BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Penerapan metode *Naïve Bayes Classifier* yang digunakan sebagai penelitian pernah dilakukan oleh Fatahillah (2017). Penelitian ini mengembangkan sebuah sistem yang dapat melakukan klasifikasi tweet berisi ujaran positif dan negatif. Sistem dibuat menggunakan teknologi *Node.js* dan *Naïve Bayes Classifier* sebagai metode perhitungan klasifikasi. Berdasarkan penelitian ini, dapat diperoleh kesimpulan pengujian klasifikasi menggunakan *Naïve Bayes Classifier* menghasilkan tingkat akurasi benar 93% dan 7% tingkat akurasi salah serta dalam penelitian ini, *Naïve Bayes Classifier* dapat digunakan untuk mengklasifikasikan *tweet* berdasarkan label.

Penelitian dengan topik analisis *sentimen* pernah dilakukan oleh sejumlah peneliti, Sunardi dkk. (2018), menganalisa *sentimen* terhadap data angket mahasiswa untuk mengetahui kepuasan mahasiswa dalam proses pendidikan menggunakan metode *Naïve Bayes Classifier*. Klasifikasi pada penelitian ini dibagi kedalam kelas positif, negatif dan netral. Penelitian ini berhasil memperoleh data sebanyak 800 data mahasiswa yang mengisi angket. *Sampel* yang digunakan pada penelitian ini sebanyak 100 data angket mahasiswa. Berdasarkan penelitian ini, dapat diperoleh kesimpulan klasifikasi data menggunakan metode *Naïve Bayes Classifier* dengan tingkat akurasi tinggi *precision* 75%, *recall* 75% dan *accuracy* 80%.

Pada penelitian lain, Tuhuteru dan Iriani (2018) juga melakukan analisa Sentimen pada Perusahaan Listrik Negara Cabang Ambon Menggunakan Metode SVM (*Support Vector Machine)* dan *Naïve Bayes Classifier* untuk membadingkan hasil akurasi tertinggi. Pengumpulan data pada penelitian ini diambil menggunakan metode *snipping* dengan jumlah data yang berhasil di peroleh yaitu sebanyak 1491 data. Pembagian data latih dan data uji pada penelitian ini menghasilkan 491 *tweet* sebagai data latih dan 1000 *tweet* sebagai data uji, pengklasifikasian dengan SVM (*Support Vector Machine)* dan NBC (*Naïve Bayes Classifier*) menggunakan dataset yang sama. Berdasarkan penelitian ini, klasifikasi menggunakan metode *Naïve Bayes Classifier* dengan 2 fold pada proses *validasi* menghasilkan akurasi sebesar 67,2% dengan *sentimen* positif 67%, *sentimen* netral 19% dan *sentimen* negatif 14%. Sementara itu, pada metode klasifikasi SVM (*Support Vector Machine*) menghasilkan akurasi sebesar 76,42%, dengan *sentimen* positif sebesar 24%, *sentimen* netral 29% dan *sentimen* negatif 47%.

Pada terhadap *tweet* juga pernah dilakukan oleh Nova dkk. (2019), yang melakukan penelitian terhadap Transportasi KRL Commuter Line menggunakan Metode *Naïve Bayes Classifier*, KNN (*K-Nearest Neighbor*) dan *Decision Tree*. Tujuan penelitian ini untuk mengetahui *sentimen* yang berkembang pada masyarakat terhadap Transportasi KRL Commuter Line dengan membandingkan tingkat akurasi pada ketiga metode klasifikasi tersebut. Adapun teknik pengumpulan data pada penelitian ini dengan mengambil data pada *Twitter* secara *random* dalam kondisi paling baru (*ter-update*) menggunakan *Rapidminer* sebagai

alat untuk mengambil data. Pengujian data set sebanyak 127 dengan mengkategorikan *sentimen* ke dalam tiga kelas yaitu positif, negatif dan netral. Dari penelitian ini, didapatkan akurasi pada *Naïve Bayes Classifier* sebesar 80%, *precision* 66,67%, *sensivity* 100% dan *specificity* 66,67%. Pada metode KNN (*K-Nearest Neighbor*) akurasi yang didapat sebesar 80%, *precision* 100%, *sensivity* 50% dan *specificity* 100%. Sedangkan pada metode *Decision Tree* akurasi sebesar 100%, *precision* 100%, *sensivity* 100% dan *specificity* 100%.

Penelitian lain yang menggunakan metode *Naïve Bayes Classifier* dalam mengklasifikasikan *sentimen* adalah Ayuningtyas (2020), untuk mengetahui *sentimen* masyarakat pada komentar halaman Facebook dan ulasan Google terhadap STMIK Akakom Yogyakarta yang saat ini beralih bentuk UNIVERSITAS TEKNOLOGI DIGITAL INDONESIA. Pada penelitian ini, menganalisa data berbentuk *tekstual* maupun *non-tekstual*. Data *non-tekstual* akan terlebih dahulu diubah kedalam bentuk teks. Data yang diperoleh akan digolongkan ke salah satu kelas *sentimen* positif, negatif atau netral. Pada tahap *preprocessing* data, tahap cleaning penggunaan emoji melalui tahap *konversi* dan *preprocessing* tidak melalui tahap *stemming*. Penelitian ini menerapkan metode SMOTE (*Synthetic Minority Over-Sampling Technique*) untuk membentuk kata *sintetis* dari kelas *minoritas* dan kesimpulan dari penelitian ini berhasil mengumpulkan sebanyak 377 data dengan tingkat *true positif* untuk negatif sebesar 100% dan tingkat akurasi klasifikasi bernilai 89,73%.

Berbeda dari Fatahillah (2017), Sunardi dkk. (2018), Tuhuteru dan Iriani (2018), Nova dkk. (2019), Ayuningtyas (2020), Fatchan dan Sugeng (2022)

menganalisa *sentimen* masyarakat terhadap Ibu Tri Rismaharini yang terpilih menjadi Menteri Sosial pada media *Twitter*. Sebanyak 3000 data berhasil diperoleh dari *crawling data*. Penelitian ini untuk mengetahui tingkat akurasi metode *Naïve Bayes Classifier*. *Sentimen* yang terbentuk berdasarkan data tersebut mengklasifikasikan setiap komentar ke dalam kelas positif dan negatif dengan membagi data penelitian menjadi tiga *data testing*, yakni 30%, 25%, dan 20%. Jumlah akurasi tertinggi pada penelitian ini adalah *data testing* 30% dan *data training* 70%. Sehingga dalam proses pengklasifikasian *Naïve Bayes Classifier* pada penelitian ini memiliki tingkat akurasi sebesar 90,3%.

Penelitian ini membahas mengenai opini masyarakat terhadap Pemindahan Ibu Kota baru Negara Indonesia yang akan berlangsung pada tahun 2024 untuk mengetahui *sentimen* yang berkembang di masyarakat seiring berjalannya waktu hingga pemindahan Ibu kota baru dilaksanakan yakni pada tahun 2024. Pengumpulan data pada penelitian ini dilakukan dengan *crawling data* pada *Twitter* seperti yang dilakukan oleh Fatchan dan Sugeng (2022). Dengan menerapkan metode SMOTE (*Synthetic Minority Over-Sampling Technique*) untuk mengatasi data yang tidak seimbang seperti yang dilakukan oleh Ayuningtyas (2020). Namun, yang membedakan pada penelitian ini dengan sebelumnya yang dilakukan oleh Ayuningtyas (2020) adalah pada *preprocessing* data, tahap *cleaning* penggunaan emoji dihapus karena dapat menganggu proses *preprocessing* dan penelitian ini melalui tahapan stemming. Penelitian ini terbagi menjadi tiga kelas *sentimen* diantaranya *sentimen* positif, negatif dan netral dengan menggunakan metode *Naïve Bayes Classifier*.

Tabel 2. 1 Perbandingan Penelitian

No	Nama	Objek	Tujuan	Metode
1.	Fatahillah (2017)	*Cuitan* (*Twee*t) pada *Twitter*	Mengembangkan sebuah sistem yang dapat melakukan klasifikasi *tweet* berisi ujaran positif dan negatif.	*Naïve Bayes Classifier*
2.	Sunardi dkk. (2018)	Angket mahasiswa STMIK PPKIA	Mengetahui kepuasan mahasiswa dalam proses pendidikan.	*Naïve Bayes Classifier*
3.	Tuhuteru dan Iriani (2018)	Perusahaan Listrik Negara Cabang Ambon	Mengetahui *sentimen* masyarakat di Pulau Ambon terhadap kondisi kelistrikan.	*a. Naïve Bayes Classifier* b. SVM (*Support Vector Machine*)
4.	Nova dkk. (2019)	Transportasi KRL Commuter Line	Mengetahui sentimen masyarakat terhadap penggunaan KRL Communer Line Jabodetabek.	*a. Naïve Bayes Classifier* b. KNN (*K-Nearest Neigbor*) *c. Decision Tree*
5.	Ayuningtyas (2020)	STMIK Akakom Yogyakarta	a) Mengetahui *sentimen* terhadap STMIK Akakom Yogyakarta berdasarkan komentar Facebook dan ulasan Google. b) Mengetahui tingkat akurasi *Naïve Bayes Classifier* dalam mengklasifikasikan *sentimen*.	*Naïve Bayes Classifier*
6.	Sugeng dan Fatchan (2022)	Ibu Tri Rismaharini	Mengetahui tingkat akurasi metode *Naïve Bayes Classifier* dalam menentukan klasifikasi *sentimen* data komentar pada media *Twitte*r terhadap Ibu Tri Rismaharini yang terpilih menjadi Menteri.	*Naïve Bayes Classifier*

Tabel 2. 2 Lanjutan Tabel 2.1

No	Nama	Objek	Tujuan	Metode
7.	Astutik	Ibu Kota Baru Negara Indonesia Tahun 2024	a) Mengetahui *sentimen* Pemindahan Ibu Kota Negara Indonesia Tahun 2024 pada media sosial *Twitter*. b) Mengetahui tingkat akurasi metode *Naïve Bayes Classifier* dalam klasifikasi analisis *sentimen*. c) Mengetahui apakah metode *Naïve Bayes Classifier* dapat diklasifikasikan dalam analisis *sentimen*.	*Naïve Bayes Classifier*

2.2 Dasar Teori

Pada bagian ini berisi definisi, penjelasan dan uraian yang di peroleh dari berbagai referensi yang di publikasikan pada media sosial berkaitan dengan topik penelitian.

2.2.1 IKN (Ibu Kota Nusantara)

Melalui Naskah Akademik Rancangan Undang-Undang Tentang Ibu Kota Negara yang di susun oleh Surhaso Monoarfa (2020) selaku Menteri Perencanaan Pembangunan Nasional / Kepala Badan Perencanaan Pembangunan Nasional Republik Indonesia, Ibu kota baru Negara Indonesia yang akan berlangsung pada tahun 2024 bertempat di Penajam Paser Utara (PPU) dan sebagian Kutai Kartanegara di Kalimantan Timur memiliki konsep diantaranya :

1. *City Beautiful Movement*, sebagaimana di kemukan oleh Daniel H. Burnham (1910) yang menekankan pada perbaikan kota dengan mempercantiknya (*beautification*).
2. *Radiant city*, melalui pemikiran Le Corbusier (1924) yaitu mewujudkan kota kompak yang terbagi/terseparasikan secara tersusun dengan pola sederhana dan rasional.
3. *Garden city*, dirumuskan oleh Ebenezer Howard (1876) merancang sebuah kawasan menjadi kota yang hidup, energik dan aktif yang dihiasi oleh keindahan dan suasana kawasan desa (pinggiran kota).
4. *Green City*, yaitu konsep pengembangan kota yang ramah lingkungan dengan memanfaatkan sumber daya air dan energi secara efesien.
5. *Eco City*, mengurangi segala jenis polusi buangan gas karbon (*zero carbon activity*), dan mempersatukan harmonisasi kota dengan lingkungan alami.
6. *Smart City*, konsep pengembangan dan pengelolaan kota dengan pemanfaatan TIK (*Teknologi Informasi dan Komunikasi*).
7. *Intelligent City*, konsep *smart city* dengan upaya untuk mengubah komunitas menjadi lebih baik dan kreatif serta terlibat dalam proyek pengembangan komunitas pintar.

Dikutip dari IKN.go.id, menurut Suharso (2022) dalam Rapat Panitia Khusus Ibu Kota Negara yang diselenggarakan pada gedung Dewan Perwakilan Rakyat (DPR), Presiden Joko Widodo telah memilih nama Ibu kota negara baru dengan nama NUSANTARA. Nama NUSANTARA dipilih karena kata tersebut merupakan sebutan (*nama*) seluruh wilayah kepulauan Indonesia dan telah dikenal

sejak sebelum Indonesia merdeka. Presiden Joko Widodo menekankan bahwa Ibu kota Nusantara nantinya tidak hanya sebagai simbol identitas bangsa, tetapi juga sebagai *representasi* kemajuan bangsa.

2.2.2 Media Sosial

Menurut Boyd dalam Nasrullah (2015) media sosial sebagai kumpulan perangkat lunak yang memungkinkan individu maupun komunitas untuk berkumpul, berbagi, berkomunikasi dan dalam kasus tertentu saling memiliki kekuatan pada UGC (*user-generated content*), dimana konten dihasilkan oleh pengguna, bukan oleh *editor* sebagaimana di instansi media massa.

Disisi lain, Van Djik dalam Nasrullah (2015) juga menyatakan bahwa media sosial merupakan *platform* media yang memfokuskan pada eksistensi pengguna yang memfasilitasi mereka dalam beraktifitas maupun berkolaborasi. Karena itu media sosial dapat dilihat sebagai medium (*fasilitator*) online yang menguatkan hubungan antar pengguna sekaligus sebuah ikatan sosial.

Berdasarkan pengertian yang telah diuraikan dari beberapa ahli, pada intinya dengan sosial media dapat dilakukan berbagai aktifitas dua arah dalam berbagai bentuk pertukaran, kolaborasi dan saling berkenalan dalam bentuk tulisan, *visual* maupun *audiovisual.* Sosial media diawali dari tiga hal, yaitu *Sharing, Collaborating dan Connecting* (Puntoadi, 2011).

2.2.3 Twitter

Twitter merupakan jejaring sosial bersimbol burung berwarna biru yang membatasi penggunanya untuk mengirim sebuah *tweet* dengan batas 140 kata, tidak lebih. *Twitter* didirikan oleh Jack Dorsey dan diresmikan pada tahun 2006 tepatnya pada bulan maret. saat ini, *Twitter* sudah sangat dikenal orang di dunia. Alasan mengapa *Twitter* sangat di gemari terlebih di Indonesia, karena *Twitter* menyediakan banyak informasi, dibandingkan dengan aplikasi lain, informasi pada *Twitter* lebih cepat sampai ke masyarakat berkat adanya *fitur trending*. Fitur yang terdapat di *Twitter* diantaranya :

1. *Tweet*, *tweet* merupakan kicauan yaitu untuk mengirim dan melihat *kicauan* setiap pengguna *Twitter*. Pada fitur ini pengguna dapat menyukai *kicaua*n ataupun membalas *kicauan* tersebut.
2. *Retweet*, merupakan fitur untuk memposting kembali sebuah *tweet*. Fitur *retweet* ini dengan cepat membagikan *tweet* kepada semua pengikut pengguna.
3. *Profile*, fitur ini juga merupakan fitur utama dari *Twitter*. Fitur ini dapat melihat *following, followers, avatar twitter, bio twitter, tweet* dan lainnya.
4. *Following*, fitur *following* merupakan fitur untuk mengikuti teman ataupun kerabat di *Twitter*. Dengan fitur ini, kita dapat bertukar informasi dengan melihat *kicauan* pengguna satu sama lain.
5. *Followers*, fitur ini untuk melihat siapa saja yang mengikuti di *Twitter*.
6. *Avatar*,fitur ini menawarkan setiap pengguna untuk dapat membuat karakter *animasi 3D* yang menggambarkan dirinya.

7. *Bio*, merupakan fitur yang digunakan untuk mengetahui pesan singkat pada akun *Twitter* yang terdapat di *profile*.
8. *Private Account*, merupakan fitur penguncian akun. Setiap akun yang dikunci oleh penggunanya, maka akun tersebut tidak dapat dilihat oleh *public* karena terlindungi.
9. *Top Trending*, merupakan fitur yang memperlihatkan penggunanya untuk melihat *kicauan* apa yang paling populer dan sering *dikicaukan* oleh pengguna *Twitter*.

2.2.4 Python

Bagi sebagian orang beranggapan bahwa bahasa pemograman ini diambil berdasarkan nama bintang melata, namun anggapan tersebut salah. Python merupakan bahasa pemograman *interpretatif* multiguna dengan filosofi perancangan yang berfokus pada tingkat keterbacaan kode. Tujuan khusus bahasa pemograman Python ini dikembangkan karena untuk membuat *source code* mudah dibaca. Python juga memiliki *library* yang lengkap sehingga memungkinkan *programmer* untuk membuat aplikasi yang mutakhir dengan menggunakan *source code* yang tampak sederhana (Ljubomir Perkovic, 2012).

2.2.5 Crawling Data

Crawling adalah proses pengambilan data dari media sosial yang kemudian dikumpulkan menjadi satu untuk di evakuasi dan dibentuk agar menjadi sebuah penelitian. Proses *crawling data* terbagi menjadi dua cara, yaitu dengan menggunakan *API* dan tanpa *API* (Tineges, 2021).

Sebagai contoh pengambilan data pada *Twitter* menggunakan *API Key Twitter* dibantu dengan bahasa pemograman Python. *API Key Twitter* merupakan API (*Application Programming Interface*), yang berisi sekumpulan perintah, fungsi, komponen dan juga *protokol* yang disediakan untuk mempermudah program pada saat memiliki membangun suatu sistem perangkat lunak. *API Key Twitter* terdiri dari *consumer keys, consumer secret, acces key,* dan *acces secret.*

Proses pengambilan data pada media sosial *Twitter* dalam penelitian ini menggunakan (*snscrape*). Dengan menggunakan *snscrape,* peneliti dapat menarik data seluruh *tweet_id* untuk *query* search tertentu. Untuk melakukan *crawling data,* maka terlebih dahulu mengimport *snscrape.modules.twitter as sntwitter.*

2.2.6 Analisis *Sentimen*

Setiap pendapat seseorang baik itu perasaan, sikap maupun pikiran yang dapat diungkapkan dikenal sebagai *sentimen.* Analisis *sentimen* merupakan aplikasi *text mining* yang banyak digunakan untuk mengkategorikan *sentimen* terhadap suatu obyek. Menurut Liu (2008), *sentiment analysis* (analisis *sentimen*) atau sering disebut juga dengan *opinion mining* (penambangan opini) adalah studi komputasi untuk mengenali dan mengekspresikan opini, sentimen, evaluasi, sikap, emosi, *subjektifitas,* penilaian atau pandangan yang terdapat dalam suatu teks.

Analisis *sentimen* pada suatu kalimat menggambarkan bagian pertimbangan penilaian terhadap *entitas* atau kejadian tertentu (Pang, dkk. 2008). *Entitas* adalah produk, layanan, topik, isu, orang, organisasi atau peristiwa yang menjadi objek target pada kalimat sentimen (Liu, 2012).

Tugas dasar dalam analisis *sentimen* adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur/tingkat aspek dan menentukan apakah pendapat yang dikemukakan dalam dokumen, kalimat atau fitur *entitas*/aspek bersifat positif, negatif atau netral. Lebih lanjut *sentiment analysis* dapat menyatakan emosional sedih, gembira, atau marah (Liu, 2012). Secara umum, analisis *sentimen* terbagi menjadi tiga jenis, yaitu *document level*, *sentence level* serta *entity and aspect level.*

1. *Document Level*

   Pada level ini adalah tahapan untuk mengklasifikasi data apakah secara keseluruhan opini mengekspresikan sentimen positif, negatif dan netral. Sebagai contoh, pada penelitian ini terdapat sebuah *tweet* pada *Twitter* yang menanggapi suatu persoalan mengenai pemindahan Ibu kota baru negara Indonesia, kemudian sistem akan menentukan secara keseluruhan opini tersebut apakah termasuk positif atau negatif. Jenis analisis ini mengasumsikan setiap dokumen untuk mewakili opini terhadap satu *entitas.*

2. *Sentence Level*

   Pada level ini akan menentukan setiap kalimat menyatakan pendapat positif, negatif atau netral. Netral diartikan sebagai tidak ada opini.

3. *Entity and Aspect Level*

   Pada level ini menentukan sentimen positif, negatif atau netral berdasarkan target (aspek) dalam suatu kalimat opini. Target opini sangat penting karena dapat membantu penganalisa untuk memahami analisis sentimen lebih baik.

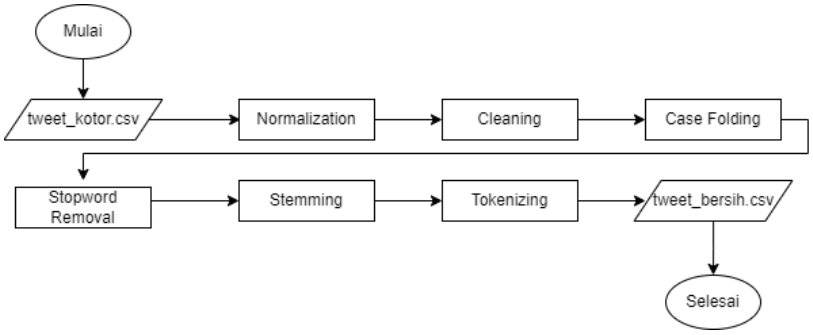
Namun, seringkali analisis level dokumen atau level kalimat tidak menemukan apa yang sebenarnya disukai dan tidak disukai berdasarkan opini tersebut.

2.2.7 Text Mining

Text mining merupakan teknik yang digunakan untuk menangani permasalahan klasifikasi, *clustering*, *information* dan *information retrival* (Berry dan Kogan, 2010). *Text mining* dilakukan pada data berjumlah besar, dimensi besar, struktur teks yang kompleks dan tidak lengkap, serta data yang memiliki *noise* tinggi. Tahapan dalam *text mining* secara umum adalah *text preprocessing*, *feature selection* dan pembobotan (*term weighting*) (Feldman dan Sanger, 2007).

1. Pra-process *(preprocessing)*

Sebelum melakukan analisa *text mining*, terlebih dahulu mempersiapkan teks setelah itu baru digunakan pada proses utama. Proses mempersiapkan teks dokumen disebut dengan proses *text preprocessing*. Fungsi dari *text preprocessing* adalah untuk mengubah data teks dari tidak terstruktur menjadi terstruktur (Sutami, 2015). Tahapan yang dilakukan dari *text preprocessing* pada penelitian ini dapat dilihat pada *Gambar 2.1*



Gambar 2. 1 *Preprocessing*

Tahapan *text preprocessing* pada penelitian ini terdiri dari :

a. Normalization

Tahapan normalisasi dilakukan perubahan kata tidak baku atau terdapat kesalahan ejaannya menjadi kata yang baku dengan menggunakan kamus normalisasi yang dibuat manual berdasarkan pengecekkan data secara keseluruhan. Contoh, “yg” merupakan singkatan dari kata penghubung “yang”.

b. Cleaning

Cleaning merupakan proses membersihkan kata dan karakter pada data yang tidak digunakan untuk mengurangi noise pada proses klarifikasi. Kata yang perlu dibersihkan seperti *username, mention* dan sebagainya. Sedangkan karakter yang dibersihkan pada data seperti (@#$%^&*()_+”:{}<.,?!~/[] !), angka, *link url* (*http://link.com*) *hashtag (#),* dan *mention (@username.*

c. Case Folding

Case Folding merupakan proses penyamaan format huruf dalam sebuah dokumen. Tahapan ini dilakukan karena tidak semua dokumen teks konsisten dalam penggunaan huruf kapital. Oleh karena itu diperlukan *case folding* dalam mengubah keseluruhan teks dalam suatu dokumen dari huruf kapital menjadi bentuk standar berupa huruf kecil (Izzah & Girsang 2021).

d. Stopword Removal

Stopword Removal atau yang disebut dengan *filtering* merupakan proses menghilangkan kata yang tidak memiliki arti dan tidak memberikan dampak pada proses pengklasifikasian *sentimen*. Contoh dari *stopword removal* yakni ke, dari,

pada, antara, dan kata-kata lainnya yang tidak memiliki arti (Izzah & Girsang 2021).

e. *Stemming*

Stemming merupakan proses mengubah kata yang ditempelkan menjadi kata dasar. Contohnya seperti, kata ‘membaca’ setelah melalui proses *stemming* menjadi ‘baca’ (Aribowo et al 2021).

f. *Tokenizing*

Tokenizing merupakan proses untuk membagi kata pada kalimat menjadi beberapa bagian. Contohnya, kalimat ‘semoga lebih baik lagi’ setelah melalui proses *tokenizing* menjadi ‘semoga, lebih, baik, lagi’. (Patel 2021).

2. Pemilihan fitur (*feature selection*)

Data hasil *preprocessing* dapat digunakan untuk tahap pemilihan fitur. Pada tahap *preprocessing*, menghilangkan kata-kata yang tidak relevan disebut *feature selection* (Feldman dan Sanger, 2007). Salah satu fungsi tahap pemilihan fitur (*feature selection*) adalah pemilihan term atau kata yang mewakili pada suatu dokumen yang akan dianalisis dengan melakukan pembobotan terhadap setiap kata/*term*. *Term* dapat berupa kata atau *frase* dalam suatu dokumen yang digunakan untuk mengetahui frekuensi dari dokumen tersebut.

3. Pembobotan Kata (*Term Weighting*)

Pada dasarnya tahap pembobotan dilakukan untuk mendapatkan nilai dari kata/*term* yang telah di ekstrak. *Term* dapat berupa kata atau *frase*. Setiap kata yang memiliki nilai akan menentukan klasifikasi *sentimen*. Pada penelitian ini menggunakan pembobotan data TF-IDF (*Term Frequency Inverse Document*).

a. *Document Frequency*

DF (*Document Frequency*) merupakan jumlah dokumen yang mengandung suatu *term* tertentu.

b. *Inverse Document Frequency*

IDF (*Inverse Document Frequency*) lebih fokus dengan munculnya *term* pada keseluruhan lokasi teks. *Term* yang jarang muncul pada keseluruhan koleksi term dinilai berharga.

c. TF-IDF (*Term Frequency Inverse Document Frequency*)

TF-IDF (*Term Frequency Inverse Document Frequency*) merupakan pembobotan yang dilakukan seelah ekstrasi artikel berita. Proses metode TF-IDF (*Term Frequency Inverse Document Frequency*) adalah dengan menghitung bobot integrasi antara TF(*Term Frequency*) dan IDF (*Inverse Document Frequency*) .

2.2.8 Naïve Bayes Classifier

Naïve Bayes Classifier merupakan pengklasifikasian dengan metode *probabilitas* dan *statis*, dikemukakan oleh ilmuwan Inggris yaitu Thomas Bayes. *Naïve Bayes Classifier* memprediksi peluang dimasa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal dengan *Teorema Bayes*. *Theorema* tersebut dikombinasikan dengan *Naïve* dan diasumsikan kondisi antar atribut saling bebas. Klasifikasi Naïve Bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya (Bustami, 2013).

Ciri utama dari *Naïve Bayes Classifier* ini adalah asumsi yang sangat kuat (*naif*) akan *independensi* dari masing-masing kondisi atau kejadian. Keuntungan menggunakan metode *Naïve Bayes Classifier* dalam penelitian adalah hanya membutuhkan jumlah *data training* yang kecil untuk menentukan *estimasi parameter* yang diperlukan. Karena diasumsikan sebagai *variabel independen*, maka hanya varian dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari *matriks kovarians* (Santoso, 2017). Pengklasifikasian *Naïve Bayes* dilakukan dengan memilih probabilitas akhir (*posterior*) tertinggi dari masing-masing kelas (Simatupang dkk.2016).

Persamaan dari teorema *Bayes* yakni :

$$P(H|X) = \frac{\mathrm{P(X \mid } H). P(X)}{P(X)} \qquad (2.1)$$

Keterangan :

X : Data dengan class yang belum diketahui

H : Hipotesis data X merupakan suatu *class* spesifik

P(H | X) : Probabilitas hipotesis H berdasar kondisi X (*posteriori probability*)

P(H) : Probabilitas hopotesis H (*prior probability*)

P(X | H) : Probabilitas X berdasarkan kepada kondisi pada hipotesis H

P(X) : Probabilitas X

Proses klasifikasi pada *Teoerma Naïve Bayes* memerlukan petunjuk untuk menentukan kelas yang sesuai dengan *sampel* yang dianalisis. Sehingga *Teorema Bayes* seiring dengan perkembangan, menyesuaikan sebagai berikut.

$$P(C|F1,\dots Fn) = \frac{\mathrm{P}(\mathrm{F1},\dots \mathrm{Fn}|\ C)}{P(F1,\dots .Fn)}.P(C) \quad (2.2)$$

Keterangan :

C : Merupakan variabel yang mempresentasikan kelas

F1,.Fn : Mempresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi

Pada rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (*Posterior*) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, yang disebut degan *prior*). Kemudian dikali dengan peluag kemunculan karakteristik sampel pada kelas C (*likelihood*) dan dibagi dengan peluang kemunculan karakteristik *sampel* secara *global*.

$$Posterior = \frac{\text{prior } x\ likehood}{evidence} \quad (2.3)$$

Keterangan :

Evidence : Merupakan nilai yang selalu tetap untuk setiap kelas pada satu sampel

Posterior : Nilai *posterior* akan dibandingkan dengan nilai *posterior* kelas lainnya untuk menentukan kelas apa *sampel* akan diklasifikasi.

Penjabaran mengenai rumus *Bayes* dilakukan dengan menjabarkan (C, F1,..., Fn)

$$P(C|F1,\dots Fn) = \mathrm{P}(\mathrm{C}) \prod_{i=1}^{n} P(Fi|C) \quad (2.4)$$

Probabilitas P(C|F1), ...P(C|Fn) dapat dihitung dari data latih yang ada.Fi melambangkan nilai atribut Fi dari data C. Atribut dapat berupa nilai kategorikal atau *kontinyu*. Untuk klasifikasi dengan data kontinyu digunakan rumus *Densitas Gauss* :

$$P(Xi = xi \mid Y = \text{yi}) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(\text{xi}-\mu)^2}{2\sigma^2}} \quad (2.5)$$

Keterangan :

P : Peluang

Xi : Atribut ke-i

xi : Nilai atribut ke-i

Y : Kelas yang dicari

yi : Sub kelas yang dicari

μ : *Mean*, menyatakan rata-rata dari seluruh atribut

σ : *Deviasi Standar*, menyatakan varian dari seluruh atribut.

2.2.9 SMOTE (*Synthetic Minority Over-Sampling Technique*)

Dalam mengklasifikasikan sebuah dataset dengan jumlah yang banyak, tentu tidak selamanya dapat berjalan dengan lancar tanpa adanya sebuah permasalahan. Salah satu permasalahan umum yang terjadi adalah ketidakseimbangan data (*Imbalance dataset*), dimana terdapat rasio yang tidak *proporsional* disetiap kelas. Ketidakseimbangan kelas dapat ditemukan diberbagai bidang termasuk penyaringan spam email, deteksi penipuan, dan lainnya. Pada

data yang tidak seimbang, hampir semua algoritma klasifikasi akan menghasilkan akurasi yang jauh lebih tinggi untuk kelas *mayoritas* daripada kelas *minoritas*.

Metode SMOTE (*Synthetic Minority Over-Sampling Technique*) merupakan metode yang menangani ketidakseimbangan kelas. Teknik ini *mensintesis sampel* baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara *sampling* ulang sampel kelas *minoritas*. Kelas *minoritas* keadaan kelas yang memiliki jumlah data lebih sedikit.

2.2.10 Pengukuran Kualitas Klasifikasi

Untuk mengukur suatu kualitas klasifikasi data, adalah dengan cara membandingkan nilai *aktual* (nilai sebenarnya) dan nilai prediksi. Penelitian ini menggunakan *confusion matrix* untuk mengukur kualitas klasifikasi. *Confusion Matrix* menurut Han dan Kamber (2011) diartikan sebagai suatu alat yang memiliki fungsi untuk melakukan analisis apakah classifier tersebut baik dalam mengenal *tuple* dari kelas yang berbeda.

Nilai dari *True-Positif* dan *True-Negatif* memberikan informasi ketika *classifier* dalam melakukan klasifikasi data bernilai benar, sedangkan *False-Positive* dan *False-Negative* memberikan informasi ketika *classifier* salah dalam melakukan klasifikasi data (Han dan Kamber, 2011).

Pada *confusion matrix* berisikan informasi mengenai hasil klasifikasi *aktual* yang telah di prediksi oleh sistem klasifikasi dan dievaluasi menggunakan data dalam sebuah *matrix.*

Tabel 2. 3 Visualisasi *Confusion Matrix*

		Kelas Hasil Prediksi	
		Positif	Negatif
Kelas sebenarnya	Positif	TP (True)	FN (False Negatif)
	Negatif	FP (False)	TN (True Negatif)

Keterangan :

1. TP (*True Positif*) merupakan data kelas positif yang terdeteksi benar.
2. FP (*False Positif*) merupakan data kelas negatif yang terdeteksi sebagai data positif.
3. FN (*False Negatif*) merupakan data kelas positif yang salah diklasifikasikan sehingga tergolong ke dalam data negatif.
4. TN (*True Negatif*) merupakan data kelas negatif yang terdeteksi benar.

Berdasarkan nilai TP (*True Positif*), FP (*False Positif*), FN (*False Negatif*) dan TN (*True Negatif*), maka dapat diperoleh nilai TPR (*True Positive Rate*) , FPR (*Fallout atau False Positive Rate*) , *Precision, Recall, F-Measure, MCC,* ROC (*Receiver Operationg Characteristics*), PRC (*Presisi Recall*).

1. TPR (*True Positive Rate*), disebut juga dengan *Recall* atau *Sensitivity* bekerja dengan cara menunjukkan perbandingan data positif yang diklasifikasikan dengan benar ke dalam kelas positif terhadap seluruh data dari kelas positif.

$$TPR = \frac{\text{TP}}{TP + FN} \tag{2.6}$$

2. FPR (*Fallout atau False Positive Rate*), menunjukkan proporsi data negatif yang masuk kedalam prediksi positif.

$$FPR = \frac{\text{FP}}{FP + TN} \tag{2.7}$$

3. *Precision*, merupakan tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

$$Precision = \frac{\text{TP}}{TP + FP} \tag{2.8}$$

4. *Recall*, merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

$$Recall = \frac{\text{TP}}{TP + FN} \tag{2.9}$$

5. *F-Measure*, merupakan perhitungan evaluasi dalam temu kembali informasi yang mengkombinasikan *recall* dan *precision*. Nilai *recall* dan *precision* pada suatu keadaan dapat memiliki bobot yang berbeda. Ukuran yang menampilkan timbal balik antar *recall* dan *precision* adalah *F-Measure* yang merupakan *bobot harmonic mean* dari *recall* dan *precision*.

$$F1 = 2 \text{ x } \frac{\text{Precision x Recall}}{Precision + Recall} \tag{2.10}$$

6. *Accuracy*, merupakan tingkat kedekatan antara nilai prediksi dengan nilai *aktual* (kelas sebenarnya).

$$Accuracy = \frac{\text{TP} + \text{TN}}{TP + TN + FP + FN} \tag{2.11}$$

7. *MCC*, digunakan dalam pembelajaran mesin sebagai ukuran kualitas klasifikasi biner (dua kelas). Dalam hal ini memperhitungkan positif dan negatif benar dan salah dan umumnya dianggap sebagai ukuran seimbang yang dapat digunakan bahkan jika kelas memiliki ukuran yang sangat berbeda.
8. ROC (*Receiver Operationg Characteristics*), merupakan nilai terpenting yang dikeluarkan oleh WEKA. WEKA memberikan gambaran tentang bagaimana kinerja pengklasifikasian secara umum.
9. *Error rate* atau *Misclassification Rate*, menunjukkan tingkat kesalahan klasifikasi.

$$\text{Error Rate} = 1 - \text{Accuracy} \tag{2.12}$$

2.2.11 Jupyter Notebook

Jupyter Notebook merupakan singkatan dari tiga bahasa pemrograman yaitu Ju (*Julia*), Py (*Python*), dan *R*. Jupyter Notebook juga merupakan *tools* yang populer digunakan untuk melakukan pengolahan data bagi seorang *data scientist* yang memungkinkan untuk mengintegrasikan antara kode dengan *output* didalam satu dokumen secara interaktif yang berisi *live code*, persamaan, *visualisasi* dan teks naratif. Terdapat sejumlah kolom menu pada Jupyter Notebook, diantaranya :

1. File

Fungsi menu *File* adalah membuat notebook baru atau membuka notebook yang telah dibuat sebelumnya. Tersedia juga fitur *Save* and *Checkpoint* untuk

membuat *checkpoint* yang akan kembali ke titik terakhir jika terjadi sesuatu yang tidak diinginkan.

2. View

Fungsi menu *View* adalah mengaktifkan atau mematikan tampilan *header* dan *toolbar*. Pada menu *View* ini, pengguna dapat menyalakan atau mematikan *Line Numbers* didalam *cells*.

3. Edit

Menu *Edit* memiliki fungsi bagi pengguna untuk dapat melakukan *cut, copy* atau *paste* dari *cell* yang tersedia. Selain itu, menu *edit* pada Jupyter Notebook ini berfungsi membagi, menyatukan dan menghapus *cell*.

4. Insert

Fungsi dari menu *Insert* adalah memasukkan *cell* diatas atau dibawah *cell* yang dipilih.

5. Kernel

Menu *Kernel* digunakan untuk mengerjakan *kernel* atau bahasa pemograman yang berjalan.

6. Cell

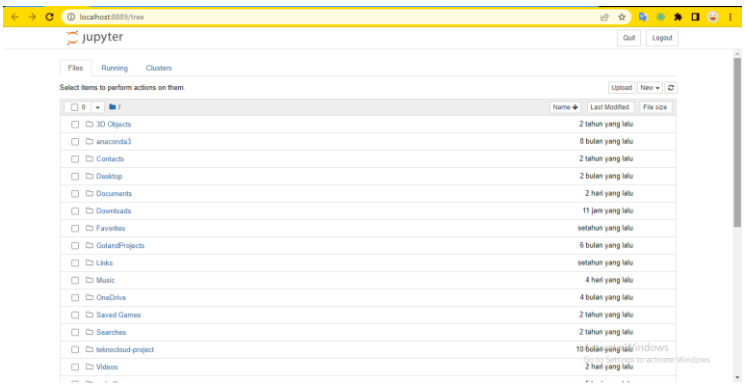
Fungsi menu *Cell* untuk menjalankan satu, beberapa atau bahkan seluruh *cell* yang tersedia.

7. Widget

Menu *Widget* berguna untuk menambah atau menghapus *widget* yang ada. *Widget* merupakan *JavaScript* yang dipakai untuk membuat konten *cell* menjadi dinamis.

8. Help

Menu terakhir pada Jupyter Notebook adalah *Help*. *Help* berfungsi bagi pengguna apabila memerlukan bantuan atau ingin mempelajari *notebook* lebih dalam. Berbagai hal seputar *keyboard shortcut*, *UI* hingga refensi materi ada dalam menu *help*.



Gambar 2.2 Jupyter Notebook

2.2.12 WEKA (*Waikato Environment for Knowledge Analysis*)

WEKA (*Waikato Environment for Knowledge Analysis*) merupakan sebuah perangkat lunak yang menerapkan berbagai *algoritma machine learning* untuk melakukan beberapa proses yang berkaitan dengan sistem temu kembali informasi atau data mining. Pada WEKA (*Waikato Environment for Knowledge Analysis*) memiliki beberapa fitur unggulan, diantaranya :

1. Classification

Pada WEKA (*Waikato Environment for Knowledge Analysis*), terdapat banyak algoritma yang mendukung untuk proses klasifikasi sebuah objek. Pengguna dapat melakukan *load dataset*, pemilihan algoritma untuk klasifikasi,

kemudian akan diberikan beberapa representasi data yang mewakili hasil akurasi, tingkat kesalahan dari proses klarifikasi.

2. Regression

Regression merupakan proses yang dapat melakukan prediksi terhadap berbagai pola yang sudah terbentuk sebelumnya yang dijadikan sebagai model data. Tujuan dari *regression* adalah menciptakan suatu variabel baru yang mewakili suatu representasi perkembangan data pada masa yang akan datang.

3. Clustering

Clustering bertujuan untuk mengelompokkan data dan juga menjelaskan hubungan/relasi yang ada diantara data tersebut serta memaksimalkan kesamaan antar satu kelas/*cluster*.

4. Association Rules

Association Rules merupakan metode yang digunakan untuk menemukan berbagai relasi antara banyaknya variabel yang terdapat didalam sebuah basis data dengan jumlah yang besar.

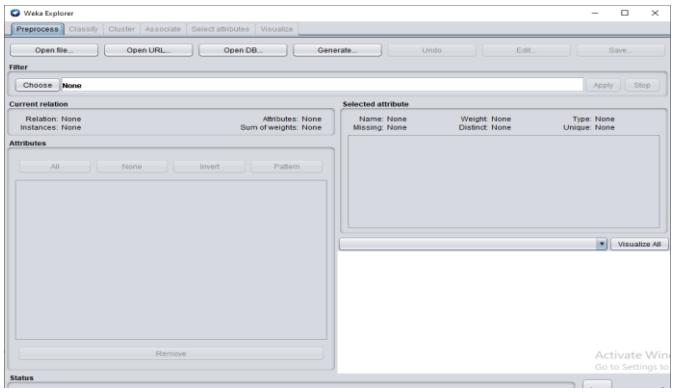
5. Visualization

Visualization dalam WEKA (*Waikato Environment for Knowledge Analysis*) pada proses *data mining*, menghasilkan bentuk gambar atau *chart*.

6. Data Preprocessing

WEKA (*Waikato Environment for Knowledge Analysis*) menyediakan fitur untuk melakukan data *preprocessing* yakni *stemming* dan *stopword removal*. Namun, pada proses *stemming* dan *stopword* yang ada pada WEKA (*Waikato Environment for Knowledge Analysis*) berbasis bahasa Inggris. Sehingga apabila

ingin implementasi menggunakan bahasa di luar bahasa Inggris, diharuskan untuk melakukan proses *preprocessing* data diluar WEKA (*Waikato Environment for Knowledge Analysis*).



Gambar 2.3 Aplikasi (Software) *WEKA*