BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pusataka

Agar dapat membedakan penelitian yang dilakukan dengan penelitian sebelumnya maka peneliti melakukan studi pustaka yang telah dilakukan peneliti - peneliti terdahulu, adapun tinjauan pustaka terdahulu.

Tinjauan pertama merupakan penelitian yang ditulis oleh Budi Kurniawan, Mochammad Ali Fauzi dan Agus Wahyu Widodo dengan judul Klasifikasi Berita Twitter Menggunakan Metode Improved Naïve Bayes pada tahun 2017. Data diambil berasal dari twitter yang kemudian berita diklasifikasikan menjadi 8 yaitu ekonomi, entertainment, olahraga, teknologi, kesehatan, makanan, otomotif, travel. Pada proses klasifikasi dilakukan mulai preprocessing kemudian proses *case folding, tokenizing, filtering dan stemming*. Setelah Proses preprocessing selesai dilakukan perhitungan kemunculan kata dan probabilitas, kata yang muncul akan 1 sedangkan yang tidak akan diberi nilai 0. Setelah itu baru dilakukan klasifikasi. Jumlah data latih mempengaruhi performa, semakin banyak data latih semakin baik hasilnya.

Tinjauan yang kedua penelitian yang ditulis oleh Dio Ariadi dan Kartika Fithriasari dengan judul Klasifikasi Berita Indonesia Menggunakan Metode *Naive Bayesian Classification* dan *Support Vector Machine* dengan *Confix Stripping Stemmer* yang ditulis pada tahun 2015. Data yang diklasifikasikan adalah berita bahasa Indonesia dengan sumber data adalah artikel berita dari koran online kompas.com yang akan dibagi menjadi 12 kategori antara lain berita nasional, internasional, olahraga, sains, edukasi, ekonomi, tekno, entertaintment, otomotif, health, properti, dan travel. Artikel diambil dari Januari hingga Desember 2014, yang kemudian data tersebut dibagi menjadi data *traning* dan data *testing* dengan porsi 70:30.

Tinjauan ketiga penelitian yang ditulis oleh Rusdi Efendi, Reza Firsandaya Malik, Jeni Mila Sari U pada tahun 2012 dengan judul Klasifikasi Dokumen Berbahasa Indonesia Menggunakan Naive Bayes Classifier. Objek yang diteliti adalah dokumen berbahasa Indonesia dengan 30 dokumen latih dan 30 dokumen uji yang kemudian akan

diklasifikasikan menjadi 6 katagori yaitu, ekonomi, kesehatan, olahraga, teknologi, politik dan pendidikan.

Tinjauan yang keempat ditulis oleh Naufal Riza Fatahillah pada tahun 2017 dengan judul penelitian Implementasi Algoritma Naive Bayes Classifier Pada Sosial Media (Twiiter) Terhadap Ujaran Kebencian Berbahasa Indonesia. Dalam proses pencariannya didasarkan hastag tertentu yang selanjutnya diklasifikasikan negatif atau positif.

Tinjauan kelima ditulis oleh Septian Narsa Putra pada tahun 2019 dengan judul Klasifikasi Topik Berita dan Analisis Sentimen pada tweets Divisi Humas Polri dengan Metode Naïve Bayes Classifier. Pada penelitian tersebut menerapkan proses text mining serta menggunakan algoritma *Naïve Bayes Classifier* untuk mengklasifikasi topik dan sentiment dari *tweets* tersebut. Penelitian ini menggunakan bantuan API twitter untuk proses pengambilan *tweets* untuk keperluan data training dan data testing. Dalam penelitian tersebut dibagi menjadi tiga kelas topic yang memiliki sentiment disetiap topiknya yaitu Positf, negative dan netral.

Adapun pada penelitian ini, sistem akan melakukan klasifikasi terhadap berita yang diupload pada akun twitter dengan topik berita Komponen Cadangan Kementrian Pertahanan dan Keamanan RI kemudian dilakukan analisis untuk mengetahui sentimen dari topik berita tersebut yang dibagi menjadi berita positif dan berita negatif. Metode yang digunakan adalah *Naive Bayes Classifier*.

Tabel 2. 1 : Tinjauan Pustaka 1

Penulis	**Objek**	**Metode**	**Hasil**
Budi Kurniawan, Mochammad Ali Fauzi, Agus Wahyu Widodo (2017)	Berita Twitter	Metode Improved Naïve Bayes	Aplikasi ini mampu mengkasifikasikan delapan kategori yaitu: ekonomi, entertainment, olahraga, teknologi, kesehatan, makanan, otomotif, travel. Data yang diambil berupa judul dari berita.

Dio Ariadi dan Kartika Fithriasari (2015)	Berita bahasa Indonesia	Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer	Berita dikategorikan sebagai berikut berita nasional, internasional, olahraga, sains, edukasi, ekonomi, tekno, entertaintment, otomotif, health, properti, dan travel. Tiap kategori akan diambil sebanyak 100 artikel sehingga data artikel keseluruhan berjumlah 1200. Hasil yang didapatkan pada saat data testing pada masing-masing pengukuran performa akurasi, precision, recall, dan F-Measure sebesar 82,2%; 83,9%; 82,2%; dan 82,4%.
Rusdi Efendi, Reza Firsandaya Malik, Jeni Mila Sari U (2012)	Dokumen Berbahasa Indonesia	Naive Bayes Classifier	Perangkat lunak pada penelitian ini telah dapat mengklasifikasikan dokumen berita dengan akurasi perangkat lunak ini mencapai 86,67% dengan menggunakan 60 dokumen yang terdiri dari 30 dokumen latih dan 30 dokumen uji untuk enam kategori berita yaitu, ekonomi,

			kesehatan, olahraga, teknologi, politik dan pendidikan.
Naufal Riza Fatahillah (2017)	Tweet pada twitter	Naive Bayes Classifier	Mendeteksi tweets negatif atau positif berdasarkan hastag.
Septian Narsa Putra (2019)	Klasifikasi Topik Berita dan Analisis Sntimen pada tweets Divisi Humas Polri	Naive Bayes Classifier	Mengklasifikasikan berita berdasarkan topik kemudian dicari sentimen dari setiap topik.
Fajriyah,Nurul (2021)	Klasifikasi Beritan dan Analisis Sentimen Terhadap Komponen Cadangan Kementrian Pertahanan RI Menggunakan Metode *Naïve Bayes Classifier*	Naive Bayes Classifier	Sistem yang akan dibuat mampu mengklasifikasi topik berita menggunakan data tweets pada media social Twitter, kemudian melakukan analisis topik berita untuk memperoleh hasil sentimennya yang terbagi menjadi 2 kategori yaitu positif dan negatif. Adapun topik berita yang diambil yaitu

			berita terkait komponen cadangan kementerian pertahanan dan keamanan republik indonesia.

2.2 Dasar Teori

2.2.1 Machine Learning

Menurut Jhon McCarthy, salah satu pakar AI atau *Artifical Intelegence*, AI adalah "*the science and engineering of making intelligent machiness*". Beberapa definisi AI yang lain ialah:

- Sebuah cabang ilmu komputer (Computer Science) yang berkaitan dengan bagaimana mensimulasikan kecerdasan pada komputer.
- Kemampuan mesin untuk meniru kecerdasan manusia.
- Sebuah sistem yang mampu melakukan pekerjaan dimana pekerjaan tersebut memerlukan kecerdasan manusia, seperti visual perception, speech recognition, decision-making, dan penerjemah bahasa.

Machine Learning pada mulanya diperkenalkan oleh Arthur Samuel pada tahun 1959 melalui jurnal yang berjudul "Some Studies in Machine Learning Using the Game of Checkers".(IBM Journal of research and Development).

Machine learning memerlukan sebuah model yang didefinisikan berdasarkan parameter-parameter tertentu. Proses learning adalah eksekusi program komputer untuk mengoptimasi parameter-paramater dari model tersebut, dengan memanfaatkan data training atau past experience.

Machine Learning mempelajari teori agar komputer mampu "belajar" dari data. Machine Learning melibatkan berbagai disiplin ilmu seperti statistika, ilmu komputer, matematika bahkan neorologi. Machine Learning menggunakan teori-teori statistika untuk membentuk model matematika, model dapat bersifat *predictive* (untuk memprediksi masa depan) atau *descriptive* (untuk memperoleh pengetahuan dari data), atau juga bisa gabungan dari *predictive* dan *descriptive*. (Rifkie Primartha, 2018)

Machine Learning adalah ilmu yang mempelajari tentang algoritma komputer yang bisa mengenali pola-pola didalam data, dengan tujuan mengubah beragam macam data menjadi suatu tindakan yang nyata dengan sesedikit mungkin campur tangaan manusia. Dengan *Machine Learning*, kita dapat menciptakan mesin (komputer) yang "belajar" dari data yang ada, selanjutnya dia bisa membuat keputusan secara mandiri tanpa perlu

diprogram lagi. Secara umum, *Machine Learning* berada dibawah payung ***Artifical Intelegence/AI***, (Kecerdasan Buatan).

Inti dari Machine Learning adalah bagaimana membuat komputer dapat menyelesaikan berbagai persoalan dan dapat belajar sendiri seperti manusia belajar sesuatu. Adapun secara umum algoritma machine learning dapat dikelompokkan menjadi 4 kategori, yaitu :

1. *Supervised Learning*

   Pada *supervice learning* menggunakan sebuah algoritma untuk mempelajari mapping function antara input dengan output. *Superviced learning* dapat bermanfaat untuk memprediksi sesuatu dengan bantuan training dataset. Ada beberapa algoritma yang sudah dikembangkan dan terkait dengan *superviced learning* diantaranya :

   - *Decision tree*
   - *Naïve Bayes Classifier*
   - *Artificial Neural Network*
   - *Support Vector Machine*
   - *Linear Regression*
   - *Logistic Regression*
   - *CART*
   - *KNN (K-Nearest Neighbor)*
   - *dsb*

2. *Unsupervised Learning*

   Unsupervised Learning bermanfaat untuk kasus-kasus dimana kita ingin menemukan relasi implisit (*implicit relationships*) dari unleabeled dataset yang disediakan. Beberapa algoritma yang dikelompokkan dalam *unsuperviced learning* antara lain:

- *K-Means*
- *Hierarchical Clustering*
- *DBSCAN*
- *Fuzzy C-Means*
- *Self-Organizing Map*
- *dsb*

3. *Reinforcement Learning*

   Reinforcement learning merupakan metode belajar yang dipengaruhi *feedback* dari lingkungan teknik learning yang *iterative* (berulang-ulang) dan adaptive (menyesuaikan). Beberapa algoritma yang dikelompokkan dalam *reinforcment learning* yaitu :

   - *Genetic Algorithm (GA)*
   - *Dynamic Programing (DP)*
   - *Generalized Policy Iteration (GPI)*
   - *Monte Carlo*

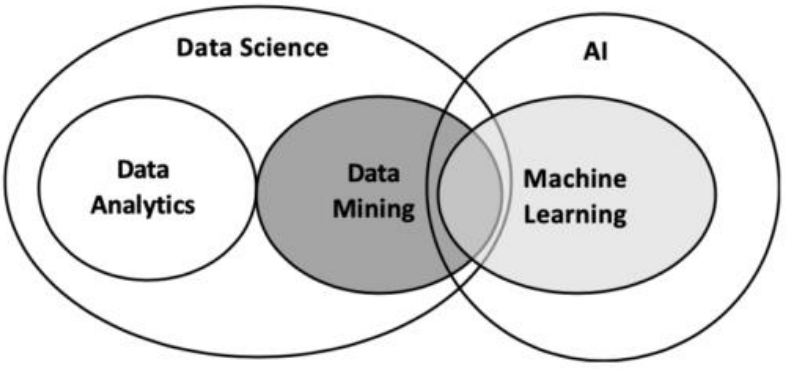
4. *Deep Learning*

   Deep learning merupakan metode belajar yang memanfaatkan *artifical neural networks* yang berlapis-lapis (multi-layer). *Artifical neural networks* ini dibuat mirip dengan otak manusia, dimana neuron-neuron terkoneksi satu sama lain sehingga membentuk sebuah jaringan neuron yang sangat rumit. Beberapa algoritma yang termasuk dalam kategori *deep learning* antara lain :

   - *Convolutional Networks*
   - *Restricted Boltzmann Machine (RBM)*
   - *Deep Belief Network (DBN)*
   - *Stacked Autoencoders*

2.2.2 Data Science

Data science merupakan ilmu multidisipliner yang memepelajari tentang upaya mendapatkan pemahaman yang lebih dari berbagai macam data. Tujuannya agar didapatkan kesimpulan dari informasi yang terkandung di dalam data sehingga orang dapat mengambil keputusan dan tindakan yang tepat. Adapula yang menyebut ilmu ini dengan data istilah ***Data Analytics***, dimana batas pemisahannya memang tidak terdefinisi dengan jelas. Pada dasarnya ***Data Analytics*** lebih difokuskan pada memecahkan permasalahan atau menjawab pertanyaan yang lebih spesifik . *Data Analytics* berusaha memberikan kesimpulan dari apa yang telah diketahui sebelumnya. berbeda dengan *Data Science* yang berupaya menghasilkan pemahaman lebih luas, yaitu menemukan hal-hal penting apa saja yang terkandung di dalam data, dan hal-hal yang potensial bisa dijadikan pertanyaan lebih lanjut. *Data Science* lebih sering digunakan untuk monetisasi, artinya yang mendapatka keuntungan finansial, karena informasi yang bernilai tinggi dapat menghasilkan uang. (Dios Kurniawan, 2019).



Gambar 2. 1Ruang Lingkup Data Science

2.2.3 Sentiment Analysis

Berdasarkan Kamus Besar Bahasa Indonesia, sentimen merupakan pendapat atau pandangan yang didasarkan pada perasaan yang berlebih-lebihan terhadap sesuatu. Analisis sentimen (sentiment analysis) atau opinion mining adalah penggalian konteks

pada teks untuk mengidentifikasi informasi subjektif dan membantu pelaku bisnis untuk memahami sentimen sosial terhadap brand, produk, atau jasanya dengan mengamati percakapan online.

Sentimen Analisys merupakan proses memahami atau mengekstrak dan mengolah data textual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam satu kalimat opini apakah cenderung positif atau negatif.

Analisa sentimen atau yang biasa disebut opinion mining merupakan salah satu cabang penelitian Text Mining. Opinion mining adalah riset komputasional dari opini, sentimen dan emosi yang diekspresikan secara tekstual. Jika diberika suatu set dokumen teks yang berisi opini mengenai suatu obyek, maka opinion mining bertujuan mengekstrak atribut dan komponen dari obyek yang telah diberikan komentar pada setiap dokumen dan untuk menentukan apakah komentar tersebut bermakna negatif atau positif (Fahlah, 2015).

Berdasarkan Liu (2012), analisis sentimen dibagi menjadi tiga kategori berdasarkan levelnya, yaitu:

- Document level
- di mana tugas pada level dokumen adalah mengklasifikasikan apakah keseluruhan dokumen memiliki sentimen positif atau negatif.
- Sentence level, di mana tugas pada level kalimat adalah mengklasifikasikan apakah setiap kalimat memiliki sentimen positif, negatif, atau netral. Netral berarti tidak ada opini apa pun.
- *Entity* and *aspect level* atau dapat disebut *feature level*, adalah level di mana yang diklasifikasikan adalah opininya.

Hal ini didasarkan oleh ide bahwa opini terdiri dari sebuah sentimen (positif atau negatif) dan sebuah target opini. Selain itu, analisis sentimen juga dapat dibagi menjadi dua kategori umum, yaitu:

- *Coarse-grained sentiment analysis*, yaitu proses analisis pada level dokumen secara keseluruhan.
- *Fined-grained sentiment analysis*, yaitu proses analisis pada level kalimat.

2.2.4 Twitter

Twitter merupakan salah satu media sosial dengan layanan *microbloging* yang terkenal dimana memungkinkan para penggunanya untuk menulis sesuatu atau yang

biasa disebut tweet. Twitter juga digunakan untuk mengutarakan opini publik maupun berita resmi dari suatu instansi atau dari akun kantor berita. Twitter dibangun oleh Jack Dorsey pada tahun 2006 dengan alamat http://www.twitter.com, jika seseorang ingin menggunakan twiiter seseorang harus terlebih dahulu memiliki akun, untuk registrasinya dapat dilakukan pada alamat tersebut dengan mendaftarkan alamat e-mail atau nomor telfon kemudian memasukkan password untuk kebutuhan keamanan akun.

Sistem situs ini berbasis pesan pendek yang ditampilkan pada profil pengguna dan dikirimkan pada pengguna lain yang telah menjalin pertemanan yang disebut dengan "followers" atau " pengikut" (yudha,2018). Pengguna dapat menulis pesan berdasarkan topik dengan tanda #(tagar). Sedangkan untuk menyebut atau membalas pesan dari pengguna lain bisa menggunakan tanda @(diikuti nama akun yang akan dibalas).

Layanan pada situs Twitter di antaranya :

1. Twipic, yakni seputar aplikasi untuk mengupload foto dan otomatis memberitahu teman lewat postingan di Twitter.
2. Twitterific dan *Tweetie* adalah aplikasi iphone yang memungkinkan pengguna menampilkan dan posting ke Twitter lewat iphone.
3. Twitter Gadget oleh Google, antar muka desktop bagi Twitter.

2.2.5 Crawling Data Twitter

Crawling adalah proses pengambilan sejumlah besar halaman web dengan cepat ke dalam suatu tempat penyimpanan lokal dan mengindeksnya berdasar sejumlah kata kunci. *Crawling* data di *twitter* dapat menggunakan dua sistem pencarian, by user dan by keyword. Pencarian menggunakan by keyword yaitu pencarian menggunakan penggalan kata maupun hashtag. Sedangkan pencarian dengan by user yaitu pencarian berdasarkan nama akun user twitter (Sembodo, dkk. 2016).

Dalam penelitian ini untuk memperoleh data tersebut menggunakan library yang telah di sediakan oleh bahasa pemrograman python yaitu *snscraper*. *Snscraper* merupakan salah satu library yang dimiliki oleh bahasa pemrogramman Python yang digunakan untuk mengambil data.

Dirilis pada 8 Juli 2020, *snscrape* adalah library yang digunakan untuk mengakses informasi untuk media sosial, salah satunya twitter. Snscrape dapat memperoleh informasi seperti profil pengguna, tagar, serta tweet yang dapat

ditemukan tanpa menggunakan API Twitter. Snscrape memudahkan pengguna untuk mengambil data tweet tanpa batasan secara real time.

Data tweet yang telah dikumpulkan akan diseleksi mana yang cocok untuk dijadikan data training maupun data prediksi testing karena tidak semua data yang telah dikumpulkan dapat digunakan.

2.2.6 Text Mining

Text Mining adalah proses ekstraksi pola (informasi dan pengetahuan yang berguna) dari sejumlah besar sumber data tak terstruktur. Penambangan teks memiliki tujuan dan menggunakan proses yang sama dengan penambangan data, namun memiliki masukan yang berbeda. Masukan untuk penambangan teks adalah data yang tidak (atau kurang) terstruktur, sedangkan masukan untuk penambangan data adalah data yang terstruktur (Ronen Feldman, 2006). *Text mining* dapat diterapkan untuk pelacakan topik, katagorisasi dan analisis sentimen.

Text mining cenderung mengarah pada bidang penelitian data mining. Data Mining merupakan gabungan sejumlah disiplin ilmu komputer, yang didefinisikan sebagai proses penemuan pola-pola baru dari kumpulan-kumpulan data sangat besar, meliputi metode-metode yang merupakan irisan dari *artificial intelligence, machine learning, statistic,* dan *database system*. Secara umum, kegunaan data mining dapat dibagi menjadi dua yaitu deskriptif dan prediktif. Deskriptif berarti data mining digunakan untuk mencari polapola yang dapat dipahami manusia yang menjelaskan karakteristik data. Sedangkan prediktif berarti data mining digunakan untuk membentuk sebuah model pengetahuan yang akan digunakan untuk melakukan prediksi. Oleh karena itu, tidak mengherankan bahwa *text mining* dan *data mining* berada pada tingkat arsitektur yang sama. Penambangan teks dapat dianggap sebagai proses dua tahap yang diawali dengan penerapan struktur terhadap sumber data teks dan dilanjutkan dengan ekkstraksi informasi dan pengetahuan yang relevan dari data teks terstruktur ini menggunakan teknik dan alat yang sama dengan penambangan data (Fahlahah,2015).

Dalam proses penambangan teks terdapat proses pengubahan data yang belum terstruktur menjadi data yang terstruktur seesuai dengan kebutuhan, proses ini disebut *Text Prepocessing*, dengan tahapannya adalah *Case Folding, Tokenizing, Filtering, Stemming, Tagging* dan *Analyzing*.

Case Folding adalah proses mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf "a" sampai dengan "z" yang diterima dan karakter selain huruf dihilangkan. Tahapan *Tokenizing* adalah proses pemotongan kalimat (*string*) berdasarkan kata peyusunnya. Kemudian tahap *filtering*, merupakan tahapan mengambil kata - kata penting dari hasil token dan membuang sebagian kata tertentu. Tahapan *Stemming* merupakan tahapan pemetaan dan penguraian suatu kata menjadi kata dasarnya. Tahapan *tagging* merupakan tahap memberikan label dari setiap kategori. Yang terakhir adalah tahap *analyzing* yaitu tahap untuk mencari seberapa jauh keterhubungan antar kata - kata setiap dokumen.

Kunci dari proses pada text mining adalah menggabungkan informasi yang berhasil diekstraksi dari berbagai sumber (Hearst, 2003). Sedangkan menurut (Harlian, 2006) text mining didefinisikan sebagai kata yang berupa text yang biasanya sumber data didapatkan dari dokumen, dengan tujuan adalah mencari kata-kata yang dapat mewakili isi dari dokumen tersebut yang nantinya dapat dilakukan analisa hubungan antar dokumen. Proses text mining yang khas meliputi kategorisasi teks, text clustering, ektraksi konsep/entitas, produksi taksonomi granular, sentiment analysis, penyimpulan dokumen, dan pemodelan relasi entitas yaitu, pembelajaran hubungan antara entitas (Bridge, 2011). Pendekatan manual text mining secara intensif dalam laboratorium pertama muncul pada pertengahan 1980-an, namun kemajuan teknologi telah memungkinkan ranah tersebut untuk berkembang selama decade terakhir.

Text mining adalah bidang interdisipliner yang mengacu pada pencarian informasi, pertambangan data, pembelajaran mesin, statistik, dan komputasi linguistic. Dikarenakan kebanyakan informasi (perkiraa umum mengatakan lehih dari 80%) saat ini disimpan sebagai teks, text mining diyakini memiliki potensi nilai komersial tinggi (Bridge, 2011).

2.2.7 Klasifikasi

Klasifikasi merupakan kata serapan dari bahasa Belanda, *classificatie*, yang sendirinya berasal dari bahasa Prancis *classification*. Istilah ini menunjuk kepada sebuah metode untuk menyusun data secara sistematis atau menurut beberapa aturan atau kaidah yang telah ditetapkan.

Bagian sangat penting dalam data mining adalah teknik klasifikasi, klasifikasi dapat didefinisikan sebagai proses untuk menyatakan suatu objek data sebagai salah satu kategori (kelas) yang telah didefinisikan sebelumnya (Zaki et al.2013).

Di dalam KBBI, klasifikasi adalah penyusunan bersistem dalam kelompok atau golongan menurut kaidah atau standar yang ditetapkan. Secara harfiah bisa pula dikatakan bahwa klasifikasi adalah pembagian sesuatu menurut kelas-kelas. Menurut Ilmu Pengetahuan, Klasifikasi adalah Proses pengelompokkan benda berdasarkan ciri-ciri persamaan dan perbedaan.

Menurut Prasetyo (2012), klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya kedalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasiifikasi terdapat dua proses yang dilakukan yaitu dengan membangun model untuk disimpan sebagai memori dan menggunakan model tersebut untuk melakukan pengenalan dan klasifikasi atau prediksi pada suatu data lain supaya diketahui dikelas mana objek data tersebut dimasukkan berdasarkan model yang telah disimpan dalam memori (Riyan, 2014).

2.2.8 *Naive Bayes Classifier*

Algoritma *Naïve Bayes Classifier* merupakan salah satu pengklasifikasi statistik, dimana pengklasifikasi ini dapat memprediksi probabilitas keanggotaan kelas suatu data yang akan masuk ke dalam kelas tertentu, sesuai dengan perhitungan probabilitas. Pengklasifikasi bayes didasari oleh *theorema bayes* yang ditemukan oleh Thomas Bayes pada abad ke-18. *Teorema Bayes* adalah teorema yang digunakan dalam statistika untuk menghitung peluang suatu hipotesis. Bayes merupakan teknik prediksi berbasis *probalistic* sederhana yang berdasar pada penerapan *theorema Bayes* dengan asumsi independensi yang kuat (Eko Prasetyo, 2012). Definisi lain mengatakan *Naive Bayes* merupakan pengklasifikasian dengan metode probalistik dan statistik yang dikemukakan oleh ilmuan Inggris Thomas Bayes, yaitu memperediksi peluang di masa depan dengan pengalaman di masa lalu (Alfa, 2015).

Keunggulan *Naïve Bayes* adalah sifatnya yang efektif dan cepat untuk mengolah data berjumlah besar. Karena kelebihannya itu, *Naïve Bayes* biasa digunakan untuk aplikasi seperti *spam filtering* (pendeteksi pesan sampah) dan deteksi

anomali di jaringan komputer. Algoritma Naïve Bayes bahkan dianggap sebagai standar *de facto* untuk penerapan klasifikasi teks, misalnya sentiment analysis (menentukan apakah penulis suatu komentar bernada positif, negatif atau netral).

Dalam penelitian ini yang menjadi data uji adalah data headline dari berita yang diunggah dan komentar dari setiap *tweets*. Ada dua tahap proses pada klasifikasi dokumen. Tahap pertama adalah pelatihan terhadap dokumen yang sudah diketahui katagorinya, selanjutnya pada tahap dua adalah proses klasifikasi data yang belum diketahui katagorinya.(Dios Kurniawan, 2021).

Naïve Bayes merupakan salah satu contoh metode *supervised document classification* yang membutuhkan data latih dalam melakukan klasifikasi. Selanjutnya yang perlu dilakukan adalah menghitung nilai probabilitas masing-masing katanya dengan persamaan 1.

Rumus untuk Naive bayes classifier ada pada persamaan berikut:

$$P(c_i) = \frac{fd(c_i)}{|D|}$$

Keterangan:

$P(c_i)$ = Probabilitas c_i yang merupakan kategori kelas

$fd(c_i)$ = Jumlah dokumen c_i

$|D|$ = Jumlah data latih / dokumen

Setelah mendapatkan probabilitas dari setiap kelas, selanjutnya yaitu menghitung probabilitas dari setiapfitur pada setiap kelas dengan persamaan 2 :

$$P(w_k|c_i$$

Keterangan:

$P(w_k|c_i)$ = Peluang kemunculan kata ada sebuah kelas, w_k adalah kata yang muncul pada sebuah kategori

$f(w_{ki}, C)$ = Nilai kemunculan kata w_{ki} pada kelas c_i

$P(C_i)$ = Jumlah keseluruhan kemunculan kata pada kelas c_i

$|W|$ = Jumlah data latih/dokumen

Kemudian langkah selanjutnya adalah menentukan probabilitas data uji dari

setiap kelas berdasarkan dari proses pembelajaran. Nilai probabilitas yang paling tinggi akan terpilih:

$$Vmap = \underset{\{kelas\ 0, kelas\ 1\}}{argmax} \prod_{i=1}^{n} P(w_k|c_i)\ x\ \ P(c_i)$$

Keterangan:

$P(w_k|c_i)$ = Probabilitas kemunculan kata-kata pada sebuah kelas, w_k merupakan kata yang muncul pada sebuah kategori

$P(c_i)$ = Menentukan probabilitas c_i yaitu kategori kelas

2.2.9 Python

Python adalah bahasa pemrograman komputer, sama layaknya seperti bahasa pemrograman lain, misalnya C, C++, Pascal. Java, PHP, Perl, dan lain-lain. Sebagai bahasa pemrograman, python tentu memiliki dialek, kosakata atau kata kunci (*keyword*), dan aturan tersendiri yang jelas berbeda dengan bahasa pemrograman lainnya.

Bahasa Pemrograman python disusun pada Desember 1989 oleh Guido Van Rossum di Centrum Wiskunde dan Informatica (CWI), sebuah pusat riset dibidang matematika dan sains, Amsterdam – Belanda; sebagai suksesor atau pengganti dari bahasa pemrograman pendahulunya, bahasa pemrograman ABC, yang juga dikembangkan oleh CWI oleh Leo Geurts, Lambert Meertens, dan Steven Pamberton. (Budi Raharjo, 2019)

Python merupakan Interpreted Language, artinya *Source Code* (kode program) yang ditulis akan dijalankan baris-per-baris oleh python interpreter. Ada empat tipe data dasat yang ada di python: ***string*** (teks), ***integer*** bilangan bulat), ***float*** (bilangan desimal misalnya 0,999), dan ***boolean*** (ya/tidak atau true/false).

Sesuai survey yang dilakukan oleh Kaggle (sebuah komunikasi data science yang dimiliki oleh Google), python adalah bahasa nomor satu paling populer yang digunakan untuk *Machine Learning* dan *Artificial Intelegence.* Popularitas Python ini salah satunya adalah karena tersedianya banyak *library* (modul-modul siap pakai) yang mendukung kegiatan *Data Analytics* dan *Machine Learning*, misalnya Matpotlib, Numpy, Pandas, Scikit-Learn, dan *library* tambahan Python lainnya.

2.2.10 Jupyter Notebook

Jupyter Notebook adalah sebauh aplikasi open source yang fungsinya untuk membuat dan berbagi dokumen yang berisi persamaan matematika, visualisasi data, tautan dan lain sebagainya. Pada Jupyter notebook input dan output berada pada satu line kode sehingga dapat dengan mudah dilakukan visualisai dari setiap kode tersebut. Jupyter Notebook dikenal sebelumnya sebagai IPython Notebook dan dalam waktu dekat akan berevolusi untuk mendukung ilmu data interaktif dan komputasi ilmiah di semua Bahasa pemrograman. (Avila Dkk dalam situsnya http://jupyter.org/).

Jupyter Notebook biasa juga disebut jupyter ini adalah, pengembangan dari Ipython atau Interactive Python. Jupyter Notebook ini suatu editor dalam bentuk web aplikasi yang berjalan di localhost komputer, adapun beberapa hal yang dapat dilakukan oleh Jupyter Notebook seperti menulis kode python, equations, visualisasi dan bisa juga sebagai markdown editor.