BAB II. LANDASAN TEORI

2.1 State-of-the-Art Penelitian Terdahulu

Penelitian yang dilakukan oleh (Yonathan Sari Mahardika & Eri Zuliarso, 2018) adalah melakukan analisis sentimen terhadap pemerintahan joko widodo pada media sosial twitter menggunakan algoritma naïve bayes classifier. Hasil dari anlisa sentiment ini yaitu Metode Naive Bayes Classifier dalam melakukan klasifikasi tweet sentimen negatif dan positif dengan 300 data latih dan 100 data uji mendapat hasil akurasi sebesar 97% dan dengan hasil akurasi yang cukup tinggi yaitu 97% maka metode Naive Bayes Classifier dapat digunakan untuk melakukan klasifikasi tweet dengan sentimen negatif dan positif secara otomatis..

Penelitian yang dilakukan oleh (Faisal Rahutomo, Imam Fahrur Rozi dan Haris Setiyono, 2018) adalah melakukan implementasi support vector machine pada analisa sentimen twitter berdasarkan waktu. Hasil dari analisa sentiment ini yaitu dalam preprocessing teks pada data twitter berjalan dengan baik dan memerlukan beberapa proses untuk menjadikan tweet lebih optimal untuk selanjutnya diproses pengklasifikasian. Implementasi algoritma support vector machine memiliki persentase akurasi rata-rata sebesar 66% yang dibagi menjadi 3 klasifikasi dengan validasi data menggunakan k-fold cross validation sebanyak 10 bagian data.

Penelitian yang dilakukan oleh (Faisal Rahutomo, Annisa Taufika Firdausi dan Nur rochmanshah, 2019) adalah pengembangan sistem analisa keberpihakan media online berdasarkan trend waktu menggunakan naïve bayes classifier. Hasil dari analisa sentimen ini yaitu bahwa dari hasil pengujian dengan Algoritma Naïve Bayes Classifier dapat digunakan untuk mengklasifikasikan kategori berita. Pada analisa keberpihakan media online berdasarkan trend waktu, dapat diambil kesimpulan bahwa Media A maupun media B berpihak kepada pasangan A dengan Nilai Keberpihakan 72 berita positif untuk pasangan A di Media A, dan 64 Berita Positif untuk pasangan A di Media B. Sedangkan untuk pasangan B, nilai keberpihakan di Media A adalah 50 berita positif dan di Media B adalah -4 berita positif atau 4 berita negatif. Nilai perhitungan accuracy dari pengujian data testing dan data training dengan pembagian merata untuk setiap kategori di setiap sub dataset adalah pada besaran 27%, 50%, 40%, dan 33%.

Penelitian yang dilakukan oleh (Hermawan Arief Putranto, Onny Setyawati dan Wijno, 2016) adalah pengaruh phrase detection dengan POS-Tagger terhadap akurasi klasifikasi sentimen menggunakan SVM. Hasil dari penelitian ini yaitu dengan menggunakan HMM POS-Tagger didapatkan peningkatan nilai akurasi pada proses klasifikasi menggunakan SVM dengan pendekatan berbasis deteksi frasa, lebih kurang sebesar 6% pada Dataset I dan sekitar 3% pada Dataset II. Hal ini dibuktikan dengan banyaknya kalimat dan paragraf yang terklasifikasi dengan benar, sesuai dengan kelas sentimennya, baik pada Dataset I maupun pada Dataset II..

Penelitian yang dilakukan oleh (Imam Fahrur Rozi, 2016) adalah implementasi rule-based document subjectivity pada sistem opinion mining. Hasil dari penelitian ini adalah berupa teks yang masing-masing kata didalamnya sudah memiliki tag. Pada hasil proses POS Tagging kemudian diterapkan rule. Dari pengujian didapatkan nilai precission dan recall untuk proses document subjectivity terbaik adalah 0.99 dan 0.88.

Penelitian yang dilakukan oleh (Nitin Sablok, Bebeto Agung Hardono dan Derry Alamsyah) adalah part-of-speech (pos) tagging bahasa indonesia menggunakan algoritma Viterbi. Hasil dari penelitian ini adalah Algoritma Viterbi dapat digunakan untuk melakukan Part-of-Speech (POS) tagging pada bahasa Indonesia, Probabilitas inisialisasi berpengaruh terhadap label (tag). Nilai probabilitas inisialisasi yang kecil memiliki kemungkinan yang kecil untuk dipilih sebagai label (tag) pada kata. Tingkat akurasi yang dihasilkan pada 10 kali pengujian yang dilakukan menghasilkan akurasi yang tidak jauh berbeda. Dengan rata – rata akurasi adalah 93,23018 % dan standar deviasi sebesar 0,260541273.Penambahan kata ‘zz’ untuk mengelompokkan kata yang tidak terdapat pada korpus (kata asing) tidak berpengaruh terhadap hasil akurasi.

Penelitian yang dilakukan oleh (Evasaria M. Sipayung, Herastia Maharani dan Ivan Zefanya) adalah perancangan sistem analisis sentiment komentar pelanggan menggunakan metode naïve bayes classifier. Hasil dari analisa ini adalah kehandalan sistem analisis sentimen dengan metode NBC ini dari 175 data latih dapat dibagi menjadi dua bagian, perhitungan klasifikasi kategori dan perhitungan klasifikasi sentimen, untuk hasil dari akurasi sistem memiliki nilai accuration

kategori 77.14% dan untuk precision sentimen 99.12%, recall sentimen 72.9%, dan accuration sentimen 75.42%, sehingga sistem mampu mengembalikan dokumen dan kecocokan data yang tinggi.

Penelitian yang dilakukan oleh (Ghulam Asrofi Buntoro, 2017) adalah analisis sentimen calon gubernur dki Jakarta 2017 di twitter. Hasil dari analisa ini adalah Setelah dilakukan analisis sentimen, terlihat berapa banyak sentimen yang ditujukan kepada calon Gubernur DKI Jakarta 2017. Nilai akurasi tertinggi didapat saat menggunakan metode klasifikasi Naïve Bayes Classifier (NBC) untuk klasifikasi data AHY, dengan nilai rata-rata akurasi mencapai 95%, nilai presisi 95%, nilai recall 95% nilai TP rate 96,8% dan nilai TN rate 84,6%. Dalam penelitian ini juga dapat diketahui metode klasifikasi Naïve Bayes Classifier (NBC) lebih tinggi akurasinya untuk klasifikasi sentimen Tweet Bahasa Indonesia dibandingkan dengan metode klasifikasi Support Vector Machine (SVM).

Penelitian yang dialakukan oleh (Hartanto, 2017) adalah text mining an sentiment analisis twitter pada gerakan LGBT. Hasil dari analisa ini adalah didapat mengenai gerakan LGBT cukup konsisten dengan kondisi terkini, hasil dari wordclous simetris dengan histogram analisis sentiment, dimana frekuensi kata seperti “kumpul kebo”, “moral”, “Indonesia”, “perbincangan”, “dibui”, “pindah”, “ditantang”, “garis batas” terbilang tinggi, dan sebanyak 379 tweet beropini netral, 79 menyatakan positif dengan gerakan LGBT dan 27 menyatakan sikap negative.

Penelitian yang dilakukan (Ahmad Fathan Hidayatullah dan Azhari) adalah analisis sentimen dan klasifikasi kategori terhadap tokoh publik pada twitter. Hasil dari analisa ini adalah akurasi pengujian klasifikasi dengan fitur term frequency diperoleh sebesar 79,91% sedangkan fitur TF-IDF didapatkan akurasi sebesar 79,68%. Klasifikasi menggunakan tools RapidMiner dengan Naive Bayes dan fitur term frequency diperoleh sebesar 73,81% sedangkan dengan fitur TF-IDF diperoleh sebesar 71.11%. Klasifikasi dengan Support Vector Machine menghasilkan akurasi 83,14% untuk fitur term frequency dan 82,69% untuk fitur TF-IDF.

Tabel 2. 1 State-of-the-Art Penelitian Terdahulu

No.	Judul	Penulis/Jurnal	Univ/Tahun	Permasalahan	Classifier	Hasil
1.	Analisa Sentimen Terhadap Pemerintahan Joko Widodo Pada Media Sosial Twitter Menggunakan Algoritma Naïve Bayes Classifier	Yonathan Sari Mahardika, Eri Zuliarso	Universitas Stikubank, 2018	Twitter merupakan media sosial yang sedang populer saat ini, disini publik bebas berkomentar dan menulis apapun. Tidak jarang pubik berkomentar dengan kata – kata kasar bahkan ujaran kebencian. Pemerintahan Joko Widodo menuai banyak komentar, ada yang memuji, mengkritik dan menghina.	Naïve Bayes Classifier	Akurasi sebesar 97%.
2.	Implemetasi Support Vector Machine Pada Analisa Sentimen Twitter Berdasarkan Waktu	Faisal Rahutomo, Imam Fahrur Rozi, Haris Setiyono	Politeknik Negeri Malang, 2018	Dalam menentukan kategori positif, negatif atau netral suatu tanggapan masyarakat di twitter dapat dilakukan dengan manual dengan cara membaca setiap tweet. Hal ini tentu membutuhkan banyak waktu dan menghabiskan banyak tenaga. Pada penelitian ini menggunakan algoritma	Support Vector Machine	Nilai rata-rata akurasi 66%, presisi 67% dan recall 66%.

				klasifikasi Support Vector Machine untuk melakukan klasifikasi data tweet menjadi sentimen positif, negatif atau netral		
3.	Pengembangan Sistem Analisa Keberpihakan Media online Berdasarkan Trend Waktu Menggunakan Naïve Bayes Classifier	Faisal Rahutomo, Annisa Taufika Firdausi, Nur Rochmanshah	Politeknik Negeri Malang, 2018	Banyak media online dalam penyajian berita di setiap hari, secara terang-terangan maupun tidak, berpihak kepada salah satu Pasangan Presiden dan Wakil Presiden. Penyajian Media Online yang tidak netral ataupun tidak objektif tidak hanya merugikan bagi pihak Pasangan Presiden dan Wakil Presiden, tetapi juga mampu memberikan perspektif berbeda bagi pembaca atau masyarakat kepada pasangan Presiden dan Wakil Presiden terkait	Naïve Bayes Classifier	Nilai perhitungan accuracy dari pengujian data testing dan data training dengan pembagian merata untuk setiap kategori di setiap sub dataset adalah pada besaran 27%, 50%, 40%, dan 33%.
4.	Pengaruh Phrase Detection dengan POS-Tagger terhadap Akurasi Klasifikasi Sentimen	Hermawan Arief Putranto, Onny Setyawati, Wijono	JNTETI, 2016	Pada proses filtering, kata tunggal hasil proses tokenisasi dibandingkan dengan kata yang ada di dalam	POS TAGGING HMM dan SVM	Dengan menggunakan HMM POS-Tagger didapatkan

	menggunakan SVM			stoplist. Apabila ada kata yang sama, maka kata di dalam array akan terhapus secara otomatis. Namun, proses filtering konvensional ini bisa menyebabkan terhapusnya ciri penting yang dapat mempengaruhi hasil klasifikasi sentimen, khususnya pada dokumen berbahasa Indonesia. Hal ini terjadi karena pada saat proses tokenisasi, kalimat dalam dokumen dipecah menjadi kata tunggal, sehingga apabila ada gabungan kata yang membentuk frasa, komputer tidak akan mengenali frasa tersebut. Hal ini menyebabkan terhapusnya kata dalam kalimat secara otomatis karena kata tersebut juga berada dalam stoplist, walaupun		peningkatan nilai akurasi pada proses klasifikasi menggunakan SVM dengan pendekatan berbasis deteksi frasa, lebih kurang sebesar 6% pada Dataset I dan sekitar 3% pada Dataset II. Hal ini dibuktikan dengan banyaknya kalimat dan paragraf yang terklasifikasi dengan benar, sesuai dengan kelas sentimennya, baik pada Dataset I maupun pada Dataset II

				merupakan bagian dari frasa		
5.	Implementasi Rule-Based Document Subjectivity Pada Sistem Opinion Mining	Imam Fahrur Rozi	Politeknik Negeri Malang, 2013	Permasalahan yang pertama dihadapi dalam mengembangkan sistem opinion mining adalah menentukan apakah suatu teks tergolong kalimat opini atau bukan (document subjectivity). Pada penelitian ini dikembangkan sistem rule-based document subjectivity. Kalimat pertama kali akan diproses menggunakan Hidden Makov Model Part-ofspeech (POS) Tagging	Hidden Makov Model Part-ofspeech (POS) Tagging	Hasil dari proses tersebut berupa teks yang masing-masing kata didalamnya sudah memiliki tag. Pada hasil proses POS Tagging kemudian diterapkan rule. Dari pengujian didapatkan nilai precission dan recall untuk proses document subjectivity terbaik adalah 0.99 dan 0.88.
6.	Part-of-Speech (POS) Tagging Bahasa Indonesia Menggunakan Algoritma Viterbi	Nitin Sabloak, Bebeto Agung Hardono, Derry Alamsyah	STMIK GI MDP Palembang, 2016	Pelabelan Kata dapat dilakukan berbasis aturan (rule based) dan probabilitas (probability-based) dari sebuah model yang dibangun. Beberapa penelitian POS tagging pada	POS-Tagging Viterbi	Tingkat akurasi yang dihasilkan pada 10 kali pengujian yang dilakukan menghasilkan akurasi yang tidak jauh

				bahasa Inggris memiliki nilai akurasi yang tinggi. Bahasa Indonesia memiliki struktur yang lebih kompleks dari bahasa Inggris. Hal ini dilandasi oleh berbagai budaya yang melatarbelakangi bangsa Indonesia. Penelitian POS tagging berbasis rule-based sudah memberikan hasil yang baik untuk bahasa Indonesia, sementara penggunaan berbasis probabilitas mengalami kendala.		berbeda. Dengan rata – rata akurasi adalah 93,23018 % dan standar deviasi sebesar 0,260541273.
7.	Perancangan Sistem Analisa Sentimen Komentar Pelangggan Menggunakan Metode Naive Bayes Classifier	Evasaria M. Sipayung, Herastia Maharani, Ivan Zefanya.	UNSRI, 2016	Hotel XYZ mengalami kesulitan untuk mendapatkan makna atau kesimpulan dari keseluruhan komentar yang diberikan pelanggan terhadap produk dan layanan hotel dikarenakan banyaknya komentar yang ada, pertahun mencapai 675 komentar. Sistem analisis	Naïve Bayes Classifier	Tingkat akurasi dalam penentuan kategori adalah sebesar 77.14% dan 75.42% dalam penentuan sentimen memiliki tingkat precision 99.12% dan recall 72.9%.

				sentiment analysis system bertujuan untuk membantu pihak hotel dalam mendapatkan makna dari komentar yang banyak dengan menggunakan metode Naive Bayes Classifier (NBC). Metode ini mengelompokan komentar berdasarkan kategori-kategori yang ditinjau oleh hotel. Komentar dibagi berdasarkan sentimen positif dan negatif, sehingga dapat dievaluasi kepuasan pelanggan terhadap produk dan jasa yang disediakan secara terkomputerisasi dan spesifik.		
8.	Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter	Ghulam Asrofi Buntoro	Universitas Muhammadiyah Ponorogo 2017	Pemilihan Gubernur DKI Jakarta 2017 ramai diperbincangkan di dunia nyata maupun dunia maya, khususnya di media sosial Twitter. Semua orang bebas berpendapat atau beropini tentang calon Gubernur DKI Jakarta	Naïve Bayes Classifier	Naïve Bayes Classifier (NBC), dengan nilai rata-rata akurasi mencapai 95%, nilai presisi 95%, nilai recall 95% nilai TP rate 96,8%

				2017 sehingga memunculkan banyak opini, tidak hanya opini yang positif atau netral tapi juga yang negatif. Media sosial khususnya Twitter sekarang ini menjadi salah satu tempat promosi atau kampanye yang efektif dan efisien.		dan nilai TN rate 84,6%
9.	Text Mining Dan Sentimen Analisis Twiter Pada Gerakan LGBT	Hartanto	Universitas Widya Dharma, 2017	Gerakan LGBT yang muncul di penghujung tahun 2016 dan masih hangat sampai penelitian ini ditulis. Perilaku dan gerakan LGBT sukar untuk dikategorikan sebagai fenomena yang benar – benar mengalami masa puncak dan mengalami masa turun. Hal tersebut disebabkan oleh tipikal gerakannya yang murni bersifat latent dan asimetris. Sehingga sangat susah ditebak dan dideskripsikan memakai metode yang biasa.	Naïve Bayes Classifier	sebanyak 379 tweet beropini netral, 79 menyatakan positif dengan gerakan LGBT dan 27 menyatakan sikap negative

10.	Analisis Sentimen Dan Klasifikasi Kategori Terhadap Tokoh Publik Pada Twitter	Ahmad Fathan Hidayatullah, Azhari SN	Universitas Islam Indonesia dan Universitas Gadjah Mada, 2014	Hasil akurasi pengujian klasifikasi dengan fitur term frequency diperoleh sebesar 79,91% sedangkan fitur TF-IDF didapatkan akurasi sebesar 79,68%. Klasifikasi menggunakan tools RapidMiner dengan Naive Bayes dan fitur term frequency diperoleh sebesar 73,81% sedangkan dengan fitur TF-IDF diperoleh sebesar 71.11%. Klasifikasi dengan Support Vector Machine menghasilkan akurasi 83,14% untuk fitur term frequency dan 82,69% untuk fitur TF-IDF.	SVM, Naïve Laplace Smooting, TF-IDF	Hasil akurasi pengujian klasifikasi dengan fitur term frequency diperoleh sebesar 79,91% sedangkan fitur TF-IDF didapatkan akurasi sebesar 79,68%. Klasifikasi menggunakan tools RapidMiner dengan Naive Bayes dan fitur term frequency diperoleh sebesar 73,81% sedangkan dengan fitur TF-IDF diperoleh sebesar 71.11%. Klasifikasi dengan Support Vector Machine menghasilkan akurasi 83,14% untuk fitur term frequency

						dan 82,69% untuk fitur TF-IDF.

*terdapat beberapa sumber >5 tahun, tetapi tetap disertakan karena alasan relevansi dengan penelitian.

2.2 Analisa Sentimen

Sentiment analysis merupakan salah satu bidang dari Natural Languange Processing (NLP) yang membangun sistem untuk mengenali dan mengekstraksi opini dalam bentuk teks.

Informasi berbentuk teks saat ini banyak terdapat di internet dalam format forum, blog, media sosial, serta situs berisi review. Dengan bantuan sentiment analysis, informasi yang tadinya tidak terstruktur dapat diubah menjadi data yang lebih terstruktur.

Data tersebut dapat menjelaskan opini masyarakat mengenai produk, merek, layanan, politik, atau topik lainnya. Perusahaan, pemerintah, maupun bidang lainnya kemudian memanfaatkan data-data tersebut untuk membuat analisis marketing, review produk, umpan-balik produk, dan layanan masyarakat.

Sentiment analysis kemudian akan membedakan teks menjadi dua kategori, yakni fakta dan opini. Fakta merupakan ekspresi objetif mengenai sesuatu. Sementara opini adalah ekpresi subjektif yang menggambarkan sentimen, perasaan, maupun penghargaan terhadap suatu hal.

2.3 *PHP (Hypertext Preprocessor)*

PHP pertama kali dibuat oleh Rasmus Lerdorf pada tahun 1995. Pada waktu itu PHP masih bernama Form Interpreted (FI), yang wujudnya berupa sekumpulan skrip yang digunakan untuk mengolah data formulir dari web.

PHP disebut bahasa pemrograman server side karena PHP diproses pada komputer server. Hal ini berbeda dibandingkan dengan bahasa pemrograman client-side seperti JavaScript yang diproses pada web browser (client).

Menurut Madcoms (2013) PHP adalah merupakan singkatan dari “Hypertex Preprocessor”. pada awalnya PHP merupakan kependekan dari personal home page (situs personal) dan PHP itu sendiri pertama kali dibuat oleh Rasmus Lerdof pada tahun 1995, dan pada saat PHP masih bernama FI (form interpreter), yang wujudnya berupa sekumpulan Script yang digunakan untuk mengolah data form dari web. Selanjutnya Rasmus merilis kode tersebut untuk umum. PHP adalah sebuah bahasa Scripting yang terpasang pada HTML.

2.4 *Text Mining*

Text mining adalah proses ekstraksi informasi dari data sumber yang belum terstruktur (Fidyawan, 2017). *Text mining* memiliki tujuan untuk memproses teks agar menjadi informasi yang diperoleh dari peramalan pola dan kecenderungan melalui pola statistik(Luqyana et al., 2018). Input untuk *text mining* adalah data yang tidak (atau kurang) terstruktur, seperti dokumen word, PDF, kutipan teks. *Text mining* digunakan untuk menangani masalah *classification, clustering, information extraction* dan *information retrival.*

2.5 *Text Preprocessing*

Preprocessing text merupakan tindakan menghilangkan karakter-karakter tertentu yang terkandung dalam dokumen, seperti koma, tanda petik dan lain-lain serta mengubah semua huruf kapital menjadi huruf kecil. Selain itu, dalam tahap *text preprocessing* ini dilakukan *tokenization*. *Text mining* dalam prakteknya mencari pola-pola tertentu, mengasosiasikan suatu bagian teks dengan yang lain berdasarkan aturan-aturan tertentu, katakata yang dapat mewakili sehingga dapat dilakukan analisa keterhubungan antar satu dengan yang lain (Siregar et al., 2017). Berikut tahapan-tahapan proses didalam *text mining*:

a. Cleaning

Data cleaning merupakan proses pembersihan kata dengan menghilangkan delimiter tanda baca, hastag, mention, angka, ur. Pembersihan kata bertujuan untuk mengurangi noise.

b. Case Folding

Case Folding bertujuan untuk mengubah setiap bentuk kata menjadi sama. Hal ini dilakukan dengan mengubah kata menjadi lower case atau huruf kecil.

c. Tokenizing

Tahap *Tokenizing* adalah tahap pemotongan tiap kata dalam kalimat atau parsing dengan menggunakan spasi sebagai delimiter yang akan menghasilkan token berupa kata. Pada tokenizing terdapat beberapa proses yang harus dilakukan yaitu merubah semua hruf besar menjadi kecil (*text to lowercase*). Proses selanjutnya adalah penguraian, proses penguraian yang dimaksud adalah membagi *text* menjadi kumpulan kata tanpa memperhatikan keterhubungan antara kata satu dengan kata lain serta peran dan posisinya pada kalimat.

d. Normalisasi Bahasa

Pada tahapan Pre-processing dilakukan normalisasi bahasa terhadap kata tidak baku. Tahapan ini bertujuan untuk mengembalikan bentuk penulisan dari masing-masing kata yang sesuai dengan Kamus Besar Bahasa Indonesia (KBBI). Proses ini dilakukan dengan mencocokkan setiap kata pada dokumen data latih maupun data uji dengan kata yang ada pada kamus Bahasa tidak baku (Darma, 2017).

e. Filtering

Tahap Filtering adalah tahap penyaringan kata yang didapat dari *Tokenizing* yang dianggap tidak penting atau tidak memiliki makna dalam proses *Text mining* yang disebut *stopword*. *Stopword* berisi katakata umum yang sering muncul dalam sebuah dokumen dalam jumlah banyak namun tidak memiliki kaitan dengan tema tertentu. Contoh stopwords adalah “yang”, “di”, “dan”, dll.

f. Stemming

Tahap stemming adalah tahap mengembalikan kata-kata yang diperoleh dari hasil Filtering ke bentuk dasarnya, menghilangkan imbuhan awal (prefix) dan imbuhan akhir (sufix) sehingga didapat kata dasar.

2.6 Pembobotan Laplace Correction

Menurut Dimas (2018 : 9) Dalam proses prediksi untuk menghindari probabailitas 0 (nol) yang dapat menyebabkan Naïve Bayes Classifier tidak dapat mengklasifikasi sebuah data inputan dengan baik maka digunakan teknik Laplace Correction. Yaitu sebuah teknik yang menambahkan nilai 1 pada setiap kombinasi atribut. Untuk jumlah data yang banyak (hingga ribuan) teknik ini akan sangat akurat karena tidak akan membuat perbedaan yang berarti pada estimasi probabilitas. Dengan persamaan sebagai berikut :

$P(X_k|C) = P(Xk|C)+1/P(C)+|V|$

Keterangan :

P(Xk|C) = Probabilitas tiap atribut dari Xk P(C) = Total probabilitas dalam Xk

|V| = Jumlah kemungkinan nilai dari Xk.

2.7 *Algoritma Metode Naïve Bayes*

Algoritma Naive Bayes merupakan sebuah metoda klasifikasi menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris Thomas

Bayes. Algoritma Naive Bayes memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dr Naïve Bayes Classifier ini adalah asumsi yg sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian. Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifier lainnya. Hal ini dibuktikan pada jurnal Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. "Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages." (2009), mengatakan bahwa "Naïve Bayes Classifier memiliki tingkat akurasi yg lebih baik dibanding model classifier lainnya".

$$P(C|d) = P(C) \prod_{i=1}^{n} P(Wi|C)$$

$$P(Wi|C) = \frac{count(wi.c) + 1}{count\ (c) + |V|}$$

Keterangan :

C	: Class
D	: Dokumen
Wi	: Kata ke-i
(Wi|C)	: Jumlah kata wi dalam C
Count (C)	: Jumlah kata di class C
|V|	: Jumlah Vocabulary (Semua Kata)

2.8 *Algoritma Viterbi*

Algoritma Viterbi adalah algoritma dynamic programming untuk menemukan kemungkinan rangkaian status yang tersembunyi (biasa disebut Viterbi path) yang dihasilkan pada rangkaian pengamatan kejadian, terutama dalam lingkup HMM.

Untuk menemukan sebuah rangkaian status terbaik, q = (q q .....q), untuk rangkaian observasi (o1 o2 or), perlu didefinisikan kwantitas:

(1) $\delta t(i) = \max P[q1q2....qt-1, qt = i, o1\ q1,q2...qt-1\ o_2\ o_t\ |\ \lambda]$

δt(i) adalah rangkaian terbaik, yaitu dengan kemungkinan terbesar, pada waktu t dimana perhitungan untuk pengamatan t pertama dan berakhir pada status i. Dengan menginduksi, didapat:

(2) $\delta t+1(j) = [\max \delta t(i)ij] . bj(o1+1)$

Untuk mendapatkan kembali rangkaian status, perlu adanya penyimpanan hasil yang memaksimalkan persamaan (2), untuk tiap i dan j, dengan menggunakan tabel $A_r(j)$. Prosedur lengkap untuk menemukan kumpulan status-status terbaik bisa dirumuskan sebagai:

1. Inisialisasi

   $\delta_1 (i) = \Pi_i b_i(o_1)$, $1 \leq i \geq N\, A_r(1) = 0$

2. Rekrusif

   $\delta t(i) = \max [\delta t-1(i)aij]bj(ot)$

   $1 \leq i \leq N$

   $2 \leq t \leq T$, $1 \leq j \leq N$

   $Ar(j) = \arg \max [\delta t-1(i)aij]$

   $1 \leq i \leq N$

   $2 \leq t \leq T$, $1 \leq j \leq N$

3. Terminasi

   $P^* = \max [\delta T(i)]$

   $1 \leq i \leq N$

   $qT^* = \arg \max [\delta T(i)]$

   $1 \leq i \leq N$

4. Lintasan status

   $q^* = Ar(t+1)(q^*)$

   t t+1

   t = T-1, T-2, ..., 1.