# BAB II
# LANDASAN TEORI

## 2.1 Tinjauan Pustaka

Ada beberapa penelitian yang dilakukan tentang analisa sentimen menggunakan algoritma *Naive Bayes* di antaranya adalah penelitian dari Fajar Darwis Dzikril Hakimi yang berjudul "Sistem Analisis Sentimen Publik tentang Opini Pemilihan Kepala Daerah Jawa Timur 2018 pada Dokumen Twitter Menggunakan *Naive Bayes Classifier*". Penelitian tersebut membuktikan bahwa algoritma *Naive Bayes* memiliki performa dengan hasil akurasi sebesar 98,99% presisi sebesar 93,44% *recall* sebesar 97,78% dan *f-measure* sebesar 95,56% (Hakimi, 2018).

Penelitian lain yang membuktikan performa dari Algoritma *Naive Bayes* adalah penelitian dari Fauziah Afshoh yang berjudul "Analisa Sentimen Menggunakan *Naive Bayes* untuk Melihat Persepsi Masyarakat Terhadap Kenaikan Harga Jual Rokok pada Media Sosial Twitter". Penelitian tersebut menghasilkan nilai *precision* terbesar yaitu 76%, nilai terbesar dari *recall* yaitu 93%, nilai terbesar dari *accuracy* pada saat melakukan pengujian yaitu 88%. Algoritma *Naive Bayes* memiliki performa yang baik, namun dalam penelitian dijelaskan salah satu permasalahan yang membuat penelitian bekerja tidak maksimal adalah data yang tidak seimbang mangakibatkan hasil dari klasifikasi sistem tidak memuaskan (Afshoh, 2017).

Masalah ketidakseimbangan data tidak hanya ditemukan pada algoritma *Naïve Bayes*, namun ditemukan pula pada algoritma klasifikasi yang lain. Berdasarkan hasil penelitian dari Siti Rahmi Kurniasari yang berjudul "Implementasi SVM dan Asosiasi untuk *Sentiment Analysis* Data Ulasan The Phoenix Hotel Yogyakarta pada Situs Tripadvisor" *dataset* yang digunakan dalam penelitian memiliki perbandingan kelas yang tidak seimbang, sehingga hasil yang diperoleh kurang maksimal dengan nilai akurasi sebesar 84,77%. Peneliti menyarankan Jika ditemukan kasus ketidakseimbangan data (*imbalanced dataset*), sebaiknya

dilakukan penanganan khusus dengan menggunakan metode atau cara tertentu agar dapat menghasilkan hasil klasifikasi yang optimal (Kurniasari, 2018).

Salah satu solusi yang dapat dilakukan untuk meningkatkan kinerja Algoritma pada data tidak seimbang adalah dengan menerapkan Metode *Ensemble.* Salah satu penelitian terkait Metode *Ensemble* adalah penelitian dari Yoga Pristyanto yang berjudul "Penerapan *Metode Ensemble* untuk Meningkatkan Kinerja Algoritma Klasifikasi pada *Imbalanced Dataset*". Penelitian tersebut menguji bagaimana pengaruh Metode *Ensemble* yakni *Adaptive Boosting* terhadap kinerja Algoritma *Decision Tree (DT), Support Vector Machine,* dan *Naive Bayes*. Berdasarkan hasil pengujian, metode *Ensemble* dengan penambahan *Adaptive Boosting* dapat meningkatkan nilai kinerja algoritma hingga 10,13% (Pristyanto, 2019).

Hasil penelitian tersebut dikuatkan oleh peneliti Lila Dini Utami dan Romi Satria Wahono yang berjudul "Integrasi Metode *Information Gain* untuk Seleksi Fitur dan *Adaboost* untuk Mengurangi Bias pada Analisis Sentimen *Review* Restoran Menggunakan Algoritma *Naive Bayes*". Dalam penelitian disebutkan bahwa *Naive Bayes* membuktikan tingkat akurasi yang bagus saat klasifikasi dianggap seimbang, akan tetapi akurasi menjadi tidak akurat saat menghadapi sentimen klasifikasi yang kompleks (Utami & Wahono, 2015). Berdasarkan hasil penelitian, penggunaan *Naive Bayes* pada data *review* restoran Bahasa Inggris tanpa mengkombinasikan dengan metode lain menghasilkan akurasi sebesar 70% dan AUC=0,500 sama halnya jika *Naive Bayes* dikombinasikan dengan metode *Information Gain*, akurasi yang dicapai hanya 70% dan AUC=0,500, itu membuktikan bahwa *Information Gain* tidak mempengaruhi akurasi terhadap *Naive Bayes.* Akan tetapi hasil yang berbeda didapatkan jika Algoritma *Naive Bayes* dan *Information Gain* dikombinasikan dengan metode *ensemble* yakni Algoritma *Adaboost*, akurasi Algoritma *Naive Bayes* meningkat sebesar 29,5% menjadi 99,5% dan AUC=0,995 (Utami & Wahono, 2015).

## 2.2 Kerangka Pemikiran

Penyusunan penelitian penerapan *Adaptive Boosting* pada Naive Bayes untuk analisa sentimen pengguna Twitter terhadap ketua umum parpol di Indonesia ini disusun melalui beberapa tahapan yaitu:

1. Latar belakang masalah

   Tahapan paling awal, yaitu menelusuri latar belakang penerapan *Adaptive Boosting* pada Naive Bayes untuk analisa sentimen pengguna Twitter terhadap ketua umum parpol di Indonesia.

2. Perumusan masalah

   Menyimpulkan dari latar belakang masalah yang ada menjadi suatu perumusan masalah yang akan diangkat untuk menjadi bahan penelitian. Rumusan masalah yang diangkat adalah Bagaimana penerapan *Adaptive Boosting* pada Naive Bayes untuk analisa sentimen pengguna Twitter terhadap ketua umum parpol di Indonesia?

3. Pengumpulan data tertulis dan tidak tertulis

   Pengumpulan data yang dilakukan yaitu dengan observasi dan studi literatur di perpustakaan.

4. Penguasaan Dasar Aplikasi

   Melakukan percobaan pembuatan alur sistem dengan menggunakan *tools Rapidminer* agar dapat lebih menguasai fitur-fitur dari aplikasi yang akan digunakan sebagai *tools* penelitian.

5. Observasi Aplikasi

   Mencari referensi alur sistem klasifikasi khususnya analisa sentimen yang telah ada dan hasil penelitian baik dari jurnal, buku dan lain sebagainya sebagai referensi untuk membangun sistem.

6. Klasifikasi dan analisa data sentimen

   Melakukan pembuatan sistem untuk klasifikasi serta analisa data sentimen dengan menggunakan data latih yang telah dikumpulkan.

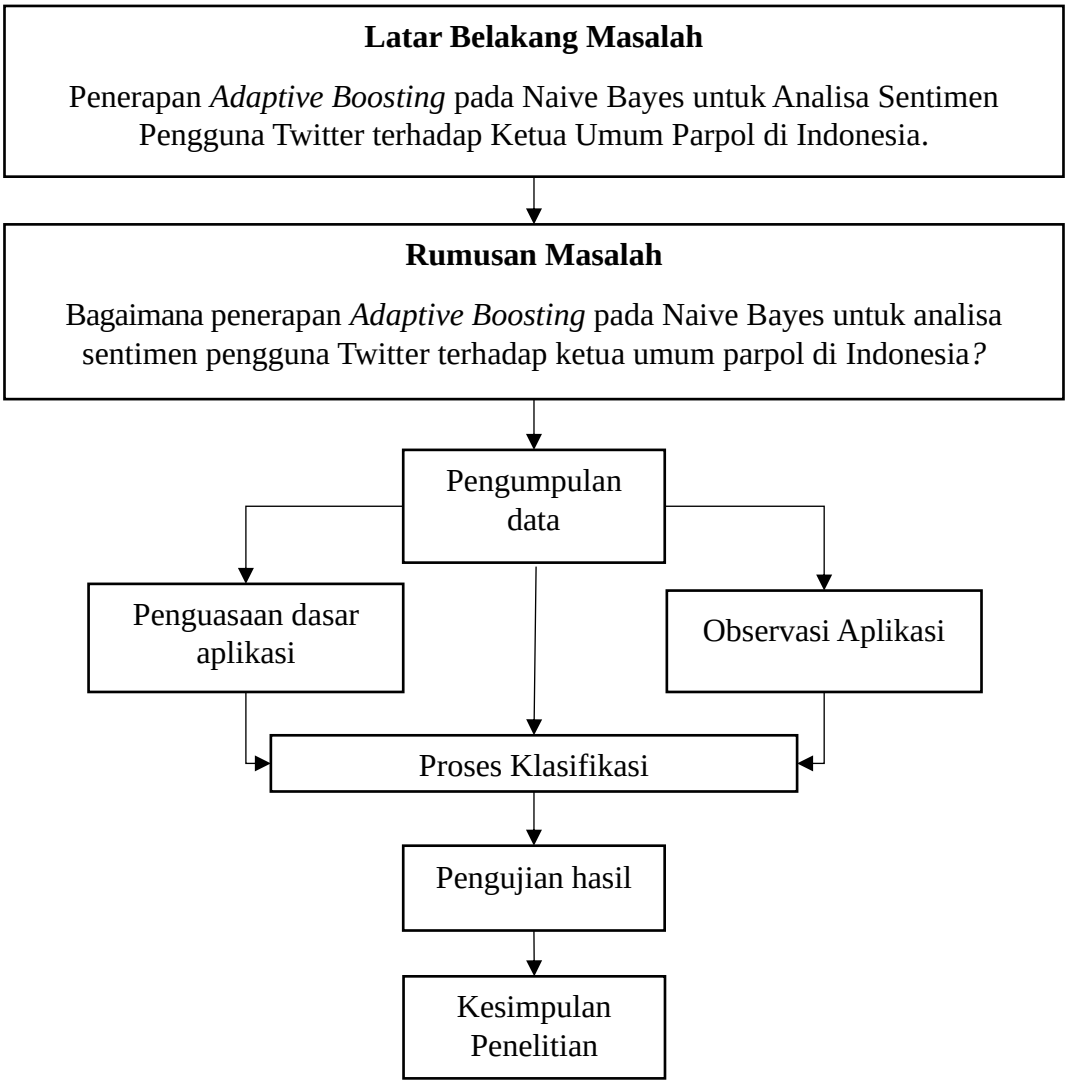
7. Pengujian Sistem

   Melakukan pengujian sistem yang telah dibangun dengan menggunakan data uji untuk mengetahui performa dari sistem yang telah dibangun, kemudian melakukan perbaikan-perbaikan yang diperlukan untuk mendapatkan alur sistem yang memiliki performa maksimal.

8. Evaluasi Hasil

   Melakukan evaluasi hasil dengan memperhatikan performa sistem.

Alur proses penelitian tersebut digambarkan dengan Kerangka pemikiran yang disajikan pada Gambar 2.1.




Gambar 2.1 Kerangka Pemikiran

## 2.3 Dasar Teori

### 2.3.1 Analisa Sentimen

Analisa sentimen/*opinion mining* yaitu sebuah bidang studi yang menganalisis opini seseorang, sentimen, evaluasi, penilaian, sikap dan emosi terhadap suatu entitas seperti produk, jasa, organisasi, masalah, peristiwa dan lain lain (Setyawan, dkk., 2016). Analisa sentimen dapat dimanfaatkan untuk mengetahui pandangan masyarakat terhadap hal tertentu. Dalam analisa sentimen akan

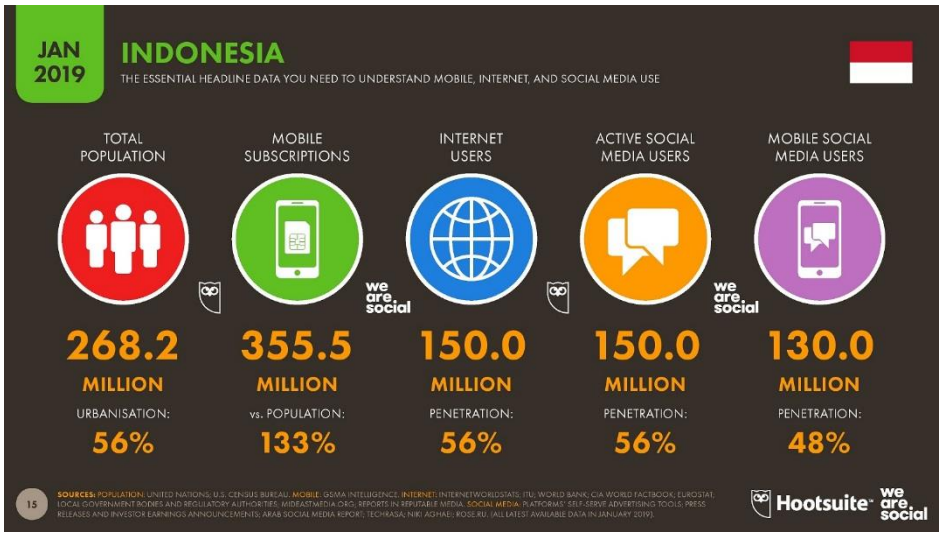
dilakukan pengklasifikasikan orientasi suatu teks ke dalam 2 kategori yakni kategori positif atau negatif (Thelwall, dkk., 2010).

Terdapat beberapa kategori dalam melakukan analisis sentimen, kategori-kategori tersebut yakni: *keyword-spotting, lexical affinity, concept-based,* dan *machine learning* (Cambria, dkk., 2013). *Keyword-spotting* dan *lexical affinity* bersifat *striclyruled*/berbasis aturan-aturan kaku yang sudah didefinisikan sebelumnya, sedangkan *concept-based,* dan *machine learning* dapat mencari pola aturan secara mandiri (Cambria, dkk., 2013).

*Machine learning* atau pembelajaran mesin merupakan pendekatan dalam *Artificial Intelligence* yang banyak digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan masalah atau melakukan otomatisasi (Tanaka & Okutomi, 2014). Inti dari *machine learning* adalah untuk membuat model matematis yang menggambarkan pola-pola data (Jan & Gotama, 2018).

### 2.3.2 Twitter

Masyarakat dapat mengungkapkan sentimen-sentimen terhadap suatu hal melalui berbagai media baik media konvensional maupun media digital. Salah satu media yang sering digunakan adalah media sosial. Dalam laporan "*Digital Around the World* 2019", tertuang bahwa dari total 268,2 juta penduduk di Indonesia, sebanyak 150 juta orang di antaranya adalah pengguna media sosial, sehingga angka penetrasinya sekitar 56 persen yang disajikan pada Gambar 2.2.




Gambar 2.2 Data Pengguna Internet di Indonesia (www.hootsuite.com)

Berbagai media sosial yang digunakan penduduk dunia yaitu Youtube, Facebook, Instagram, Twitter, Pinterest, Snapchat, Path, Tumblr dan Reddit. Dari semua media sosial tersebut, Twitter merupakan media sosial dimana semua orang dapat masuk dan terlibat dalam diskusi tanpa adanya pengaruh dari *gatekeeper* (Achsa, 2018). Media sosial seperti Twitter telah membuat ruang privat menjadi ruang koneksi dan tanpa isolasi seperti halnya individu yang berhubungaan dengan ranah politik, dan melibatkan diri dalam aktifitas pemerintahan dan sosial (Fuchs, 2017).

Berdasarkan informasi dari situs resmi Twitter, Twitter adalah layanan bagi teman, keluarga, dan teman sekerja untuk berkomunikasi dan tetap terhubung melalui pertukaran pesan yang cepat dan sering. Pengguna memposting *tweet* (kicauan), yang dapat berisi foto, video, tautan, dan teks. Dalam Twitter ada beberapa istilah dasar, diantaranya:

1. *Tweet*

   *Tweet* adalah setiap pesan yang diposting ke Twitter dan dapat berisi foto, video, tautan, serta teks.
2. *Retweet*

   *Retweet* adalah *tweet* yang Anda teruskan ke pengikut Anda.
3. *Mention* (@/Sebutan)

   *Mention* (sebutan) adalah *tweet* yang berisi nama pengguna orang lain di mana pun di dalam isi *tweet*.
4. *Hashtag* (#)

   *Hashtag* yang ditulis dengan simbol # digunakan untuk mengindeks kata kunci atau topik di Twitter. Fungsi ini dibuat di Twitter dan memungkinkan pengguna untuk mengikuti topik yang diminati dengan mudah.

### 2.3.3 Parpol (Partai Politik)

Berdasarkan UU Nomor 2 Tahun 2011 tentang Perubahan atas UU Nomor 2 Tahun 2008 tentang Partai Politik, parpol adalah organisasi yang bersifat nasional dan dibentuk oleh sekelompok warga negara Indonesia secara sukarela atas dasar kesamaan kehendak dan cita-cita untuk memperjuangkan dan membela kepentingan politik anggota, masyarakat, bangsa dan negara, serta memelihara keutuhan Negara Kesatuan Republik Indonesia berdasarkan Pancasila dan Undang-Undang Dasar Negara Republik

Indonesia Tahun 1945. Parpol memiliki peran yang sangat penting dalam pembentukan kepemimpinan sebuah negara, karena hanya melalui parpol pasangan calon presiden dan wakil presiden (berdasarkan UUD 1945 Pasal 6A ayat (2)) beserta wakil rakyat di Dewan Perwakilan Rakyat dan Dewan Perwakilan Rakyat Daerah (berdasarkan UUD 1945 Pasal 22E ayat (3)) diusulkan.

Berdasarkan data dari KPU terdapat 27 parpol nasional yang terdaftar sebagai peserta Pemilu Tahun 2019, di mana 9 partai berhasil memenuhi ambang batas parlemen atau *parliamentary threshold* sebagaimana disajikan pada Tabel 2.1.

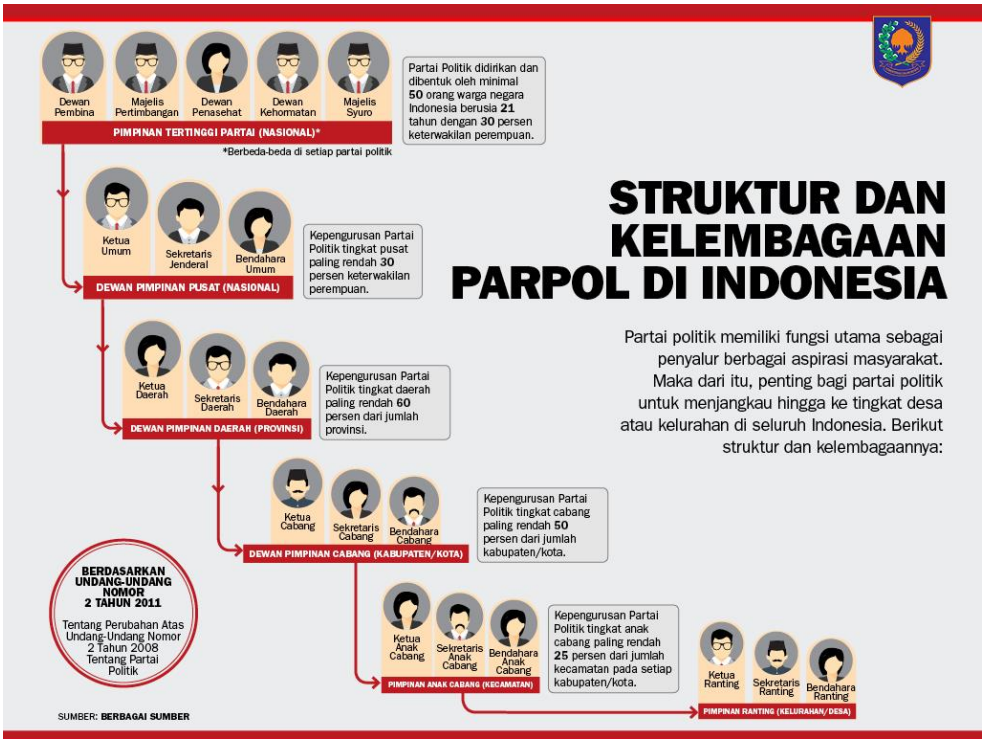
Tabel 2.1 Rekapitulasi Hasil Penghitungan Suara Tingkat Nasional Pemilu 2019

| No | Nama Parpol | Suara Sah | Suara sah (%) |
|---|---|---|---|
| 1 | Partai Demokrasi Indonesia Perjuangan (PDI-P) | 27,053,961 | 19.33 |
| 2 | Partai Gerakan Indonesia Raya (Partai Gerindra) | 17,594,839 | 12.57 |
| 3 | Partai Golongan Karya (Partai Golkar) | 17,229,789 | 12.31 |
| 4 | Partai Kebangkitan Bangsa (PKB) | 13,570,097 | 9.69 |
| 5 | Partai Nasional Demokrat (Partai NasDem) | 12,661,792 | 9.05 |
| 6 | Partai Keadilan Sejahtera (PKS) | 11,493,663 | 8.21 |
| 7 | Partai Demokrat | 10,876,507 | 7.77 |
| 8 | Partai Amanat Nasional (PAN) | 9,572,623 | 6.84 |
| 9 | Partai Persatuan Pembangunan (PPP) | 6,323,147 | 4.52 |

*Sumber: Laporan Ambang Batas dan Perolehan Kursi KPU Tahun 2019 (www.kpu.go.id)*

Berdasarkan pasal 222 Undang Undang Nomor 7 Tahun 2017 tentang Pemilihan Umum bahwa pasangan calon presiden dan wakil presiden diusulkan oleh parpol atau gabungan parpol peserta pemilu yang memenuhi persyaratan perolehan kursi paling sedikit 20 persen dari jumlah kursi DPR atau memperoleh 25 persen dari suara sah nasional dalam pemilu anggota DPR sebelumnya. Hal ini menandakan 9 parpol atau gabungan dari parpol tersebut yang memiliki pengaruh paling kuat dalam pengusulan pasangan calon presiden dan wakil presiden 5 tahun yang akan datang.

Dalam struktur organisasi parpol, ketua umum merupakan pimpinan tertinggi dalam Dewan Pimpinan Pusat sebagaimana dapat dilihat pada Gambar 2.3.




*Sumber: Tim Publikasi Katadata – 2018 (www.katadata.co.id)*

Gambar 2.3 Struktur dan Kelembagaan Parpol di Indonesia

Jabatan pimpinan tertinggi sebuah parpol berbeda-beda tergantung AD/ART yang ditetapkan oleh masing-masing partai. Data mengenai nama ketua umum dan jabatan pimpinan tertinggi 5 parpol dengan perolehan suara terbanyak di Pemilu 2019 dapat dilihat di Tabel 2.2.

Tabel 2.2 Data Ketua Umum dan Pimpinan Tertinggi Partai

| Nama Parpol | Ketua Umum | Pimpinan Tertinggi Partai |
|---|---|---|
| PDI-P | Megawati Soekarnoputri | Ketua Umum |
| Partai Gerindra | Prabowo Subianto | Ketua Dewan Pembina |
| Partai Golkar | Airlangga Hartarto | Ketua Umum |
| PKB | Muhaimin Iskandar | Ketua Umum |
| Partai NasDem | Surya Paloh | Ketua Umum |

Berdasarkan data pada Tabel 2.2, dapat dilihat bahwa Partai Gerindra merupakan parpol di mana pimpinan tertinggi partai tersebut berada di Ketua Dewan Pembina. Namun berdasarkan AD/ART Partai Gerindra Tahun 2014 jabatan tersebut dipegang oleh Prabowo Subianto yang tidak lain adalah Ketua Umum dari Partai Gerindra. Sehingga dapat disimpulkan dari 5 parpol dengan perolehan suara terbanyak pada Pemilu 2019, pimpinan tertinggi partai dipegang

oleh Ketua Umum yang memiliki kewenangan penuh untuk membuat keputusan strategis dalam parpol.

### 2.3.4 Algoritma Naive Bayes

Algoritma *Naive Bayes* merupakan teknik prediksi berbasis *probabilistic* sederhana yang berdasar pada penerapan teorema Bayes (atau aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif) (Prasetyo, 2012). Algoritma *Naive Bayes* mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang paling tepat. Algoritma *Naïve Bayes* memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya karena Algortima *Naive Bayes* merupakan metode pembelajaran terawasi (*supervised learning)* dan sangat tergantung pada data pelatihan (Yuliana & Erlangga, 2017). Algoritma *Navie Bayes* dinyatakan secara matematis sebagai Persamaan 2.1 (Anggarwal, 2015):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (2.1)$$

Dengan $P(B|A)$ merupakan nilai probabilitas dari kemunculan dokumen b pada kelas a, $P(A)$ merupakan nilai probabilitas kemunculan a dan $P(B)$merupakan nilai probabilitas kemunculan dokumen b.

Proses Algoritma *Naïve Bayesian* adalah sebagai berikut (Han, dkk., 2012):

1. *Variable* D menjadi pelatihan set *tuple* dan *label* yang terkait dengan kelas. Seperti biasa, setiap *tuple* diwakili oleh vektor atribut n-dimensi, X = (x1, x2, ..., xn), ini menggambarkan pengukuran n dibuat pada *tuple* dari atribut n, masing-masing, A1, A2, ..., An.
2. Jika ada kelas m, C1, C2, ..., Cm. Diberi sebuah *tuple*, X, *classifier* akan memprediksi X yang masuk kelompok memiliki probabilitas *posterior* tertinggi, kondisi-disebutkan pada X. Artinya, Algoritma *Naive Bayes* memprediksi bahwa X *tuple* milik kelas Ci dengan kondisi sebagaimana disajikan pada Persamaan 2.2.

   $$P(C_i|X) > \left(C_j|X\right) \; dimana \; 1 \leq j \leq m, j \neq i \qquad (2.2)$$

   Jadi memaksimalkan P (Ci | X). Ci kelas yang P (Ci | X) dimaksimalkan disebut hipotesis posteriori maksimal dengan Persamaan 2.3.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (2.3)$$

Keterangan:

P(Ci|X) : Probabilitas hipotesis Ci jika diberikan fakta atau *record* X *(Posterior probability)*

P(X|Ci) : mencari nilai parameter yang memberi kemungkinan yang paling besar *(likelihood)*

P(Ci) : *Prior probability* dari X *(Prior probability)*

P(X) : Jumlah *probability tuple* yg muncul

### 2.3.5 Data Tak Seimbang

Masalah data yang tidak seimbang ditandai dengan lebih banyaknya sampel pada kelas tertentu daripada kelas lainnya. Pada klasifikasi dengan dua kelas, masalah data tak seimbang adalah suatu kondisi dimana salah satu kelas diwakili oleh jumlah data sampel yang lebih besar dibandingkan dengan jumlah data sampel pada kelas lain dengan jumlah yang bervariasi (Sun, 2007). Mayoritas algoritma klasifikasi memiliki kelemahan dalam mengklasifikasi dataset tak seimbang (Wahono, dkk., 2014). Data tak seimbang sangat mempengaruhi performa dari algoritma klasifikasi, karena algoritma klasifikasi bekerja dengan mengasumsikan distribusi kelas pada dataset relatif seimbang (Sun, 2007). Ada tiga pendekatan untuk menangani dataset tak seimbang, yaitu pendekatan pada level data, level algoritmik, dan menggabungkan (*ensemble*) metode (Yap, dkk., 2014).

Pendekatan pada level data mencakup berbagai teknik *resampling* dan sintesis data untuk memperbaiki kecondongan distribusi kelas data latih. Pada pendekatan level algoritmik, metode utamanya adalah menyesuaikan operasi algoritma yang ada untuk membuat pengklasifikasi (classifier) agar lebih konduktif terhadap klasifikasi kelas minoritas (Zhang, dkk., 2011). Pendekatan dengan menggabungkan *ensemble* memiliki tujuan yang sama dengan pendekatan level algoritmik, yaitu memperbaiki algoritma pengklasifikasi tanpa mengubah data (Peng & Yao, 2010).

### 2.3.6 Algoritma Adaptive Boosting (AdaBoost)

*Boosting* merupakan sebuah keluarga *ensemble* yang meliputi banyak algoritma, dimana Algoritma *AdaBoost* merupakan salah satu yang paling populer. Tujuan dari dikembangkannya metode *ensemble* adalah untuk meningkatkan akurasi prediksi klasifikasi (Wezel & Potharst, 2007). Salah satu ide utama Algoritma *AdaBoost* adalah menjaga distribusi atau set bobot (Wang, 2012).

Saat Algoritma *AdaBoost* digunakan pada klasifikasi Algoritma *Naive Bayes*, Algoritma *AdaBoost* akan meningkatkan kerja dari klasifikasi Algoritma *Naive Bayes* dengan cara mengklasifikasikan data yang masuk ke dalam *class* yang masih tidak seimbang dengan semua atribut yang ada didalam dataset (Korada, dkk., 2012).

Algoritma *Adaboost* telah sukses diterapkan pada beberapa bidang (domain) karena dasar teorinya yang kuat, prediksi yang akurat, dan kesederhanaan yang besar (Liu, dkk., 2015). Algoritma *Adaboost* dinyatakan secara matematis sebagai persamaan 2.4 (Liu, dkk. 2015).

$$F(x) = \sum_{t=1}^{T} \propto_t \; h_t\,(x) \qquad (2.4)$$

Keterangan:

$h_t(x)$ = Pengklasifikasi dasar atau lemah

$\propto_t$ = Tingkat pembelajaran (*learning rate*)

F (x) = Hasil, berupa pengklasifikasi kuat atau akhir

Dalam Algoritma *AdaBoost* kesalahan klasifikasi oleh *classifier* sebelumnya direorganisasi menjadi *classifier* selanjutnya. Algoritma dimulai dengan menetapkan bobot yang sama pada semua sampel dalam data pelatihan. Kemudian algoritma pembelajaran dijalankan untuk membentuk *classifier* data tersebut dan dilakukan pembobotan ulang setiap sampel sesuai dengan keluaran *classifier* (Korada, dkk., 2012). Bobot data yang terklasifikasi dengan benar akan diberi bobot lebih kecil, sedangkan data yang terklasifikasi tidak sesuai akan diberi bobot yang lebih besar (Korada, dkk., 2012). Dalam iterasi berikutnya, *classifier* dibangun untuk pembobotan ulang data yang berfokus pada sampel data dengan bobot yang lebih besar, sehingga akurasi klasifikasi dapat diperbaiki (Rani & Saepudin, 2013).