KINERJA NAÏVE BAYES *CLASSIFIER* PADA PENYARINGAN *SHORT MESSAGE SERVICE* (SMS) *SPAM*

(Skripsi)

Oleh

PUTRI APRICIA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

ABSTRAK

KINERJA *NAÏVE BAYES CLASSIFIER* PADA PENYARINGAN *SHORT MESSAGE SERVICE (SMS) SPAM*

Oleh

PUTRI APRICIA

Naïve Bayes juga dikenal multinomial Naïve Bayes ialah metode klasifikasi yang menggunakan metode probabilitas serta statistik. Metode ini ialah model algoritma Bayes yang disederhanakan, sesuai untuk mengklasifikasikan teks atau dokumen. Metode yang digunakan dalam penelitian ini adalah metode Naïve Bayes merupakan teknik analisis klasifikasi pada data SMS. Metode ini digunakan untuk mengklasifikasikan SMS yang ada apakah SMS tersebut *spam* atau bukan *spam (ham)*. Dalam beberapa pembagian data *training* dan data *testing* hasil akurasi yang diperoleh sebesar 97% dengan menggunakan perbandingan 60 : 40. Karena dijaman sekarang SMS merupakan media komunikasi yang paling sering digunakan. Seiring dengan meningkatnya intensitas pengunaan SMS ini dimanfaatkan oleh beberapa orang yang tidak bertanggung jawab untuk melakukan tindak kriminal seperti penipuan melalui media SMS. SMS yang disalahgunakan inilah yang disebut *spam*. Oleh karena itu pemanfaatan informasi ini memerlukan teknik analisis sehingga informasi yang dihasilkan dapat membantu banyak pihak yang ada.

Kata kunci: *Naive Bayes,* SMS spam, Klasifikasi.

ABSTRACT

NAÏVE BAYES CLASSIFIER PERFORMANCE ON SCREENING SHORT MESSAGE SERVICE (SMS) SPAM

By

PUTRI APRICIA

Naïve Bayes also known as multinomial Naïve Bayes is a classification method that uses probability and statistical methods. This method is a simplified Bayes algorithm model, suitable for classifying text or documents.The method used in this research is methodNaive Bayesis a classification analysis technique on SMS data. This method is used to classify existing SMS whether the SMS is spam or not spam (ham). In several divisions of training data and testing data, the accuracy obtained is 97% using a ratio of 60: 40.BecauseToday, SMS is the most frequently used communication medium. Coupled with the increasing intensity of SMS usage, it is used by some irresponsible people to commit criminal acts such as fraud via SMS media. This misused SMS is called spam. Therefore, the utilization of this information requires analytical techniques so that the resulting information can help many existing parties.

Keyword: Naive Bayes, SMS spam, classification

KINERJA *NAÏVE* BAYES *CLASSIFIER* PADA PENYARINGAN *SHORT MESSAGE SERVICE (SMS) SPAM*

Oleh

PUTRI APRICIA

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

Judul Skripsi	: **KINERJA *NAÏVE BAYES CLASSIFIER* PADA PENYARINGAN *SHORT MESSAGE SERVICE (SMS) SPAM***
Nama Mahasiswa	: **Putri Apricia**
NPM	: **1957031015**
Jurusan	: **Matematika**
Fakultas	: **Matematika dan Ilmu Pengetahuan Alam**

MENYETUJUI

1. Komisi Pembimbing

Dr. Khoirin Nisa, M.Si.
NIP. 19740726 200003 2 001

Dr. Muslim Ansori, M.Si.
NIP. 19720227 199802 1 001

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 19740316 200501 1 001

MENGESAHKAN

1. Tim Penguji

   Ketua : **Dr. Khoirin Nisa, M.Si.** .....

   Sekretaris : **Dr. Muslim Ansori, M.Si.** .....

   Penguji
   Bukan Pembimbing : **Prof. Ir. Netti Herawati, M.Sc., Ph.D.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung

Dr. Eng. Heri Satria, S.Si. M.Si
NIP. 19711001 200501 1 002

Tanggal Lulus Ujian Skripsi : **13 Juni 2023**

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama	: **Putri Apricia**
Nomor Pokok Mahasiswa	: **1957031015**
Jurusan	: **Matematika**
Judul Skripsi	: **KINERJA NAÏVE BAYES *CLASSIFIER* PADA PENYARINGAN *SHORT MESSAGE SERVICE (SMS) SPAM***

Dengan ini menyatakan bahwa skripsi ini adalah hasil pekerjaan saya sendiri dan semua tulisan yang tertuang dalam skripsi ini telah mengikuti kaidah karya penulisan ilmiah Universitas Lampung.

Bandar Lampung, 13 Juni 2023

Yang menyatakan,

Putri Apricia

NPM. 1957031015

RIWAYAT HIDUP

Penulis bernama Putri Apricia biasa disapa nci atau Aprici lahir di Bandar Lampung pada tanggal 10 April 2001, merupakan anak kedua dari pasangan Bapak Muhammad Amin dan Ibu Rosdiana Roni. Penulis mempunyai saudara laki-laki yang bernama Erlanda Primajaya Saputra dan Tegar Prianandа dan Saudara perempuan bernama Arsyila Romeesa Farzana.

Penulis mengawali pendidikan Taman Kanak-kanak di TK Xaverius Terbanggi Besar pada tahun 2006-2007, kemudian menempuh pendidikan sekolah dasar di SD IT Bustanul Ulum Terbanggi Besar pada tahun 2007-2013, selanjutnya pada tahun 2013-2016 penulis melanjutkan pendidikan Sekolah Menengah Pertama di SMPN 3 Terbanggi Besar dan tahun 2016-2019 penulis melanjutkan Sekolah Menengah Atas di SMAS Global Madani Bandar Lampung. Pada tahun 2019 penulis terdaftar sebagai mahasiswa S1 Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) Universitas Lampung.

Selama menjadi mahasiswa penulis aktif dalam organisasi Tapak Suci (TS). Selain aktif dalam organisasi kampus, pada tahun 2021 penulis mengikuti kegiatan Kampus Merdeka yaitu program Kampus Mengajar pada bulan Juli-Desember 2021 dan ditempatkan di SDN 4 Way Laga, Panjang, Lampung. Penulis melaksanakan Kerja Praktik (KP) di BPS Provinsi Lampung pada tanggal 4 Januari 2022 sampai 12 Februari 2022. Dan melaksanakan Kuliah Kerja Nyata (KKN) Periode II tahun 2022 di Desa Sumber Hadi, Kabupaten Lampung Timur, Lampung pada bulan Juni-Agustus 2022

KATA INSPIRASI

" Libatkan Allah dalam setiap urusanmu"

"Dan barang siapa yang bertakwa kepada Allah, niscaya Allah menjadikan baginya kemudahan dalam urusannya"
(Q.S At-Talaq: 4)

"Terkadang kita diuji bukan untuk menunjukkan kelemahan kita, tetapi untuk menemukan kekuatan kita."

"Ilmu tanpa akal ibarat seperti memiliki sepatu tanpa kaki. Dan akal tanpa ilmu ilbarat seperti memiliki kaki tanpa sepatu."
(Ali bin Abi Thalib)

"Maka sesungguhnya bersama kesulitan itu ada kemudahan."
(QS Al-Insyirah: 5)

"Apapun yang kamu lakukan, lakukanlah dengan sekuat tenaga."
(Marcus Tullius Cicero)

PERSEMBAHAN

Alhamdulillah hirobbil'alamin,
Puji dan syukur tiada hentinya terpanjatkan kepada Allah SWT atas ridhonya sehingga penulis dapat menyelesaikan skripsi ini. Saya persembahkan karya ini untuk :

Orang Tua Tercinta

Kepada kedua orang tuaku yang selalu memberikan doa, tenaga pikiran dan dukungan secara terus menerus untuk keberhasilanku dalam segala hal serta menjadi penyemangat terbaikku. Terimakasih atas semuanya, orang tuaku adalah orang yang sangat luar biasa.

Dosen Pembimbing dan Pembahas

Kepada dosen-dosen Pembimbing dan Pembahas yang telah sangat sabar dalam membimbing, memberikan masukan dan ide-ide yang membangun sehingga dapat menyelesaikan skripsi ini.

Sahabat - Sahabatku

Kepada sahabat-sahabatku yang selalu memberikan keceriaan, semangat dan doa serta kenangan canda tawa selama masa perkuliahan

Almamater Tercinta, Universitas Lampung.

SANCAWACANA

Puji syukur kehadirat Allah SWT, atas segala rahmat dan karunianya sehingga penulis dapat menyelesaikan tugas akhir berbentuk Skripsi yang berjudul **"Kinerja Naïve Bayes *Classifier Pada Penyaringan Short Message Service* (SMS) Spam".** Sholawat serta salam tak hentinya selalu tercurahkan kepada baginda besar kita Nabi Muhammad SAW yang kita natikan syafaatnya di yaumul akhir kelak.

Dalam penyusunan skripsi ini, penulis menyadari bahwa skripsi ini tidak akan terealisasi dengan baik tanpa adanya dukungan, bantuan, bimbingan serta arahan dari berbagai pihak. Pada kesempatan ini penulis ingin mengungkapkan rasa terima kasih sebesar-besarnya kepada:

1. Ibu Dr. Khoirin Nisa, M.Si., selaku Dosen Pembimbing I yang telah banyak memberikan bimbingan, arahan, dan kemudahan kepada penulis untuk menyelesaikan skripsi ini.
2. Bapak Dr. Muslim Ansori, M.Si. selaku Dosen Pembimbing II yang telah memberikan bimbingan dan arahan selama proses penulisan skripsi ini.
3. Ibu Prof. Ir. Netti Herawati, M.Sc., Ph.D. selaku Dosen Pembahas, terimakasih atas kesediaannya untuk membahas dan memberikan saran yang membangun untuk keberhasilan skripsi saya.
4. Alm Bapak Amanto, S.Si., M.Si. selaku dosen pembimbing akademik saya yang telah membimbing saya. Semoga bapak bahagia selalu disana.
5. Ibu Dina Eka Nurvazly, S.Pd., M.Si. selaku dosen pembimbing akademik saya yang telah membimbing saya.

6. Bapak Dr. Aang Nuryaman, S.Si., M.Si., selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
7. Bapak Dr. Eng. Heri Satria, S.Si., M.Si., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
8. Terkhusus untuk kedua orang tuaku, yang tak henti-hentinya mencurahkan doa, ridho, kasih sayang dan cinta untuk saya hingga sampai saat ini. Semoga Allah selalu menjaga, melindungi, meridhoi, dan mengumpulkan kita sekeluarga di syurganya kelak. Aaminn.
9. Saudara kandungku tersayang dan Keluarga besar yang selalu mendukung dalam menjalani Pendidikan. Semoga kalian semua dimudahkan salam segala urusan, usaha, dan rejekinya, dan dapat bermanfaat serta mebanggakan kedua orang tua serta keluarga.
10. Yadi Aprizal patner terbaikku yang selalu menemani, membantu dan memberikan dukungan. Temanku Hijer, Feby, Debi, Nada, Qorry, Zahro yang selalu mendukung dan memberikan semangat dalam menjalani pendidikan di Jurusan Matematika FMIPA Universitas Lampung.
11. Semua pihak yang tidak bisa saya sebutkan satu-persatu yang telah membantu saya dalam menyelesaikan Skripsi ini.

Akhir kata, semoga Allah SWT senatiasa melimpahkan Rahmat-Nya dan membalas segala kebaikan kepada pihak-pihak yang telah membantu penulis. Penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan. Kritik yang konstruktif sangat penulis harapkan agar dapat digunakan untuk bahan perbaikan kedepannya. Semoga laporan ini memberikan manfaat kepada kita.

Bandar Lampung, 13 Juni 2023
Penulis,

Putri Apricia

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Naïve Bayes juga dikenal multinomial Naïve Bayes ialah metode klasifikasi yang menggunakan metode probabilitas serta statistik. Metode ini ialah model algoritma Bayes yang disederhanakan, sesuai untuk mengklasifikasikan teks atau dokumen. Saat menggunakan Naïve Bayes untuk klasifikasi, nilai jenis dari sesuatu dokumen hendak ditentukan bersumber pada fitur/kata yang timbul dalam dokumen yang hendak diklasifikasikan. Permasalahan tersebut dapat diatasi dengan terdapatnya aplikasi *short message service* (SMS) *filtering* (Bakri, 2017). Namun walaupun bermacam fitur lunak *short message service filtering* banyak ada, permasalahan *spam* pula semakin tumbuh, sehingga melahirkan sebagian metode *short message service filtering* (Rajput, dkk., 2019). Salah satu tata cara *short message service filtering* yang sangat terkenal ialah tata cara Bayes (Bayesian *filtering*) (*Ham*zah, 2012).

Metode tersebut menggunakan teorema probabilitas ialah teorema Bayes, untuk memprediksi probabilitas masa depan bersumber pada data di masa lebih dahulu. Oleh sebab itu penulis berupaya mengulas kembali pelaksanaan tata cara Bayes pada proses *short message service filtering* serta setelah itu menelaah tingkatan akurasi dari Bayesian *filtering*. Berdasarkan penelitian yang dilakukan KOMINFO jumlah pengguna *instan messaging* tahun 2017 diperkirakan sebanyak 3,3 miliar akun. Dengan rincian 84,76% adalah pengguna *instan messaging* dan sebanyak 15,24% adalah bukan pengguna *instan messaging* dan mayoritas

PNS/TNI/Polri, karyawan swasta, dan Non PNS/honorer memakai pulsa lebih dari 1 juta rupiah/bulan baik di wilayah rural dan urban.

Pemakaian internet sudah jadi kebutuhan sehari-hari yang berarti dalam kehidupan berkomunikasi. Aplikasi internet sebagai sarana komunikasi, terdapat beberapa sarana di internet, salah satunya adalah pesan elektronik ataupun yang lebih diketahui selaku *short message service* (SMS). *Short message service* adalah salah satu penggunaan internet yang paling populer untuk komunikasi karena cepat, bianyanya murah dan mudah digunakan. *Short message servive* ialah media komunikasi di internet semacam buat berdiskusi (maillist), mentransfer data berbentuk file (*mail attachment*) terlebih dapat digunakan buat media periklanan sesuatu industri. Sarana *short message service* murah serta mudah dikirim ke sejumlah penerima hingga sebagian orang menggunakannya untuk mengirim *short message service* yang berisi promosi produk atau layanan, pornografi, virus, serta konten *non-esensial* lainnya ke ribuan pengguna *short message service.*

Penggunaan *short message service* yang sangat sering memiliki dampak positif dan negatif karena pada dasarnya tidak semua orang memakai *short message service* dengan benar, bahkan banyak terjadi penyalahgunaan *short message service* yang sangat berpengaruh merugikan orang lain. *Short message service* yang disalah gunakan inilah yang umumnya dikenal dengan *spam*. Untuk menanggulangi permasalahan ini dibutuhkan sesuatu *filter* (Sharma, 2018), contohnya merupakan klasifikasi, yang bisa memisahkan *spam SMS* serta *ham SMS* (Kurniawan, dkk., 2012). Ada beberapa pengklasifikasian yang bisa diaplikasikan dalam klasifikasi *spam* semacam *Decision Tree, K- Nearest Neighbor* (KNN), Naïve Bayes, ID3 serta C4. Dari kelima metode-metode tersebut, Naïve Bayes ialah tata cara statistik yang simpel serta mempunyai akurasi yang baik dan *error rate* yang *minimum* pada proses pengklasifikasian (Saad, dkk., 2012).

Pada tahun 2016 Syarli dan Asrul Azari Muin melakukan penelitian menggunakan metode Naïve Bayes untuk memprediksi kelulusan mahasiswa. Pada tahun 2013 Selvia Lorena Br Ginting dan Reggy Pasya Trinanda melakukan penelitian yang berjudul Penggunaan Metode Naïve Bayes *Classifier* pada Aplikasi Perpustakaan. Dalam penelitian ini Naïve Bayes *Classifier* dipergunakan untuk mengklasifikasikan beberapa judul dan kategori yang terdapat pada database perpustakaan kemudian pada tahun 2012 Amir *Ham*zah melakukan penelitian dengan judul Klasifikasi Teks dengan Naïve Bayes *Classifier* (NBC) untuk Pengelompokan teks berita dan abstrak akademis. Dalam penelitian tersebut Naïve Bayes *Classifier* dipergunakan untuk mengklasifikasi dokumen berita maupun dokumen akademis.

Pentingnya dilakukan penelitian ini dikarenakan banyaknya penggunan *short message service* pada zaman sekarang , maka penulis melakukan penerapan metode Naïve Bayes pada proses penyaringan *short message service* dan menganalisis tingkat akurasi dari metode Naïve Bayes tersebut. Penelitian ini dilakukan untuk mengetahui seberapa besar tingkat akurasi metode Naïve Bayes terhadap penyaringan *spam* dan *ham* pada *short message service*.

1.2 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menerapkan metode Naïve Bayes pada *short message service spam* dan *ham.*
2. Mengetahui seberapa besar tingkat akurasi metode Naïve Bayes terhadap penyaringan *spam* dan *ham* pada *short message service*.
3. Mendapatkan hasil penyaringan *spam* dan *ham* pada *short message service*.

1.3 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Menambah wawasan terkait penerapan metode *Naïve* Bayes.
2. Memberikan informasi terkait penyaringan *spam* dan *ham* yang terdapat pada *short message service*.
3. Mendapatkan hasil akurasi, *precision, recall* dan *F1-Score* penyaringan *spam* dan *ham* yang terdapat pada *short message service* .

II. TINJAUAN PUSTAKA

2.1 Konsep Peluang

Peluang ialah suatu konsep matematika yang dipakai untuk melihat besarnya probabilitas atau peluang terjadinya suatu kejadian. Peluang (probabilitas) ialah harga angka yang menunjukkan sebesar apa peluang suatu kejadian akan terjadi. Konsep peluang telah banyak diaplikasikan pada hal-hal yang bersifat sederhana contohnya permainan dadu atau pada hal yang lebih kompleks, seperti investasi, cuaca, asuransi, dan lainnya. Nilai peluang antara 0 dan 1, yang artinya peluang kejadian 0 tidak terjadi dan peluang kejadian 1 pasti terjadi. Konsep dasar peluang adalah penjabaran lebih rinci terkait besaran apa saja yang wajib dipa*ham*i.

Beberapa istilah yang harus diketahui untuk mempelajari konsep peluang adalah sebagai berikut:

1. Ruang sampel ialah himpunan semua hasil yang mungkin dari sebuah percobaan.
2. Titik sampel ialah anggota yang ada pada ruang sampel.
3. Kejadian ialah himpunan bagian dari ruang sampel.

Konsep ini didapatkan dengan melakukan percobaan. Jika suatu percobaan mempunyai N hasil percobaan yang berbeda, dan masing-masingnya memiliki kemungkinan yang sama untuk terjadi, serta jika tepat n di antara hasil percobaan itu menyusun A, maka peluang kejadian A adalah:

$$P(A) = \frac{n(A)}{n(S)} \tag{2.1}$$

dengan:

$n(A)$: Banyaknya cara atau kemungkinan terjadi.

$n(S)$: Banyaknya semua kemungkinan.

2.1.1 Peluang Bersyarat

Menurut Otaya (2016), peluang bersyarat masuk ke materi ilmu peluang matematika. Pada peluang bersyarat terdapat konsep apabila ada 2 kejadian bisa disebut sebagai kejadian bersyarat atau kejadian yang saling bergantung. Apabila kejadian A berkaitan dengan kejadian B atau sebaliknya, itu bisa disebut sebagai kejadian bersyarat atau saling bergantung. Ini dilambangkan dengan P(A|B) dibaca peluang terjadinya A bila kejadian B diketahui. Didefiniskan sebagai berikut:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0 \tag{2.2}$$

Menurut Lumbantoruan (2019), peluang bersyarat dilambangkan sebagai berikut: Peluang terjadinya kejadian A dengan syarat apabila kejadian B telah terjadi terlebih dahulu, ditulis $P(A|B)$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) \neq 0 \tag{2.3}$$

Peluang terjadinya kejadian B dengan syarat apabila kejadian A telah terjadi terlebih dahulu, ditulis $P(B|A)$:

$$P(B|A) = \frac{P(B \cap A)}{P(A)}, P(A) \neq 0 \tag{2.4}$$

dengan $P(A \cap B)$ = peluang irisan A dan B.

2.1.2 Peluang Kaidah Bayes

Kaidah Bayes adalah kaidah yang memperbaiki suatu probabilitas dengan menggunakan informasi tambahan, yaitu dari probabilitas awal yang belum diperbaiki kemudian dirumuskan menurut informasi yang ada atau tersedia saat ini, maka dibuatlah probabilitas berikutnya (Natalius & Samuel, 2010). Teori lain mengatakan bahwa Teori Bayes merupakan kesimpulan statistik yang menarik kesimpilan yang baru atau suatu probabilitas yang benar-benar mungkin. Algoritma Bayes dirumuskan sebagai berikut:

$$P(Y|X_1, \dots, X_n) = \frac{P(Y).P(X_1, \dots, X_n|Y)}{P(X_1, \dots, X_n)} \quad (2.5)$$

dengan:

$P(Y|X_1,...,X_n)$: Peluang masuknya sampel variabel tertentu dalam kelas Y (*posterior*).

$P(Y)$: Peluang munsulnya kelas Y sebelum masuknya sampel (*prior*).

$P(X_1,...,X_n|Y)$: Peluang kemunculan variabel sampel pada kelas Y (*likelihood*).

$P(X_1,...,X_n)$: Peluang kemunculan variabel sampel secara umum (*evidence*).

Bisa juga ditulis dengan:

$$Posterior = \frac{Prior\ x\ Likelihood}{Evidence} \quad (2.6)$$

Pada nilai *posterior* akan dibandingkan dengan nilai *posterior* kelas lain yang nantinya akan digunakan untuk menentukan kelas mana sampel harus diletakkan. Pada nilai *evidence* bernilai 1 atau tetap pada tiap kelas suatu sampel, maka dapat

diabaikan. Nilai *posterior* dapat dihitung dengan mengalikan nilai *prior* dan *likelihood.* Nilai *prior* ialah peluang kelas Y muncul sebelum sampel masuk, dirumuskan sebagai berikut (Watratan, dkk., 2020):

$$P(Y_n) = \frac{d_n}{D} \tag{2.7}$$

dengan:

$P(Y_n)$: Peluang munculnya kelas Y ke-n, n=1,2,3...t

d_n : Banyaknya pengamatan pada kelas ke-n.

2.2 Naïve Bayes *Classifier*

Naïve Bayes adalah salah satu metode yang bisa digunakan untuk mengklasifikasikan data, juga metode klasifikasi yang algoritmanya menggunakan probabilitas dan statistik. Ini berakar pada teorema Bayes dan memiliki asumsi independensi poin yang tertinggi untuk setiap kondisi atau peristiwa. Bayes*ian classification* ialah pengklasifikasian statistik yang dapat diaplikasikan untuk memprediski probabilitas keanggotaan suatu kelas (Borman & Wati, 2020). Teorema ini diusulkan oleh ilmuwan Inggris, Thomas Bayes, dengan memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Gandhi, dkk., 2020).

Terorema Bayes dipadukan dengan "*Naïve*" yang berarti setiap atribut/variabel bersifat bebas (*independent*) (Putri & Surahman, 2020). Keuntungan dari pengklasifikasi ialah bahwa hanya membutuhkan sejumlah kecil data pelatihan untuk memperkirakan parameter (sarana dan *varians* dari variabel) yang dibutuhkan untuk klasifikasi. Teori Naïve Bayes memiliki kemampuan klasifikasi yang mirp dengan *Decision Tree* dan *Neural Network* bahkan algoritma Naïve Bayes memiliki akurasi dan kecepatan yang tinggi saat

dioperasikan ke dalam database dengan data yang besar (Fahlapi & Rianto, 2020). Metode klasifikasi Naïve Bayes dimanfaatkan untuk mengambil keputusan dengan melakukan prediksi pada suatu kasus berdasarkan hasil dari klasifikasi yang telah didapatkan.

Algoritma Naïve Bayes adalah algoritma yang sangat populer untuk digunakan. Naïve Bayes terdiri dari beberapa jenis, termasuk Naïve Bayes Gaussian, Naïve Bayes Multivariate Bernoulli, Naïve Bayes Multinomial. Dari berbagai jenis Naïve Bayes yang diolah menggunakan bentuk data yang berbeda diperoleh pada tahap pra-pemrosesan (Pedregosa, dkk., 2011). Klasifikasi Naïve Bayes mengasumsikan bahwa suatu fitur tidak bergantung pada fitur lain, yang disebut asumsi Naïve.

Metode Naïve Bayes memiliki 2 tahap pada proses klasifikasi teks, yaitu tahap pelatihan dan tahap pengujian. Pada tahap pelatihan terdapat proses analisis terhadap sampel dokumen, yaitu pemilihan *vocabulary*. Kata yang mungkin ada atau muncul dalam koleksi dokumen sampel yang mungkin dapat menjadi representasi dokumen. Kemudian penentuan probabilitas *prior* bagi tiap kategori yang didasarkan pada sampel dokumen. Dalam tahap klasifikasi ditentukan berdasarkan nilai kategori yang berasal dari suatu dokumen seusai dengan erm yang ada pada dokumen yang diklasifikasikan.

Pada penelitian ini, metode Naïve Bayes digunakan untuk menentukan apakah SMS tersebut *spam* atau *ham* pada notasi klasifikasi Naïve Bayes ditunjukkan pada persamaan berikut ini:

$$P(A|B) = \frac{P(B|A)x\ P(A)}{P(B)} \tag{2.8}$$

dengan:

A : Hipotesis data merupakan *class* spesifik.

B : Data dengan kelas yang masih belum diketahui.

P(A|B) : Probabilitas hipotesis berdasarkan kondisi.
P(A) : Probabilitas hipotesis.
P(B|A) : Probabilitas berdasarkan kondisi pada hipotesis.
P(B) : Probabilitas B.

Setiap pesan bisa kita pertimbangkan untuk mewakili *short message service* dengan variabel {x1,x2,...,xn} dari tiap kata yang berbeda yang terdapat dalam teks. $P(Y|x1, \dots, xn)$, dimana *Y* adalah *short message service spam* atau *short message service ham*, dengan Teorema Bayes berikut:

$$\begin{aligned} P(Y|X_1, \dots, X_n) &= P(Y).P(X_1, .., X_n|Y) \\ &= P(Y).P(X_1|Y).P(X_2, .., X_n|Y, X_1) \\ &= P(Y).P(X_1|Y).P(X_2|Y, X_1).P(X_3, .., X_n|Y, X_1, X_2) \\ &= P(Y).P(X_1|Y).P(X_2|Y, X_1) \dots P(X_n|Y, X_1, X_2 \dots X_{n-1}) \end{aligned} \quad (2.9)$$

Semakin banyak variabel dari data yang ada, semakin kompleks faktor kondisional yang mempengaruhi nilai probabilitas. Di bawah asumsi itu, persamaan berikut ini dapat kita gunakan dalam pengkalsifikasian.

$$P(Y|X_1, \dots, X_n) = P(Y).P(X_1|Y).P(X_2|Y) \dots P(X_n|Y). \quad (2.10)$$

Dengan ini kita dapat mempertimbangkan probabilitas *short message service* menjadi *spam*, jika kita mengetahui probabilitas sebelumnya dari *short message service* menjadi *spam*, dan probabilitas untuk kata tertentu dalam *short message service spam* atau *ham*. Sejauh asumsi independensi berlaku dan perkiraan probabilitas akurat, pengklasifikasi mengadopsi kriteria ini mencapai hasil yang optimal (Duda & Hart, 1973).

Short message service memiliki cara untuk menghitung probabilitas *short message service* menjadi *spam* menggunakan istilah yang relatif sederhana, tetapi belum jelas bagaimana menghitung probabiltas bersyarat tertentu.
Pengklasifikasian ini berfungsi dengan mengambil sejumlah *short message*

service yang telah dsiberi label sebagai *spam*, dan menggunakan data tersebut untuk menghitung probabilitas *spam* kata dengan menghitung frekuensi setiap kata.

2.3 Data *Training* dan Data *Testing*

Data yang digunakanan akan dibagi menjadi dua yaitu data latih dan data uji, ini dilakukan karena akan digunakan untuk menentukan tingkat akurasi dari suatu metode. Data *training* merupakan data yang sebelumnya sudah diketahui untuk label kelompok dan dipakai untuk membangun model klasifikasi. Data *testing* merupakan data yang belum diketahui label kelompoknya. Data *training* digunakan untuk mencari model yang tepat dan untuk melatih algoritma. Data *testing* dipakai untuk mengukur keakuratan *classifier* apakah berhasil melakukan klasifikasi dengan benar (Witten, dkk., 2011).

Data *training* dan data *testing* biasanya dibagi dengan perbandingan data 60:40 yang berarti data *training* 60% dan data *testing* 40%, 70:30 yang berarti data *training* 70% dan data *testing* 30%, 80:20 yang berarti data *training* 80% dan data *testing* 20%, dan 90:10 yang berarti data *training* 90% dan data *testing* 10%. Untuk menghitung jumlah data *training* dan data *testing* digunakan persamaan berikut:

$$\text{Jumlah data } training = \text{proporsi data } training \; x \; N \tag{2.11}$$

$$\text{Jumlah data } testing = N - \text{jumlah data } training \tag{2.12}$$

dengan:

N : Jumlah seluruh data.

2.4 *Confusion Matrix*

Metode *confusion matrix* digunakan pada tahap pengujian model yang ditampilkan hasil dari nilai model dengan table matrix. *Confusion matrix* menggunakan tabel matriks seperti pada Tabel 1 dimana *record* perbandingan hasil klasifikasi data *testing* atau data uji berdasarkan data *training* atau data latih dengan data sebenarnya (Liu, 2011). Jika dataset terdapat 2 kelas maka kelas yang pertama disebut positif dan kelas lainnya disebut negatif (Putra & Wibowo, 2020).

Tabel 1. *Confusion Matrix*

Kelas Asli	Kelas Prediksi	
	Positif	Negatif
Positif	*True Positive* (TP)	*Flase Negative* (FN)
Negatif	*False Positive* (FP)	*True Negative* (TN)

True positive merupakan jumlah dari *record* positif yang diklasifikasikan sebagai positif, *false positive* merupakan jumlah dari *record* negatif yang diklasifikasikan sebagai positif, *false negative* merupakan jumlah dari *record* positif yang diklasifikasikan sebagai negatif, *true negative* merupakan jumlah dari *record* negatif yang diklasifikasikan sebagai negatif (Putra & Wibowo, 2020). Klasifikasi *accuracy* merupakan pengukuran utama, jumlah kasus diklasifikasikan dengan benar dalam rangkaian pengujian dibagi dengan jumlah total kasus dalam rangkaian pengujian. *Precision* dan *recall* mengukur seberapa presisi dan kelengkapan klasifikasi ini pada kelas positif:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \; x \; 100\% \tag{2.13}$$

Precision (p) merupakan perbandingan kejadian secara aktual diklasifikasikan sebagai positif untuk semua ketentuan yang diklasifikasikan sebagai positif.

$$p = \frac{TP}{TP + FP}\ x\ 100\% \tag{2.14}$$

Recall (r) merupakan perbandingan kejadian secara aktual diklasifikasikan sebagai positif untuk semua ketentuan yang diklasifikasikan sebagai positif.

$$r = \frac{TP}{TP + FN}\ x\ 100\% \tag{2.15}$$

F1-score merupakan harmonic mean dari *precision* dan *recall*. Nilai terbaik dari *F1-score* 1.0 dan nilai terburuknya 0, jika F1-score memiliki skor yang baik dapat mengidentifikasi bahwa model klasifikasi memiliki precision dan recall yang baik.

$$F1 = \frac{2(Recall\ x\ Precision)}{Recall + Precision}\ x\ 100\% \tag{2.16}$$

Beberapa studi telah menemukan pengklasifikasian Naïve Bayes menjadi sangat efektif (Langley, dkk., 1992). Meskipun fakta bahwa asumsi independensinya biasanya terlalu sederhana (Domingos & Pazzani, 1996).

2.5 Tipe Naïve Bayes *Classifier*

2.5.1 Multinomial Naïve Bayes

Model ini telah banyak dipergunakan untuk menyelesaikan masalah klasifikasi dokumen, seperti *short message service*, peringkat film, dokumen, situs web, dsb. Dalam pembelajaran teks memiliki hitungan setiap kata ini digunakan untuk memprediksi kelas atau label. Algorima multinomial Naïve Bayes akan membantu mengetahui kategori suatu dokumen, apakah dokumen tersebut

termasuk dokumen penting, dokumen yang bisa dipindahkan atau malah dokumen *spam* berbahaya. Algoritma ini dapat dipergunakan untuk mengkategorikan dokumen menurut tema tertentu seperti, gaya hidup, sosial politik, atau olahraga.

Fitur yang digunakan oleh *classifier* merupakan frekuensi kata yang ada dalam dokumen tersebut. Misalnya, bila suatu dokumen terus menerus menampilkan kata "kesehatan tubuh", "pola makan", "tidur cukup" maka ini bisa dimasukkan kedalam kategori gaya hidup sehat.

2.5.2 Bernoulli Naïve Bayes

Bernoulli Naïve Bayes *ham*pir mirip dengan Multinomial Naïve Bayes, perbedaan antara keduanya terdapat pada fitur atau prediktornya. Alih-alih menggunakna frekuensi kata, algoritma Bernoulli Naïve Bayes akan menggunakan variabel boolean. Dengan kata lain, parameter yang akan dipakai untuk memprediksi kelas hanya mengambil dua nilai. Misalnya 0 dapat mewakili kata tidak muncul didokumen dan 1 sebagai kata muncul didokumen atau dengan menggunakan nilai ya atau tidak, benar atau salah.

Seperti, untuk dapat menentukan apakah dokumen tersebut masuk kedalam kategori gaya hidup, makan bisa dengan mengenali apakah kata "pola makan" muncul dalam dokumen tersebut. Jika kata tersebut muncul, maka dokumen secara otomatis diklasifikasikan sebagai dokumen kategori pola hidup, dan jika kata tersebut tidak muncul maka dokumen tersebut tidak masuk kedalam dokumen tentang gaya hidup.

2.5.3 Gaussian Naïve Bayes

Karakteristik dari algoritma Gaussian Naïve Bayes, yaitu ketika fitur atau prediktor mengambil nilai yang kontinu (tidak diskrit). Asumsikan bahwa setiap fitur mengikuti distribusi Gaussian. Saat data diplot, kurva lonceng simetris ditampilkan, kurva ini akan menunjukkan rata-rata dari setiap fitur.

2.6 Klasifikasi Teks Menggunakan Naïve Bayes

Klasifikasi teks ialah proses yang mengklasifikasika teks ke dalam kategori tententu (Damayanti & Sulistiani, 2017). Pada text mining, klasifikasi mengacu pada kegiatan menganalisis atau meninjau sekumpulan dokumen teks yang telah dikategorikan sebelumnya untuk mendapatkan model atau fungsi yang bisa dipergunakan untuk mengklasifikasikan dokumen teks lain yang tidak diketahui kelasnya ke dalam satu atau lebih kategori tersebut. Dokumen yang dipergunakan untuk belajar dinamakan contoh yang dideskripsikan oleh himpunan atribut atau variabel.

Klasifikasi termasuk pembelajaran jenis *supervised learning*, jenis lain merupakan *unsupervised learning* atau disebut juga *clustering*. Di *supervised learning*, data pelatihan terdapat pasangan data input umumnya vektor serta hasil yang diperlukan, sedangkan di *unsupervised learning* belum ditentukan sasaran hasil atau *output* yang wajib diperoleh. Proses klasifikasi teks bisa dibagi dua fase, yaitu fase *information retrieval (IR)* yakni menerima data numerik dari dokumen teks serta fase klasifikasi utama yakni dimana algoritma memproses data numerik di atas untuk menetapkan ke kategori mana teks baru akan ditempatkan (Patil & Pawar, 2012).

Pada proses klasifikasi memakai multinomial Naïve Bayes, akan dilakukan *pre-processing* SMS terlebih dahulu. Tahap *pre-processing* SMS dilakukan

untuk menghapus fitur yang tidak diperlukan serta membersihkan semua dokumen yang diberi *tag* untuk meningkatkan akurasi pengklasifikasian pada saat digunakan. Berikut adalah langkah-langkah dalam *pre-processing*:

1. Tokenisasi

Mengekstrak isi dokumen SMS ke dalam bentuk kata atau fitur lebih dikenal dengan nama token ini merupakan langkah awal dalam proses tokenisasi. Kemudian dilakukan *case folding*, yaitu penyeragamn bentuk huruf dimana semua huruf diseragamkan ke bentuk huruf kecil semua serta menghilangkan karakter-karakter tertentu seperti angka atau tanda baca lainnya.

2. Eliminasi *Stopword*

Stopwords merupakan kata yang mempunyai frekuensi kemunculan yang tinggi namun tidak mempunyai nilai informasi yang tinggi. Contohnya adalah kata *"a", "in", "is","the", "that", "this"*. Pada tahun 2003 Silvat dan Ribeirot melakukan penelitian, mereka menyimpulkan bahwa penghapusan stopwords mempunyai dampak yang penting untuk menaikkan akurasi *classifier*.

3. Lemmatisasi Fitur

Pada proses lemmatisasi dilakukan reduksi kata ke dalam bentuk kata dasar. Lemma atau kata dasar memiliki arti dan terdapat dalam kamus. Seperti kata *"talks", "talked", "talking"* bila dilakukan proses lemmatisasi maka kata dasarnya ialah *"talk"*.

2.7 *Short MessageSservice Filtering*

Short message service filtering merupakan suatu proses yang otomatis akan mendeteksi sebuah *short message service* yang ada, apakah SMS tersebut *spam* atau *ham*. Beberapa metode yang bisa dipergunakan untuk *short message service filtering* diantaranya *Keyword Filtering, Black Listing* serta *White Listing*,

Signature-Based filtering, Naïve Bayesian *filtering*. Sebagian ciri *short message service filtering* yaitu:

1. Binary *Class*\
   Binary *class* mengklasifikasikan *short message service* kedalam kelas *spam* dan legitimate *short message service*
2. Prediksi
   Short message service filtering mampu memprediksi kelas dari sebuah *short message service* yang ada.
3. Komputasi Mudah
   Sifat data *short message service* memiliki dimensi yang tinggi, oleh karena itu membutuhkan *short message service filter* yang dapat melakukan komputasi dengan mudah.
4. *Learning*
   Dapat melakukan *learning* dari *short message service* yang sudah ada.
5. Kinerja yang Bagus
   Mengurangi nilai *false positive*, mentolerir nilai *false negative* yang relatif tinggi dan memiliki akurasi yang tinggi.

2.8 *Spam*

Spam atau *junk* SMS merupakan peyalahgunaan pada pengiriman informasi elektronika untuk menunujukkan informasi iklan serta keperluan lainnya yang menyebabkan ketidaknyamanan bagi para pengguna web (Kurniawan, dkk., 2012). *Spam* ialah *unsolicited short message service* (SMS yang tidak diinginkan) yang dikirimkan ke banyak orang. Menurut Lambert (2003) *spam* didefinisikan sebagai berikut :

1. Isi atau konten dari *short message service* tidak sesuai dengan minat penerima.
2. Penerima tidak dapat menolak masuknya *short message service* yang tidak diinginkan tersebut.

3. Dari sisi penerima, pengiriman serta penerimaan pesan tersebut memberikan laba bagi pengirimnya.

Bentuk informasi *spam* yang biasanya dikenal mencakup: *spam* pos-el, *spam* pesan *instan, spam usenet news-group, spam* mesin pencari informasi web (*web search.engine spam*), *spam blog, spam* informasi di telepon genggam, *spam* forum internet, serta lain-lainnya. *Spam* ini umumnya bertubi-bertubi tanpa diminta serta tak jarang tak diketahui oleh penerimanya. *Spam* terjadi akibat murahnya biaya buat mengirimkan *spam,* biaya pengiriman satu SMS sama dengana 1000 *short message service* atau bahkan satu juta *short message service*. *Spam* dapat dikategorikan sebagai berikut:

1. *Junk* mail merupakan *short message service* yang dikirimkan secara besar-besaran yang berasal dari dalam perusahaan bisnis yang sebenarnya tidak kita harapkan.
2. *Non-commercial spam,* contohnya surat berantai atau cerita humor yang dikirimkan secara banyak tanpa tujuan komersial tertentu.
3. *Pornographic spam* yaitu *short message service* yang dikirimkan secara massal untuk mengirimkan gambar pornografi.
4. Virus *spam* merupakan *short message service* yang dikirimkan secara masal, serta mengandung virus atau Trojans

Alih-alih menghapus pesan yang diblokir, pesan tersebut dapat dikembalikan ke pengirim, dengan permintaan untuk mengirimkannya kembali ke alamat *short message service* pribadi penerima yang tidak di*filter* (Hall, 1998).

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada semester genap tahun ajaran 2023/2024 di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini adalah data online yaitu data *short message service spam* dan *ham* sebanyak 5574 data yang diambil dari web kaggle yang dapat diakses (https://www.kaggle.com/datasets/ucim/sms-*spam*-collection-dataset.)
Data ini digunakan untuk mengetahui seberapa banyak *short message service spam* dan *ham* yang ada.

3.3 Metode Penelitian

Langkah-langkah yang dilakukan dalam penelitian ini menggunakan *Python* adalah sebagai berikut:

1. *Import dataset* SMS *spam* dan *ham* yang di dapat dari situs kaggle ke dalam *Pyhton* menggunakan *Jupyter Notebook.*
2. Melakukan visualisai data SMS *spam* dan *ham*.
3. Melakukan data cleaning yaitu cek *missing value* dan cek jumlah nilai *unique* pada label dan baris terduplikat dan mentransformasikan data.
4. Melakukan *prepocesing* data, yaitu
   - *Cleaning*: membersihkan karakter atau simbol.
   - *Case folding:* mengubah data*frame* menjadi huruf kecil semua agar sama rata dan memudahkan dalam pengolahan data selanjutnya.
   - *Splitting:* pemotongan setiap kata pada data*frame*.
   - *Stemming:* proses mengubah kata menjadi kata dasar dengan menghilangkan awalan dan akhiran. Karena data*frame* yang digunakan menggunakan bahasa *inggris,* sehingga *stopwords* yang digunakan pada pengujian ini ialah *stopwords* bahasa *inggris.*
   - *Waordcloud:* mengetahui kata yang sering muncul.
   - Melakukan pembagian data *training* dan data *testing* dengan perbandingan data 60:40, 70:30, 80:20, dan 90:10.
5. Membangun model menggunakan Naïve Bayes, untuk mengetahui kelas terbaru dari data yang ada apakah masuk kekelas *spam* atau *ham*.
6. Melakukan klasifikasi menggunakan data *training* dan *testing*.
7. Melakukan evaluasi menggunkan *confusion matrix* yang didapatkan dari hasil klasifikasi dan menghitung nilai *accuracy, precision, recall, f1 score* dengan metode multinomial Naïve Bayes.

V. Kesimpulan

5.1 Kesimpulan

Berdasarkan hasil analisis data dan pembahasan yang telah dilakukan dengan menggunakan Multinomial Naïve Bayes dalam mengklasifikasikan SMS *spam* dan *ham*, maka diperoleh kesimpulan hasil dari nilai *precision, recall, f1-score* dan akurasi menggunakan *confusion matrix* dengan data SMS *spam* dan *ham* pada perbandingan data 60:40 menghasilkan nilai *precision, recall, f1-score* dan akurasi yang lebih tinggi dibandingkan dengan perbandingan lainnya. Dengan ini menjadikan perbandingan 60:40 mendapatkan model terbaik.

5.2 Saran

Berikut ini saran yang dibuat untuk pengembangan penelitian selanjutnya:

1. Bagi peneliti yang ingin melanjutkan penelitian ini, dapat memasukan data baru dan membuat apakah data yang masuk tersebut *spam* atau *ham*.
2. Penelitian ini dapat dikembangkan dengan menggunakan metode lain, menggabungkan metode lain atau membandingkan dengan metode lainnya, sehingga dapat dilihat metode mana yang pengklasifikasiannya lebih baik.

DAFTAR PUSTAKA

Bakri, M. 2017. Penerapan Data Mining untuk Clustering Kualitas Batu Bara dalam Proses Pembakaran di PLTU Sebalang Menggunakan Metode K-Means. *Jurnal Teknoinfo*. **11**(1): 6.

Borman, R. I. & Wati, M. 2020. Penerapan Data Maining Dalam Klasifikasi Data Anggota Kopdit Sejahtera Bandarlampung Dengan Algoritma Naïve Bayes. *Jurnal Ilmiah Fakultas Ilmu Komputer*. **9**(1): 25-34.

Damayanti, D. & Sulistiani, H. 2017. Sistem Informasi Pembayaran Biaya Sekolah Pada SD Ar-Raudah Bandar Lampung. *Jurnal. Teknoinfo*. **11**(2): 25.

Darwanto, D & Dinata, K. B. 2021. *Pengantar Teori Peluang*. UMKO Publishing, Lampung Utara.

Domingos, P. & Pazzani, M. 1996. Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. Proceedings of the 13th Int. Conference on Machine Learning, Bari, Italy.

Duda, R. O., & Hart, P. E. 1973. Bayes *Decision Theory. Chapter 2 in Pattern Classification and Scene Analysis*. New York, Wiley. **3**(1): 731-739.

Fahlapi, R., & Rianto, Y. 2020. Twitter Comment Prediksi Perubahan Iuran BPJS Kesehatan Tahun 2020. *Jurnal Penelitian Teknik Informatika*. **5**(1): 170-183.

Gandhi, B.S., Megawaty, D.A., & Alita, D. 2020. Aplikasi Monitoring Dan Penentuan Peringkat Kelas Menggunakan Naïve Bayes *Classifier*. *Jurnal Informatika Dan Rekayasa Perangkat Lunak (Jatika)*. **2**(1): 54-63.

Hall, R. J. 1998. How to Avoid Unwanted Short message service (SMS). *Communications of ACM*. **41**(3):88–95.

*Ham*zah, A. 2012. Klasifikasi Teks Dengan Naïve Bayes *Classifier (NBC)* Untuk Pengelompokan Teks Berita Dan Abstract Akademis. Prosiding Seminar Nasional Aplikasi Sains dan Teknologi. **3**(1): 269– 277.

Kurniawan, B., Effendi, E., & Sitompul, O.S. 2012. Klasifikasi Konten Berita Dengan Metode Text Mining. *Jurnal Dunia Teknol. Inf.* **1**(1): 14–19.

Lambert, A. 2003. *Analysis of SPAM, Master's thesis, Department of Computer Science*. University of Dublin, Trinity College.

Langley, P., Wayne, I., & Thompson, K. 1992. *Analysis of* Bayes*ia. Classifiers*. Proceedings of the 10th National Conference on AI, San Jose, California.

Liu, B. 2011. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Second Edition*. Spiger, New York.

Lumbantoruan, J. H. 2019. Buku Materi Pembelajaran Teori Peluang dan Kombinatorika. Universitas Kristen Indonesia, Jakarta.

Natalius & Samuel. 2010. Metode Naïve Bayes Classifer dan Penggunaannya Pada Klasifikasi Dokumen (Skripsi). Jurusan Sistem dan Teknologi Informasi STEI ITB, Bandung.

Otaya, L. G. 2016. Probabilitas Bersyarat, Independensi Dan Teorema Bayes Dalam Menentukan Peluang Terjadinya Suatu Peristiwa. *Tadbir: Jurnal Manajemen Pendidikan Islam*. **4**(1), 68–78.

Patil, A. S & Pawar, B. V. 2012. Klasifikasi Otomatis Situs Web Menggunakan Algoritma Naive Bayesian. Prosiding Multikonferensi Internasional Insiyur dan ilmuwan Komputer. **1**(1): 519-523

Pedregosa, F., Gael,V.,Vincent, & Bertrand, M. T. 2011. Scikit-lear:Machine learning in Pyhton. *Journal Machine Learning*. **12**(1): 2825-2830.

Putra, D., & Wibobo, A. 2020. Prediksi Keputusan Minat Penjurusan Siswa SMA Yadika 5 Menggunakan Algoritma Naive Bayes. Prosiding Seminar Nasional Riset dan Information Science (SENARIS). **2**(1): 84-92.

Rajput, A. S, Athaval e, V., & Mittal, S. 2019. Intelligent model for classification of *SPAM* and *HAM*. *International. Journal Innovation. Technology*. **8**(6): 773–777.

Saad, O. A., Darwish, & Faraj, R.. 2012. A survey of machine learning techniques for *Spam filtering*. *Journal Computer. Science*. **12**(2): 66–73.

Sharma, M. 2018. A Survey of Short message serviceSpam Filtering Methods. *Journal International Institute for Science, Technology and Education*. **7**(1):14–21.

Watratan, A. F., & Moeis, D. 2020. Implementasi Algoritma Naïve Byaes Untuk Memprediksi Tingkat Penyebaran Covid. *Jurnal of Applied Computer Science and Technology.* **1**(1): 7-14.

Witten, I. H., & Frank, E., & Jasmir. 2020. Klasifikasi Data Mining Untuk Mendiagnosa Penyakit ISPA Menggunakan Metode Naïve Bayes Pada Puskesmas Jambi Selatan. *Jurnal Manajemen Teknologi Dan Sistem Informasi (JMS).* **2**(3): 214-227.