian ini penentuan kelas klasifikasi dilakukan secara manual. Pemberian label dilakukan dengan memperhatikan isi dari *tweet* yang akan diberi label. *Labelling* manual dilakukan oleh penulis karena dalam hal

ini tidak membutuhkan keahliah khusus dalam pemberian label. *Tweet* yang mengandung informasi wisata kuliner yaitu menu dan lokasi kuliner akan diberi label sesuai dengan menu kuliner tersebut. Sedangkan *tweet* yang hanya berisi mengenai kuliner tanpa adanya informasi lokasi kuliner tersebut akan diberi label "kuliner lain", serta *tweet* yang berisi konten selain kuliner akan diberi label "bukan kuliner". Daftar kelas atau label dalam penelitian ini dapat dilihat pada Tabel 3.2.

Kode	Keterangan
0	Soto
1	Gudeg
2	Mie
3	Sate
4	Rujak
5	Pempek
6	Rendang
7	Pecel
8	Kuliner Lain
9	Bukan Kuliner

Tabel 3. 2 Kelas Kategori

Pada kelas kategori diatas didapatkan kode 0 hingga 7 yang menunjukkan kelas kategori menu kuliner. Sebuah *tweet* dapat dilabeli kode 0 hingga 7 jika *tweet* tersebut mengandung informasi menu kuliner sesuai dengan kode kelas serta mengandung informasi dimana lokasi kuliner tersebut tersedia. Lokasi tersebut dapat berupa nama tempat, nama wilayah, ataupun informasi sekitar kuliner. Sehingga *tweet* dengan kode 0 hingga 7 merupakan *tweet* tentang wisata kuliner sesuai dengan keterangan menu kuliner tersebut. Jika sebuah *tweet* hanya mengandung informasi menu kuliner baik menu dalam kelas kategori maupun menu yang bukan kelas kategori akan diberi kode label 8. Label 8 merupakan *tweet* yang berisi informasi selain wisata kuliner. Informasi tersebut dapat berupa rasa, harga, menu kuliner, hingga bentuk

makanan yang sebenarnya tidak terkait dengan informasi wisata kuliner. Sedangkan label kode 9 merupakan *tweet* yang berisi informasi selain kuliner baik wisata kuliner maupun hal lain mengenai kuliner. Informasi label 9 dapat berupa informasi pendidikan, ekonomi, sosial, politik, dan informasi lain selain kuliner. Setelah data memiliki label, data akan melalui tahap pembagian data yaitu data *training* dan data *testing*.

3.4.4 Pembagian Data

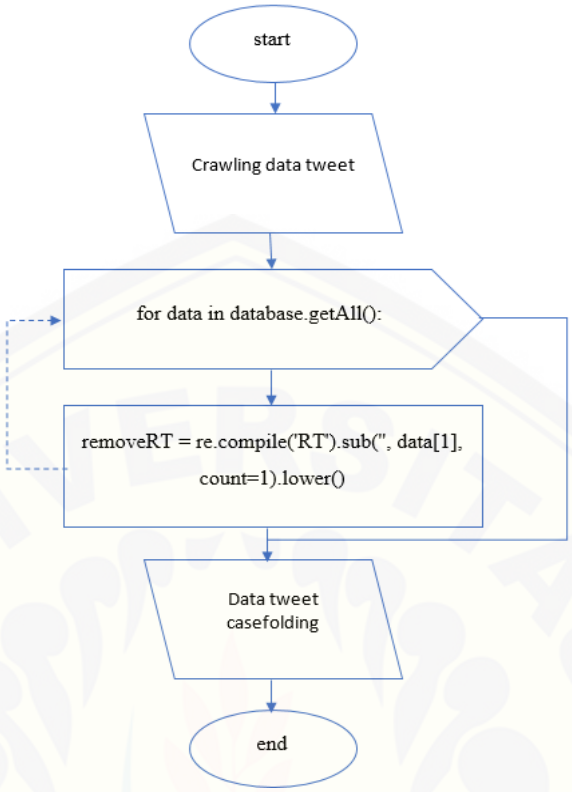
Pembagian data dilakukan untuk membagi dataset menjadi data *training* dan data *testing*. Data *training* merupakan data latih untuk melatih model yang akan dibuat. Sedangkan data *testing* merupakan data uji untuk menguji model yang telah dilatih. Penulis menggunakan pembagian dalam bentuk persentase yaitu 80% data *training* dan 20% data *testing*. Pembagian dataset ini sangat berpengaruh terhadap akurasi sistem karena semakin baik atau semakin bervariasi data training yang dimiliki maka dapat menghasilkan tingkat akurasi yang lebih baik. Setelah data dibagi menjadi data *training* dan data *testing* data akan melalui proses *prepocessing*.

3.4.5 *Preprocessing*

Prepocessing merupakan tahapan mempersiapkan data tekstual yang akan digunakan agar dapat diproses pada tahapan berikutnya. Proses yang dilakukan pada tahapan *preprocessing* yaitu:

a. *Case Folding*

Dalam *text preprocessing* kita bisa menggunakan fungsi *lower*() yang merupakan bawaan dari *Python*. Ilustrasi proses C*ase Folding* seperti kalimat “Menu Soto Betawi pada Taman Santap Rumah Kayu menjadi Menu Terfavorit.” diproses menjadi “menu soto betawi pada taman santap rumah kayu menjadi menu terfavorit”. Pada contoh kalimat tersebut mengubah huruf kapital menjadi huruf kecil (*lower)* dan menghilangkan tanda yaitu tanda baca titik yang terdapat pada akhir kalimat. *Flow chart Case Folding* dapat dilihat pada Gambar 3.5.

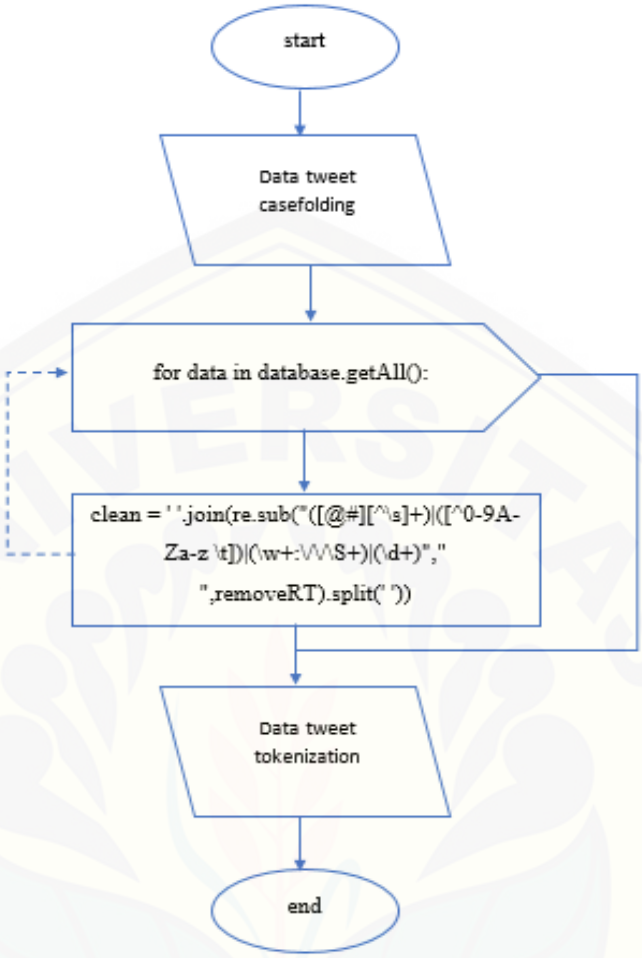


Gambar 3. 5 Flow Chart Casefolding

Pada tahap ini juga dilakukan proses pembuangan kata retweet dengan menggunakan fungsi *re.compile('RT').* Setelah data melalui proses *Case Folding* selanjutnya akan melalui proses *Tokenizing*.

b. *Tokenizing*

Melakukan proses T*okenizing* bisa menggunakan fungsi *split()* yang merupakan bawaan dari *Python*. Pada proses *tokenizing* kita akan memecah dokumen teks yang terdiri dari sekumpulan kalimat menjadi bagian-bagian kata yang disebut token. Ilustrasi proses *Tokenizing* seperti kalimat "menu soto betawi pada taman santap rumah kayu menjadi menu favorit" diproses menjadi ('menu' 'soto' 'betawi' 'pada' 'taman' 'santap' 'rumah' 'kayu' 'menjadi' 'menu' 'terfavorit'). Pada ilustrasi tersebut proses ini melakukan pemotongan string. *Flow chart Tokenizing* dapat dilihat pada Gambar 3.6.



Gambar 3. 6 Flow Chart Tokenizing

Pada saat mengubah kalimat menjadi token juga dilakukan proses pembuangan angka menggunakan fungsi python *re.sub (r"\d+","",var)*, membuang simbol menggunakan fungsi *translator(string.maketrans ("","",),string.punctuation)*, membuang karakter kosong menggunakan fungsi *strip( )*, membuang url menggunakan fungsi *(\w+:\/\/\S+)*, membuang kata *retweet* dengan menggunakan fungsi *re.compile('RT')*, membuang *emoticon* dengan menggunakan fungsi *join(re.sub("([@#][^\s]+)|([^0-9A-Za-z \t])*.Setelah data melalui proses *Tokenizing*, selanjutnya akan melalui *Stopword Removal*.

c. *Stopword Removal*

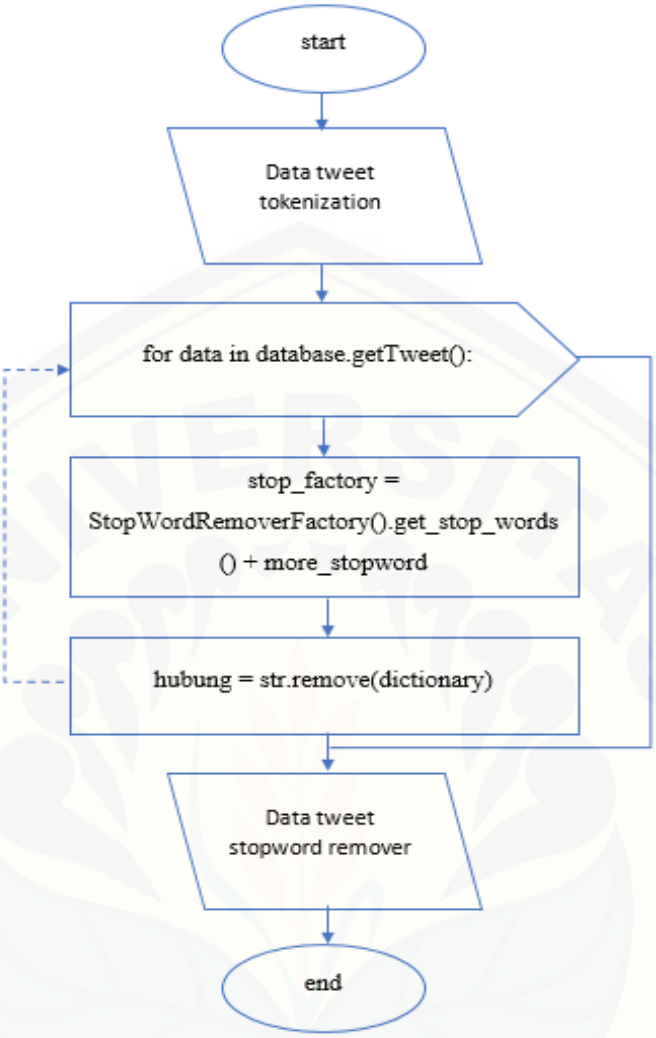
Tahapan ini akan mengambil kata-kata yang dianggap penting dari hasil *tokenizing* atau menghilangkan kata-kata yang dianggap tidak

terlalu mempunyai arti penting dalam kalimat *tweet* menggunakan pustaka Sastrawi. Pada tahap ini mengunakan fungsi *StopWordRemoverFactory().create_stop_word_remover().remove().* Daftar kamus *stopword remover* dari *Sastrawi* dapat dilihat pada Gambar 3.7.

['yang', 'untuk', 'pada', 'ke', 'para', 'namun', 'menurut', 'antara', 'dia', 'dua', 'ia', 'seperti', 'jika', 'jika', 'sehingga', 'kembali', 'dan', 'tidak', 'ini', 'karena', 'kepada', 'oleh', 'saat', 'harus', 'sementara', 'setelah', 'belum', 'kami', 'sekitar', 'bagi', 'serta', 'di', 'dari', 'telah', 'sebagai', 'masih', 'hal', 'ketika', 'adalah', 'itu', 'dalam', 'bisa', 'bahwa', 'atau', 'hanya', 'kita', 'dengan', 'akan', 'juga', 'ada', 'mereka', 'sudah', 'saya', 'terhadap', 'secara', 'agar', 'lain', 'anda', 'begitu', 'mengapa', 'kenapa', 'yaitu', 'yakni', 'daripada', 'itulah', 'lagi', 'maka', 'tentang', 'demi', 'dimana', 'kemana', 'pula', 'sambil', 'sebelum', 'sesudah', 'supaya', 'guna', 'kah', 'pun', 'sampai', 'sedangkan', 'selagi', 'sementara', 'tetapi', 'apakah', 'kecuali', 'sebab', 'selain', 'seolah', 'seraya', 'seterusnya', 'tanpa', 'agak', 'boleh', 'dapat', 'dsb', 'dst', 'dll', 'dahulu', 'dulunya', 'anu', 'demikian', 'tapi', 'ingin', 'juga', 'nggak', 'mari', 'nanti', 'melainkan', 'oh', 'ok', 'seharusnya', 'sebetulnya', 'setiap', 'setidaknya', 'sesuatu', 'pasti', 'saja', 'toh', 'ya', 'walau', 'tolong', 'tentu', 'amat', 'apalagi', 'bagaimanapun']

Gambar 3. 7 Kamus Stopword Remover Sastrawi

Ilustrasi proses S*topword Removal* seperti kalimat (‘menu’ ‘soto’ ‘betawi’ ‘pada’ ‘taman’ ‘santap’ ‘rumah’ ‘kayu’ ‘menjadi’ ‘menu’ ‘terfavorit’) diproses menjadi (‘menu’ ‘soto’ ‘betawi’ ‘taman’ ‘santap’ ‘rumah’ ‘kayu’ ‘menu’ ‘terfavorit’). Pada ilustrasi tersebut menghilangkan kata (‘pada’) dan (‘menjadi’) karena dianggap tidak terlalu mempunyai arti penting dalam kalimat tersebut. Selain menggunakan kamus yang tersedia pada kamus Sastrawi penulis juga membuat kamus tambahan dengan menambahkan pada *stop_factory*. *Flow chart Stopword Remover* dapat dilihat pada Gambar 3.8.



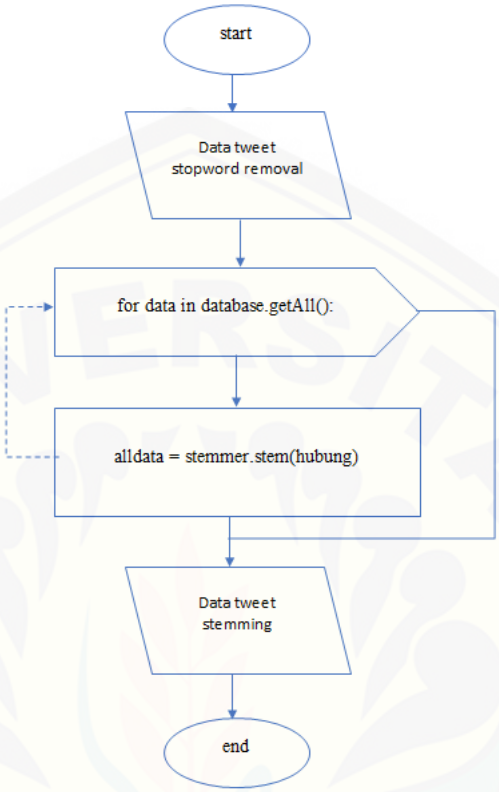
Gambar 3. 8 Flow Chart Stopword Remover

Setelah data melalui proses *Stopword Removal*, selanjutnya akan melalui proses *Stemming*.

d. *Stemming*

Stemming bertujuan untuk mentransformasikan kata menjadi kata dasarnya (*root word*) dengan menghilangkan semua imbuhan kata. Pada proses ini penulis menggunakan *library* Sastrawi dengan menerapkan fungsi *StemmerFactory ().create_stemme r().stem()* Ilustrasi proses *Stemming* seperti kalimat (‘menu’ ‘soto’ ‘betawi’ ‘taman’ ‘santap’ ‘rumah’ ‘kayu’ ‘menu’ ‘terfavorit’) diproses menjadi (‘menu’ ‘soto’ ‘betawi’ ‘taman’ ‘santap’ ‘rumah’ ‘kayu’ ‘menu’ ‘favorit’). Pada contoh kalimat tersebut menghilangkan awalan ‘ter-’ pada kata “terfavorit”

sehingga menjadi kata dasar “favorit”. *Flow chart Stemming* dapat dilihat pada Gambar 3.9.



Gambar 3. 9 Flow Chart Stemming

Data yang telah melalui tahap *preprocessing* akan dihitung nilai keterhubungan bobot kata dengan kalimat menggunakan algoritma *Importance of a Term in a Document (ITD) and Importance of a Term for expressing Sentiment (ITS).*

3.4.6 Pembobotan Kata

Pembobotan kata merupakan tahap penentuan seberapa jauh keterhubungan antar suatu kata terhadap kalimat dengan menghitung nilai atau bobot keterhubungan. Pembobotan ini dapat dilakukan dengan menggunakan algoritma *Importance of a Term in a Document* (ITD) *and Importance of a Term for expressing Sentiment* (ITS). Algoritma ini digunakan dalam proses perhitungan bobot (W) terminologi kata untuk menghitung bobot setiap kata yang paling umum digunakan. Setelah diketahui bobot masing-masing setiap kata, selanjutnya dilakukan

penyusunan model dengan menggunakan metode *Naive Bayes Classifier*.

3.4.7 Penyusunan Model

Penyusunan model digunakan untuk membangun model sebagai pembelajaran sistem. Pada penelitian ini menggunakan metode *Naive Bayes Classifier (NBC)* sebagai penyusunan model. *NBC* dalam melakukan klasifikasi terdapat dua tahap penting yaitu *training* dan *testing*. Pada tahap *training* dilakukan proses analisis terhadap sampel dokumen berupa pemilihan *vocabulary*, yaitu kata yang mungkin muncul dalam koleksi *tweet* sampel yang sedapat mungkin dapat menjadi representasi *tweet*. Selanjutnya adalah penentuan *probabilitas prior* bagi tiap kelas berdasarkan sampel *tweet*. Pada tahap klasifikasi ditentukan nilai kelas dari suatu *tweet* berdasarkan *term* yang muncul pada *tweet* yang diklasifikasi. Hasil klasifikasi yang telah didapat akan diukur nilai akurasinya dengan menggunakan *Confusion Matrix*

3.4.8 Evaluasi Hasil

Pada penelitian ini evaluasi hasil dilakukan dengan menggunakan *Confusion matrix*. *Confusion matrix* merupakan alat ukur yang standar digunakan untuk mengetahui seberapa akurat hasil perkiraan dari sistem klasifikasi. Sehingga sistem dapat digunakan dengan baik berdasarkan nilai akurasi sistem tersebut. Dengan menggunakan rumus persamaan (14), tabel mengenai uji performansi kelas prediksi dan kelas target (Fibrianda and Bhawiyuga 2018) dapat dilihat pada Tabel 3.3.

		Kelas Prediksi	
		Class = Yes	Class = No
Kelas Target	Class = Yes	TP	FN
	Class = No	FP	TN

Tabel 3. 3 Uji Performasi Biner

Keterangan :

1. *True Positif* (TP) *True Positif* adalah merupakan data yang klasifikasi riilnya positif dan diprediksi positif.
2. *True Negative* (TN) *True Negative* adalah merupakan data yang klasifikasi riilnya negatif dan diprediksi negatif.

3. *False Positif* (FP) *False Positif* adalah merupakan data yang klasifikasi riilnya negatif dan diprediksi positif.

4. *False Negative* (FN) *False Negative* adalah merupakan data yang klasifikasi riilnya positif dan diprediksi negatif.

Sedagkan tabel uji performasi untuk kelas non biner atau kelas klasifikasi kuliner dengan 10 kelas dapat dilihat pada Tabel 3.4.

		Kelas Prediksi									
		0	1	2	3	4	5	6	7	8	9
Kelas Target	0	00	01	02	03	04	05	06	07	08	09
	1	10	11	12	13	14	15	16	17	18	19
	2	20	21	22	23	24	25	26	27	28	29
	3	30	31	32	33	34	35	36	37	38	39
	4	40	41	42	43	44	45	46	47	48	49
	5	50	51	52	53	54	55	56	57	58	59
	6	60	61	62	63	64	65	66	67	68	69
	7	70	71	72	73	74	75	76	77	78	79
	8	80	81	82	83	84	85	86	87	88	89
	9	90	91	92	93	94	95	96	97	98	99

Tabel 3. 4Uji Performasi Kelas Kuliner

Berdasarkan tabel di atas dapat diketahui bahwa kolom yang terisi dengan angka sama dengan kelas prediksi dan kelas target maka kolom tersebut diklasifikasi secara benar dan angka pada kolom tersebut seharusnya merupakan angka yang tertinggi diantara angka pada kolom yang lain. Sehingga dengan demikian akan didapatkan nilai akurasi yang tinggi karena mendapatkan angka kesesuaian prediksi yang tinggi.

BAB V KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan analisis dan pengujian yang dilakukan pada bab sebelumnya, maka kesimpulan yang dapat diambil dari penelitian ini adalah sebagai berikut:

1. Klasifikasi informasi wisata kuliner dari *twitter* telah dilakukan dengan tahapan pembangunan model menggunakan *Naïve Bayes Classifier*. Data *tweet* yang sudah melewati tahapan *text preprocessing* selanjutnya dihitung menggunakan nilai probabilitasnya menggunakan algoritma *Importance of a Term in a Document* (ITD) *and Importance of a Term for expressing Sentiment* (ITS) untuk pembobotan kata yang selanjutkan akan diproses menggunakan *Naïve Bayes Classifier* sehingga mendapatkan nilai *conditional probability.* Selanjutnya diproses kedalam persamaan metode untuk dilakukan tahap *testing.* Dimana untuk mencari hasil *testing* terbaik harus mencari nilai probabilitas tertinggi.
2. Penerapan metode *Naïve Bayes Classifier* menjadi sangat optimal ketika data *training* yang digunakan memiliki jumlah data yang banyak dan data yang bervariasi. *Naïve Bayes Classifier* sangat baik dalam mengklasifikasikan teks dengan jumlah data yang kecil atau cuplikan dokumen seperti *tweet*. Penggunaan Algoritma *Importance of a Term in a Document* (ITD) *and Importance of a Term for expressing Sentiment* (ITS) dalam proses perhitungan bobot (W) terminologi kata terhadap metode *Naive Bayes Classifier* menghasilkan nilai akurasi yang tinggi. perhitungan ITD menggunakan normalisasi frekuensi kata yang muncul dalam sebuah dokumen sehingga dapat menghasilkan nilai akurasi yang lebih tinggi. Dalam penelitian ini menggunakan 5000 *dataset* dengan 4000 data *training* dan 1000 data *testing* menghasilkan nilai akurasi 86.5%.

5.2. Saran

Penulis menyarankan pengembangan penelitian lebih lanjut sistem *ektraksi* informasi wisata kuliner menggunakan metode *Naïve Bayes Classifier* sebagai berikut:

1. Untuk mendapat nilai akurasi yang lebih tinggi, diharapkan menambah proses dalam pengolahan data atau *training* data serta selalu melakukan update data karena konten pada *Twitter* selalu bertambah. Penggunaan *training* menjadikan hasil pengolahan data menjadi lebih akurat.
2. Penggunaan variasi data dengan jumlah besar menjadikan proses *training* menjadi lebih beragam, nilai yang dihasilkan akan semakin akurat jika data *training* bervariasi. Serta dapat menyeimbangkan data training lokasi wisata kuliner sehingga tingkat kesalahan prediksi kelas wisata kuliner terhadap kelas kuliner lain dan bukan kuliner dapat berkurang dan meningkatkan nilai akurasi sistem.
3. Penelitian selanjutnya dapat menggunakan penambahan data *training* secara otomatis tanpa adanya penambahan secara manual dari tahap *testing* dan dapat menggunakan server supaya proses pengolahaan data bisa lebih cepat.

DAFTAR PUSTAKA

Agus Hermanto. 2016. "Implementasi Text Mining Menggunakan Naive Bayes Untuk Penentuan Kategori Tugas Akhir Mahasiswa Berdasarkan Abstraksinya." *Konvergensi* 12: 1–10.

Benjamin Bengfort, Tony Ojeda, Rebecca Bilbro. 2018. *Applied Text Analysis with Python*. O'Reilly Media.

Deng, Zhi-hong, Kun-hu Luo, and Hong-liang Yu. 2014. "A Study of Supervised Term Weighting Scheme for Sentiment Analysis." *Expert Systems With Applications* 41 (7): 3506–13. https://doi.org/10.1016/j.eswa.2013.10.056.

Fibrianda, Mercury Fluorida, and Adhitya Bhawiyuga. 2018. "Analisis Perbandingan Akurasi Deteksi Serangan Pada Jaringan Komputer Dengan Metode Naïve Bayes Dan Support Vector Machine (SVM)." *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer* 2 (9): 3112–23.

Hanifah, Raidah, and Isye Susana Nurhasanah. 2018. "Implementasi Web Crawling Untuk Mengumpulkan Web Crawling Implementation For Collecting." *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)* 5 (5): 531–36. https://doi.org/10.25126/jtiik20185842.

Indraloka, Dwi Smaradahana, Budi Santosa, Departemen Matematika, Fakultas Matematka, Pengetahuan Alam, Institut Teknologi, and Sepuluh Nopember. 2017. "Penerapan Text Mining Untuk Melakukan Clustering Data Tweet Shopee Indonesia." *Jurnal Sains Dan Seni Its* 6 (2): 2337–3520.

JJrgens, Pascal, and Andreas Jungherr. 2016. "A Tutorial for Using Twitter Data in the Social Sciences: Data Collection, Preparation, and Analysis." *SSRN Electronic Journal* 01 (January): 1–95. https://doi.org/10.2139/ssrn.2710146.

Le, Cong Cuong, P. W.C. Prasad, Abeer Alsadoon, L. Pham, and A. Elchouemi. 2019. "Text Classification: Naïve Bayes Classifier with Sentiment Lexicon." *IAENG International Journal of Computer Science* 46 (2): 141–48.

Ni Luh Ratniasih, Made Sudarma, Nyoman Gunantara. 2017. "Penerapan Text Mining Dalam Spam Filtering Untuk Aplikasi Chat." *Teknologi Elektro* 16 (3).

Pandhu, Akhmad, and Heru Agus. 2016. "Naive Bayes Classification Pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government." *Journal of Applied Intelligent System* 1 (1): 48-55–55.

Sunada, Dwight. 2017. *Building a Naive Bayes Text Classifier and Accounting for Document Length*. Independently Published.

Wisdom, Vivek, and Rajat Gupta. 2016. "An Introduction to Twitter Data Analysis in Python," no. September: 1–6. https://doi.org/10.13140/RG.2.2.12803.30243.