**PROGRAM STUDI SARJANA SISTEM INFORMASI PROPOSAL**

**PROYEK PEMPROGRAMAN BAHASA ALAMI (11S4037)**



***Sentiment Analysis for Amazon Food Review Using Naïve Bayes***

**Disusun Oleh :**

12S17012 – Reza Oktovian Siregar

12S17032 – Angelia Regina Ginting

12S17047 – Christina Clara

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**JANUARI 2021**

**DAFTAR ISI**

[BAB I. PENDAHULUAN 6](#_Toc61033389)

[1.2 Rumusan Masalah 8](#_Toc61033390)

[1.3 Tujuan 8](#_Toc61033391)

[1.4 Manfaat 8](#_Toc61033392)

[1.5 Ruang Lingkup 8](#_Toc61033393)

[BAB II. ISI 9](#_Toc61033394)

[2.1 Analisis Data 9](#_Toc61033395)

[2.1.1 Analisa Data 9](#_Toc61033396)

[2.1.2 Analisa Metode 11](#_Toc61033397)

[2.2 Desain 13](#_Toc61033398)

[BAB III. IMPLEMENTASI 18](#_Toc61033399)

[3.1 Implementasi *Collecting Data* 18](#_Toc61033400)

[3.2 Implementasi *Load data* dan *Review Data* 18](#_Toc61033401)

[3.3 Implementasi Explore Data 20](#_Toc61033402)

[3.4 Implementasi *Text Pre-Processing* 24](#_Toc61033403)

[3.4.1 Implementasi *Convert Into Lower Cases* 24](#_Toc61033404)

[3.4.2 Implementasi Remove Stopwords 24](#_Toc61033405)

[3.4.3 Implementasi Remove HTML Tags 25](#_Toc61033406)

[3.4.4 Implementasi *Remove Special Characters* 25](#_Toc61033407)

[3.4.5 Implementasi *Lemmatization* 26](#_Toc61033408)

[3.4.6 Implementasi *Stemming* 26](#_Toc61033409)

[3.5 Implementasi Feature Extraction 27](#_Toc61033410)

[3.6 Implementasi *Modelling* 28](#_Toc61033411)

[3.7 Implementasi Evaluation 30](#_Toc61033412)

[BAB IV. PENUTUP 31](#_Toc61033413)

[4.1 Kesimpulan 31](#_Toc61033414)

[4.2 Saran 31](#_Toc61033415)

[4.3 Pembagian Tugas 31](#_Toc61033416)

[Referensi 32](#_Toc61033417)

**DAFTAR TABEL**

[Tabel 1. Variabel pada Metode Evaluasi 17](#_Toc61034650)

[Tabel 2. Pembagian Tugas 31](#_Toc61034651)

**DAFTAR GAMBAR**

[Gambar 1. Kode prgram read data 10](#_Toc61034488)

[Gambar 2. Rumus Naïve Bayes 12](#_Toc61034489)

[Gambar 3. Tahapan Pemrosesan Bahasa Alami 13](file:///E:\Kuliah\Semester%207\PBA\Proyek\Sentiment-Analysis-for-Amazon-Food-Review---Impostor\PBA-Gsl-2021-Impostor.docx#_Toc61034490)

[Gambar 4. Kode program load data 18](#_Toc61034491)

[Gambar 5. Kode program menampilkan informasi data 19](#_Toc61034492)

[Gambar 6. Kode program drop kolom 19](#_Toc61034493)

[Gambar 7. Kode Program melakukanscale nilai pada produk 20](#_Toc61034494)

[Gambar 8. Kode program split data 20](#_Toc61034495)

[Gambar 9. Kode program wordcloud 21](#_Toc61034496)

[Gambar 10. Wordcloud review buruk 22](#_Toc61034497)

[Gambar 11. Wordcloud review baik 22](#_Toc61034498)

[Gambar 12. Kode program analisis produk dengan pengguna 23](#_Toc61034499)

[Gambar 13. Visualisasi review pengguna 2000-2012 23](#_Toc61034500)

[Gambar 14. Kode program lower case 24](#_Toc61034501)

[Gambar 15. Kode program remove stopwords 24](#_Toc61034502)

[Gambar 16. Kode program Remove HTML Tags 25](#_Toc61034503)

[Gambar 17. Kode program remove special characters 25](#_Toc61034504)

[Gambar 18. Kode program lematisasi 26](#_Toc61034505)

[Gambar 19. Kode program stemming 27](#_Toc61034506)

[Gambar 20. Kode program feautre extraction 28](#_Toc61034507)

[Gambar 21. Kode program implementasi klasifikasi 28](#_Toc61034508)

[Gambar 22. Kode program kalkulasi akurasi 29](#_Toc61034509)

[Gambar 23. Implementasi heatmap 29](#_Toc61034510)

[Gambar 24. Kode program evaluasi 30](#_Toc61034511)

# **BAB I. PENDAHULUAN**

Pada bab ini akan dijelaskan mengenai latar belakang, tujuan, manfaat dan ruang lingkup proyek.

**1.1 Latar Belakang**

Diera digital di waktu sekarang membuat masyarakat memasuki gaya kehidupan yang baru dimana hampir keseluruhan kaktivitas menggunakan media elektronik. Salah satu gaya kehidupan yang cukup banyak mengalami perubahan yakni dalam kegiatan jual-beli dimana belanja secara *online* sudah menjadi kegiatan yang sudah umum dimana sebelumnya kegiatan jual-beli dengan masih konvensional dan kini sudah menjadi modern dikarenakan kehadiran internet.

Didalam *e-commerce*, *review* dari sesama pengguna menjadi sebuah acuan untuk mencari saran maupun kualitas terkait layanan dan produk sebelum melakukan pembelian. Amazon merupakan website *e-commerce* yang menyediakan berbagai kebutuhan pelanggan seperti produk makanan, produk elektronik, produk kesehatan, produk kecantikan, dan berbagai kebutuhan lainnya. Namun pada penelitian ini akan berfokus pada produk makanan. Didalam website Amazon menyediakan sebuah kolom komentar atau disebut *review* yang dapat dimanfaatkan oleh pelanggan untuk membagikan pengalaman terkait dengan produk yang dibeli. Melaui kolom komentar tersebut dapat menjadi masukan untuk pihak Amazon maupun pelanggan yang ingin melakukan pembelian terhadap produk (makanan) tersebut. Namun dikarenakan informasi yang diberikan cukup banyak maka pelanggan hanya memperhatikan *review* teratas. Didalam kolom *review* tersebut juga banyak kata-kata yang dapat menjadi masukan untuk pihak amazon maupun penjual barang yang terkait dengan kualitas barang.

Pengguna yang memanfaatkan *e-commerce* untuk transksi jual-beli semakin bertumbuh setiap waktu ditambah juga produk yang dijual semakin beragam sehingga pada akhirnya akan mengasilkan *review* yang berjumlah banyak pada setiap produk. Dikarenakan jumlah data yang sangat besar dan sangat banyak sehingga tidak mungkin untuk melakukan analisis dan menghasilkan kesimpulan dengan cara yang manual [1]. Selain dikarenakan jumlah data yang besar, masalah yang lain yakni informasi pada *review* adalah berupa opini. Opini berisikan informasi yang bersifat tidak lengkap, tidak konsisten, atau informasi yang berperilaku pada tren tertentu [2]. Banyak komentar juga menggunakan kaidah bahasa yang tidak formal. Melalui proses ekstraksi informasi dan mengembangkan sistem analisis sentimen, maka sistem tersebut nantinya dapat dimanfaatkan untuk menentukan kualitas produk dengan penilaian secara objektif dan dapat juga dimanfaatkan untuk mengatasi masalah informasi yang jumlahnya besar. Sehingga analisis sentimen ini dapat digunakan untuk menemukan polaritas sentimen dari suatu pernyataan atau kalimat dan mengkategorikan menjadi kategori positif, netral serta negatif.

Analisis sentimen merupakan sebuah bidang yang mempelajari terkait opini, tanggapan, dan emosi dari seseorang terhadap sebuah entitas yakni salah satunya *review* produk [1]. Sentimen didalam bentuk teks mengandung informasi yang bersifat penting dan dapat diambil melalui *text mining*. *Text mining* merupakan tahapan awal dari proses analisis sentiment analisis dan sangat penting karena dimanfaatkan untuk mengidentifikasi emosional melalui pernyaataan. [3]. Lalu proses selanjutnya dilakukannya analisis sentimen.

Berbagai penelitian juga semakin berkembang dalam bidang analisis sentimen yang dimanfaatkan sebagai solusi yang sesuai karena sistem yang dirancang sudah dapat secara otomatis melakukan analisis terhadap *review* yang berada diinternet dan mengekstrak informasi yang memungkinkan paling relevan terhadap pengguna. Analisis sentimen dimanfaatkan untuk dapat memantau pendapat atau kecenderungan opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung berpandangan atau beropini negatif atau positif (Bo Pang, 2008)Berdasarkan pendapat Azam & Zao (2012) bahwa dengan mendapatkan informasi yang revelan dan tepat waktu dari keseluruhan *review* itu menjadi bagian yang sangat penting. Terdapat berbagai metode untuk melakukan analisis sentiment yakni Naïve Bayes sebagai salah satu pilihan pendekatan yang tepat. Pada penelitian ini akan menggunakan metode Naive Bayes. Naïve Bayes merupakan sebuah pendekatan yang berbasis proballistik. Metode Naïve Bayes Classifier merupakan metode sederhana namun disamping itu metode ini mempunyai nilai akurasi dan performansi yang tinggi dalam mengklasifikasikan sebuah teks [4]. Disamping itu terdapat juga dukungan pendapat dalam pemilihan metode Naïve Bayes yakni pada penelitian sebelumnya menyatakan bahwa naive bayes classifier cukup membutuhkan sejumlah kecil data pelatihan untuk mengestimasi parameter (varian dari kelas) yang diperlukan untuk klasifikasi (Chandra, et al., 2016).

Hasil akhir yang diharapkan pada penelitian analisis sentimen ini untuk mengidentifikasi sentiment positif dan negatif dari para *reviewer* pada *Amazon* dengan divisi *item food*. Klasifikasi pada label sentimen analisis yakni memakai penerapan dari ilmu ilmu linguistik *Natural Language Processing* dengan menerapkan tahapan *text preprocessing, text classification,* dan *language modeling*. Hasil dari analisis ini dapat dimanfaatkan sebagai media yang dapat memberikan pertimbangan pada pelanggan maupun penjual untuk mengambil keputusan dimasa mendatang berdasarkan pengelompokkan opini yang telah dilakukan.

## **1.2 Rumusan Masalah**

Berdasarkan pemaparan latar belakang dapat dirumuskan rumusan masalah didalam proyek Pemprograman Bahasa Alami ini adalah untuk bagaimana mengidentifikasi klasifikasi *review* pada amazon dengan kategori makanan (*food)* sehingga dapat memperoleh informasi mengenai kecendurungan opini pembeli melalui pendekatan Naïve Bayes *.*

## **1.3 Tujuan**

Tujuan proyek Pemprograman Bahasa Alami ini adalah mengimplementasikan pendekatan Naïve Bayes untuk mengidentifikasi *review* pengguna amazon didalam produk kategori *Food.*

## **1.4 Manfaat**

Manfaat yang diharapkan dari penelitian ini yakni membantu pihak Amazon dan penjual produk untuk mendapatkan *feedback* penting yakni *yang* sudah diklasifikasikan sehingga dapat langsung mendapatkan penialaian terhadap kualitas produk serta juga pengguna juga mendapatkan informasi penting dan relevan mengenai kualitas produk berdasarkan opini pengalaman pembeli sebelumnya sehingga pihak-pihak yang terkait didalam amazon dapat mengambil keputusan yang tepat

## **1.5 Ruang Lingkup**

Ruang lingkup dalam pengerjaan proyek ini adalah menggunakan metodeNaïve Bayes dengan data sampel yang diperoleh dari kaggle.com. Dataset yang digunakan adalah data komentar pengguna pada Amazon pada kategori *Food.*

# **BAB II. ISI**

## **2.1 Analisis Data**

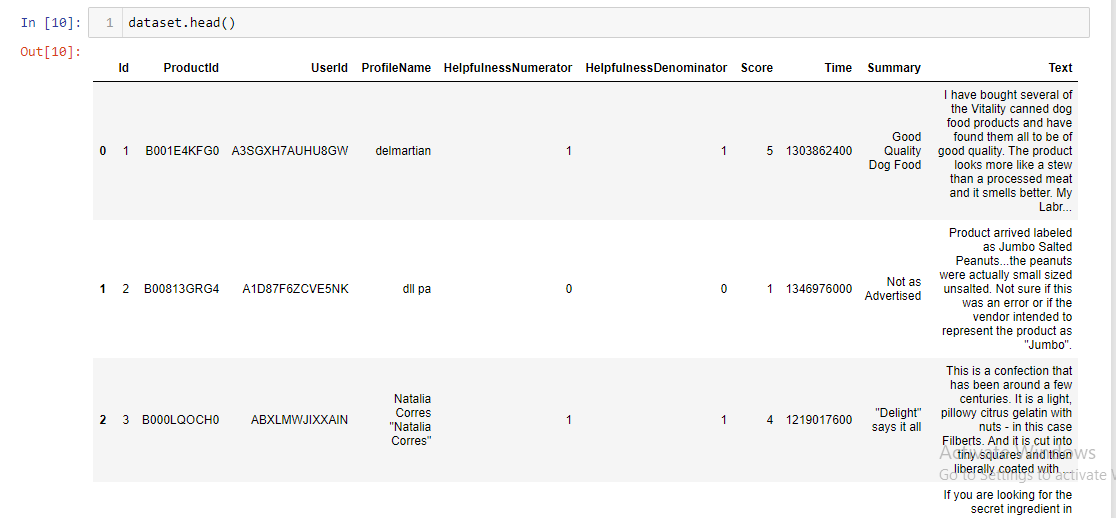
Pada subab ini akan menjabarkan hasil analisis yang telah dilakukan terhadap data serta metode yang akan dimanfaatkan pada proses implementasi pemrosesan bahasa alami. Data set yang digunakan didalam penelitian berasal dari *Amazon Fine Food review* yang diperoleh dari kaggle (<https://www.kaggle.com/snap/amazon-fine-food-reviews>).

### **2.1.1 Analisa Data**

Data ini akan berisi tentang kumpulan opini pembeli mengenai makanan sedap yang ada pada Amazon dengan jangka waktu lebih dari 10 tahun, termasuk 568.454 ulasan hingga Oktober 2012. Ulasan termasuk peringkat, informasi produk dan pengguna, dan ulasan teks biasa. Ini juga mencakup ulasan dari semua kategori Amazon lainnya.

Adapun attribut yang berada pada dataset ini yakni :

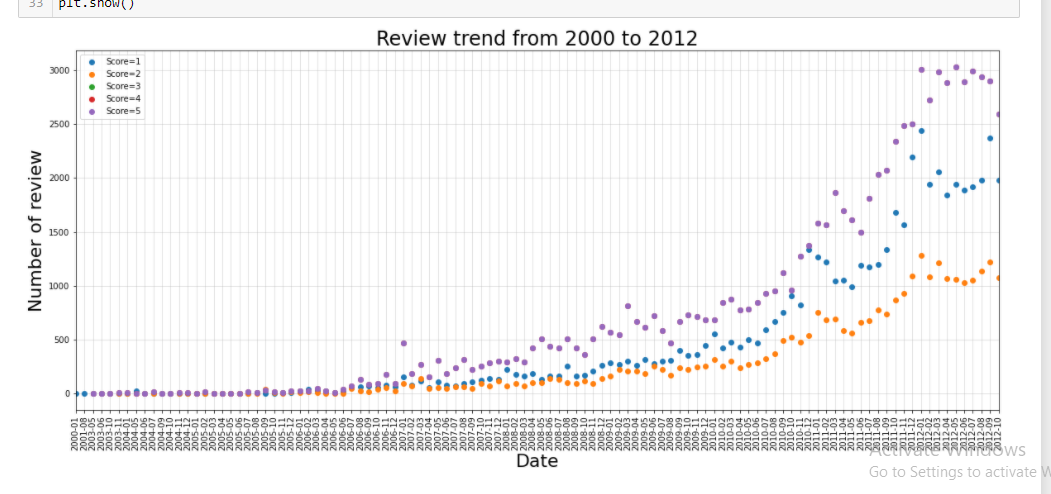
1. ProductId: Pengidentifikasi unik untuk produk
2. Id: pengenal unik untuk pengguna
3. ProfileName: Nama profil pengguna
4. HelpfulnessNumerator: Jumlah pengguna yang menganggap review bermanfaat
5. HelpfulnessDenominator: Jumlah pengguna yang menunjukkan apakah ulasan tersebut bermanfaat atau tidak
6. Score: Peringkat antara 1 dan 5
7. Time: Stempel waktu
8. Summary: Ringkasan review
9. Text : Review



Gambar 1. Kode prgram read data

Opini-Opini pengguna tersebut akan ditentukan positif dan negative menggunakan Score. Sebagai contoh dimana opini yang bernilai score 1 hingga 3 dikategorikan sebagai opini buruk dan 4 hingga 5 sebagai opini baik.

Untuk membantu analisis dapat ditinjau melalui tahapan eksplorasi data seperti melihat menganalisis tren ulasan.



Dari visualisasi diatas , peneliti dapat mengetahui dari tahun 2000 hingga 2006 jumlah tinjauan konsisten. Namun setelah itu, jumlah review mulai meningkat. Dari jumlah tersebut, sejumlah review dengan rating 5 bintang tergolong tinggi.

Dengan melakukan eksplorasi data kita dapat melihat kecendrungan seperti apa pendapat dari pembeli melalui visualisasi data berikut :

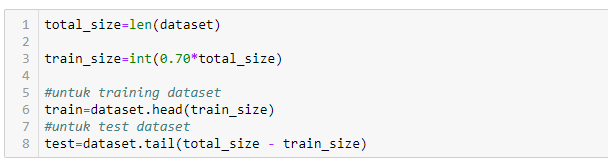


Terlihat bahwa adanya perbedaan ukuran, dan pendapat yang paling dominan yakni *good* dan *great.*

### **2.1.2 Analisa Metode**

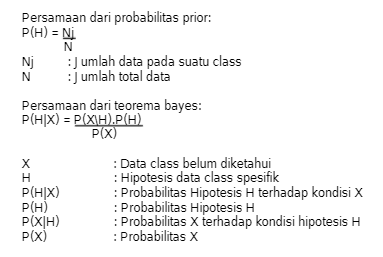
Naïve Bayes adalah teknik klasifikasi yang berlandaskan pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yakni melakukan prediksi peluang berdasarkan pengalaman di masa sebelumnya (Teorema Bayes) dengan karakteristik utama yakni asumsi yang paling kuat akan ketergantungan dari masing-masing kondisi.

Algoritma Naïve Bayes akan dimanfaatkan untuk menentukan apakah data yang akan diuji termasuk kedalam sentiment positf atau negatif. Dalam melakukan klasifikasi akan memerlukan data latih dan data testing sebagai proses mesin pembelajaran. Data latih dimanfatkan didalam metode Naïve Bayes sehingga diperoleh model klasifikasi untuk penentuan kelas pada data uji.



Naive Bayes menerapkan fungsi statistik sederhana berdasarkan teorema bayes dengan asumsi keberadaan dari suatu fitur tertentu terhadap suatu kelas yang tidak berhubungan dengan fitur lainnya (Wilianto, Pudjiantoro, & Umbara, 2017).

Rumus Naïve Bayes :



Gambar 2. Rumus Naïve Bayes

Sumber(<https://www.researchgate.net/publication/339234570_Penerapan_Algoritma_Naive_Bayes_Untuk_Analisis_Sentimen_Pada_Wisata_TMII_Berbasis_Website>)

## **2.2 Desain**

Pada bab ini akan dijabarkan mengenai tahapan pemrosesan bahasa alami yang akan diterapkan pada proyek. Adapun tahapan pemrosesan dapat dilihat pada aliran diagram pada Gambar 1.

Gambar 3. Tahapan Pemrosesan Bahasa Alami

Collecting Data

Load Data and Review Data

Explore Data

Text Pre-Processing

Evaluation

Modelling

Feature Extraction

1. **Collecting Data**

Pada tahap ini dilakukan pengumpulan data yang merupakan proses mengumpulkan dan memastikan informasi pada subjek yang akan dilakukan uji coba yang bertujuan untuk mendapatkan jawaban pertanyaan dari uji coba yang dilakukan, uji hipotesis dan mengevaluasi hasil. Proyek Sentimen Analisis akan dilakukan dengan menggunakan *dataset* yang berasal dari *kaggle* yang sudah di *download* dalam bentuk *file* .csv yaitu *Amazon Fine Food Review Sentiment Analysis* dengan 10 kolom dan 568454 baris. *Dataset* yang digunakan berisikan atribut mengenai Id, ProductId, UserId, ProfileName, HelpfulnessNumerator, HelpfulnessDenomitor, Score, Time, Summary, dan Text.

1. **Load Data**

Pada tahap ini dilakukan *load data* untuk menampilkan atribut dan isi dari *dataset* tersebut. Selain itu juga, tim proyek dapat memilih isi data dari atribut mana yang akan ditampilkan serta jumlah data yang ingin ditampilkan.

1. **Explore Data**

Pada tahap ini dilakukan ekpslorasi data untuk memungkinkan tim proyek memahami isi data yang digunakan. *Explore data* yang dilakukan menggunakan *Word Cloud* atau *Tag Cloud*, dimana *Word Cloud* ini merupakan representasi visual dari data teks dengan menggambarkan metadata atau dengan kata lain memvisualisasikan suatu bentuk teks secara bebas dengan berbagai ukuran (*font*) atau warna.

1. **Text Pre-Processing**

Pada tahap ini dilakukan *preprocessing data* dengan tujuan untuk menghilangkan *noise*, menyeragamkan bentuk kata, dan mengurangi volume kosakata. Adapun tahapan *preprocessing* ini meliputi:

1. *Convert into lower cases*

Proses ini dilakukan dengan mengkonversi semua huruf menjadi *lower case* atau huruf kecil. Hal ini dilakukan agar semua data yang masuk pada tahap berikutnya menjadi sama dan standar. Contohnya, kata “Pemrosesan Bahasa Alami” akan dikonversi menjadi kata “pemrosesan bahasa alami”. [5]

1. *Remove punctuations*

Proses ini dilakukan dengan menghapus tanda baca pada suatu kalimat. Contohnya, kata “this taffy is so good. it is very soft and chewy” menjadi “this taffy is so good it is very soft and chewy”.

1. *Remove stop words*

Proses ini dilakukan untuk memeriksa *stop word list* atau daftar kata-kata yang seharusnya dihilangkan. Proses ini dilakukan dengan membuat *token* atau *tokenize* dari kumpulan data terlebih dahulu, dimana kata yang kurang memiliki arti atau tidak memenuhi syarat sebagai kata akan dibuang. Namun, apabila kata-kata tersebut memang menjadi bagian dari sebuah kalimat, maka kata tersebut tidak dapat digunakan sebagai *token* karena kata tersebut berupa kata sambung atau keterangan. Misalnya “the”, “this”, “that”, “these”, “those”, “his”, “her”, “its”, “your”, “my”, dan lain-lain. [5] [6]

1. *Remove HTML tags*

Proses ini dilakukan untuk menghapus semua *tag* HTML yang ada pada daftar kata. Misalnya *tag* seperti “<a>”, “</a>”, “<br>”, “</br>”, dan lain-lain akan dihapus.

1. *Remove special characters*

Proses ini dilakukan untuk menghapus semua *spescial character* dari setiap kalimat. Misalnya “he’s learned about whales, india, drooping roses” menjadi “he s learned about whales india drooping roses”.

1. *Lemmatization*

Proses ini dilakukan untuk mengubah setiap *token* ke bentuk dasar umum dari kata tersebut. Pada proses ini, setiap kata akan diubah menjadi kata dasar dan membuat peningkatan fitur benar pada setiap dokumen yang mengakibatkan peningkatan pengambilan fitur yang terekstrak. Dalam kata lain, *lemmatization* adalah suatu proses untuk menemukan bentuk dasar dari sebuah kata [7] [8].

1. *Stemming*

Proses ini dilakukan dengan menghilangkan semua imbuhan, baik imbuhan awalan, sisipan, akhiran, maupun gabungan awalan dan akhiran. Proses *stemming* dilakukan juga karena pada proses sebelumnya masih terdapat kata yang sebenarnya memiliki arti yang sama namun berbeda bentuk kata. Hal ini dikarenakan pemakaian bentuk kata yang berbeda dalam setiap kalimat. Misalnya kata “disease”, “diseases”, “diseased”, akan diubah menjadi bentuk “disease” [5] [8]

1. **Feature Extraction**

*Feature extraction* merupakan sebuah proses mengekstrak informasi-informasi yang paling relevan untuk dikategorikan dengan meminimalkan variasi antar pola dalam suatu kategori dan sekaligus memaksimalkan variasi antar kategori yang berbeda dari data mentah yang tersedia [9]. *Feature extraction* yang dapat digunakan pada penelitian ini yaitu menggunakan metode *Bag of Words.*

Metode *Bag of Words* (BoW) merupakan representasi sederhana yang digunakan pada pemrosesan bahasa alami, dimana sebuah teks yang berupa kalimat atau dokumen diwakili sebagai kantung (*bag*) *multiset* dari kata-kata yang terkandung didalamnya tanpa memandang urutan kata dan tata bahasa, namun tetap mempertahankan keberagamannya. Pengertian lain dari BoW adalah sebuah model yang mempelajari sebuah kosa kata dari seluruh dokumen untuk kemudian memodelkan tiap dokumen dengan menghitung jumlah kemunculan setiap kata.

1. **Modelling**

Pada tahap ini dilakukan pembangunan sentimen analisis untuk *food review amazon* menggunakan *Naïve Bayes*. Metode *Naïve Bayes* digunakan untuk mengklasifikasikan sebuah teks, dimana metode ini menggunakan perhitungan probabilistik dengan cara melakukan penjumlahan terhadap frekuensi dan kombinasi dari sebuah dataset. Hasil dari proses klasifikasi ini kemudian akan diuji untuk menentukan tingkat akurasi sistem dalam melakukan proses klasifikasi.

1. **Evaluation**

*Evaluation* atau pengujian dilakukan dengan membagi data ke dalam data *training* dan data *testing*. Metode pengujian merupakan penilaian seberapa baik proses identifikasi pada sistem dengan menggunakan rumus umum perhitungan *accuracy*, *precision*, *recall*, dan *F-score* yang diperkenalkan oleh Baeza-Yates & Ribeiro-Neto [10].

* *Accuracy* adalah tingkat kedekatan antara nilai prediksi dengan nilai aktual.

Rumus yang digunakan untuk perhitungan *accuracy* dapat dilihat pada persamaan (1) dan (2) berikut.

(1)

(2)

* *Precision* adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem

Rumus yang digunakan untuk perhitungan *precicion* dapat dilihat pada persamaan (3) berikut.

(3)

* *Recall* adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

Rumus yang digunakan untuk perhitungan *recall* dapat dilihat pada persamaan (4) berikut.

(4)

* *F1-Score* adalah *harmonic mean* dari *precision* dan *recall*.

Rumus yang digunakan untuk perhitungan *f1-score* dapat dilihat pada persamaan (5) berikut.

(5)

Tabel 1. Variabel pada Metode Evaluasi

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Label Manual** | | |
| Positif | Negatif | Netral |
| **Hasil Identifikasi** | Positif | True Positif (TP) | False Positif (FP) | False Positif (FP) |
| Negatif | False Negatif (FN) | True Negatif (TN) | False Negatif (FN) |
| Netral | False Netral (FN) | False Netral (FNt) | True Netral (TNt) |

Keterangan:

* True Positif (TP) adalah kondisi data dengan label positif berhasil diidentifikasi sebagai kalimat positif
* False Positif (FP) adalah kondisi data dengan label positif gagal diidentifikasi sebagai kalimat positif
* True Negatif (TN) adalah kondisi data dengan label negatif berhasil diidentifikasi sebagai kalimat negatif
* False Negatif (FN) adalah kondisi data dengan label negatif gagal diidentifikasi sebagai kalimat negatif
* True Netral (TN) adalah kondisi data dengan label netral berhasil diidentifikasi sebagai kalimat netral
* False Netral (FN) adalah kondisi data dengan label netral gagal diidentifikasi sebagai kalimat netral

# **BAB III. IMPLEMENTASI**

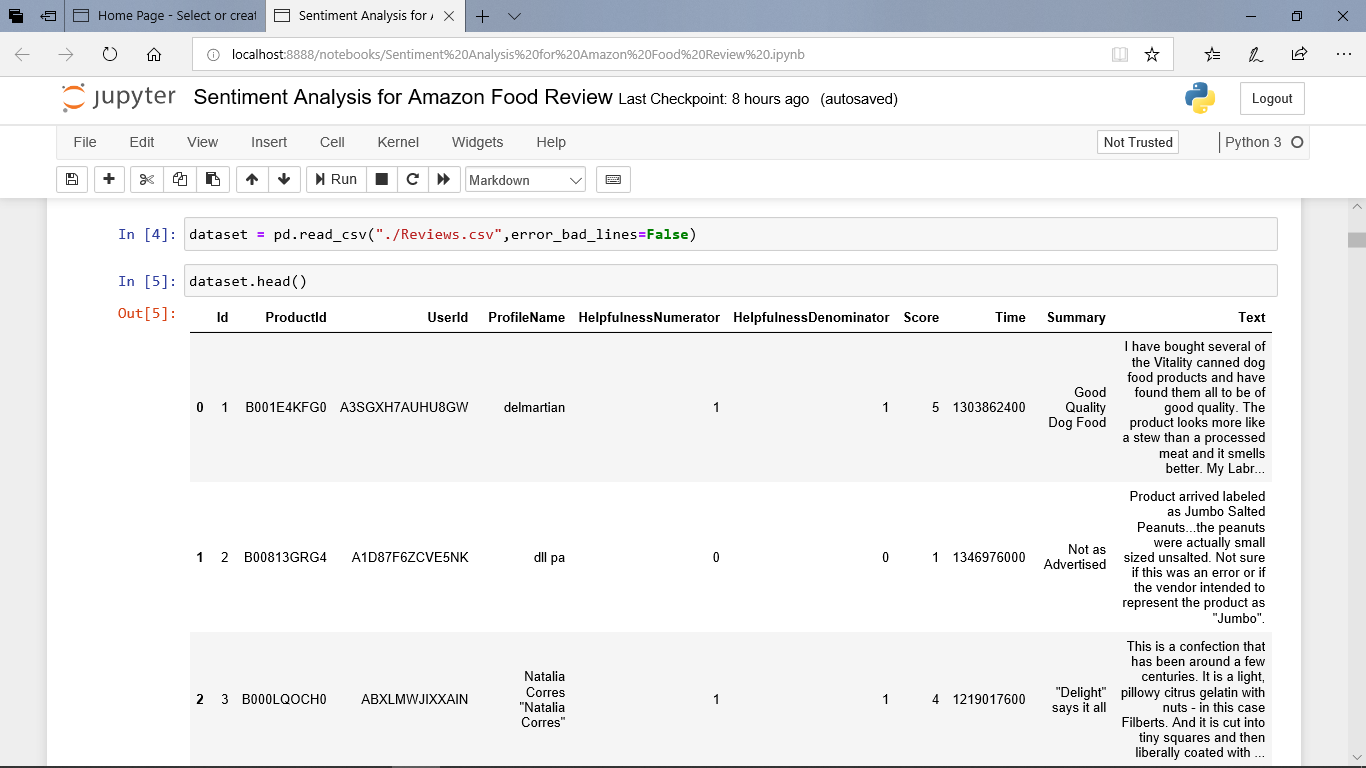
Pada bab ini akan dijabarkan mengenai implementasi proyek sentiment analisis menggunakan Naïve Bayes yang diimplementasikan pada jupyter notebook. Implementasi yang dilakukan pada proyek ini berdasarkan desain yang telah dijabarkan pada bab 2.

## **3.1 Implementasi *Collecting Data***

Data yang digunakan merupakan kumpulan *review* makanan dari pengguna Amazon, yang diperoleh melalui Kaggle dan dapat diakses pada situs berikut <https://www.kaggle.com/snap/amazon-fine-food-reviews> sebanyak 568,454 *food reviews.*

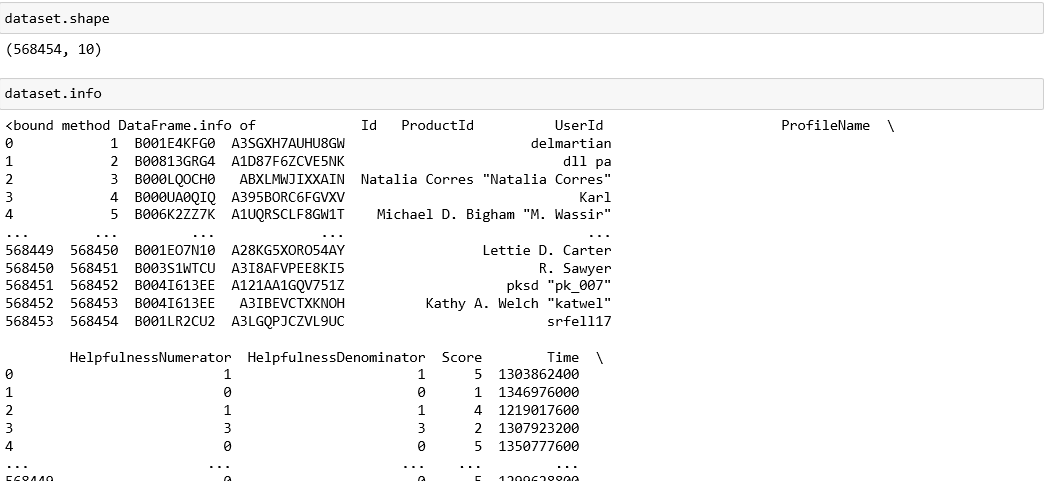
## **3.2 Implementasi *Load data* dan *Review Data***

Pada *jupyter notebook*  untuk melakukan *load* data dapat di tuliskan dengan kode program berikut, untuk melihat sebagai data yang sudah di*load* dapat mengunakan perintah .head() yang selanjutnya akan menampilkan beberapa data yang berasal dari *csv.*



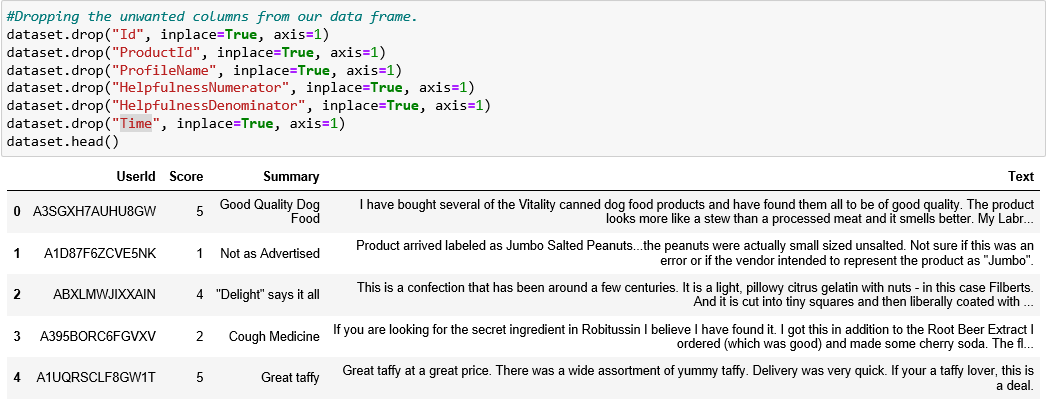
Gambar 4. Kode program load data

Perintah .shape dan .info akan menampilkan berapa banyak kolom dan baris pada dataset dan memberikan informasi mengenai dataset itu sendiri. Adapun kode program dari perintah tersebut dapat dilihat pada kode program dibawah ini.



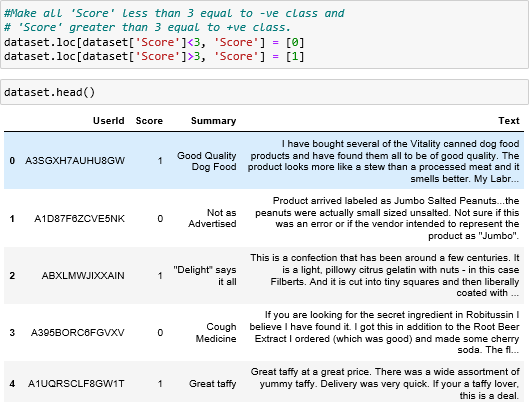
Gambar 5. Kode program menampilkan informasi data

Pada implementasi ini tim proyek melakukan *drop* pada kolom yang tidak diperlukan Id, ProductId, HelpfulnessNumerator, HelpfulnessDenominator, dan Time.



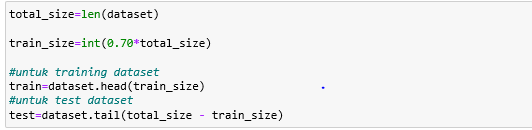
Gambar 6. Kode program drop kolom

Score yang diberikan pengguna memiliki ragam variasi disini dilakukan penskalaan dimana untuk produk yang memiliki nilai lebih atau sama dengan 3 akan mendapatkan nilai 1 dan untuk produk yang bernilai kurang dari 3 akan mendapatkan nilai 0.



Gambar 7. Kode Program melakukanscale nilai pada produk

Implementasi yang dilakukan melakukan *split* data untuk *trainee data* dan *test data* dimana pada *train data* digunakan 0.70 dari ukuran dataset sisanya akan digunakan untuk *test data.*



Gambar 8. Kode program split data

## **3.3 Implementasi Explore Data**

Pada tahap eksplorasi data akan dilakukan visualisasi review customer menggunakan Word Cloud. Word cloud (atau disebut juga tag cloud) adalah representasi visual dari data teks, biasanya digunakan untuk menggambarkan metadata atau untuk memvisualisasikan suatu bentuk teks secara bebas. Wordcloud (atau Tag cloud) adalah representasi visual dari data teks ini. Ini menampilkan daftar kata dengan berbagai ukuran font atau warna yang berguna untuk memahami istilah yang paling menonjol dengan cepat, untuk mengimplementasikan word cloud perlu dilakukan instlasi sebelumnya menggunakan pip install wordcloud.

Pada pengimplementasian *WordCloud* membutuhkan inputan *single string* dari teks, ringkasan *review* akan digabungkan menjadi single string, similarly akan dibangun melalui atribut Text.

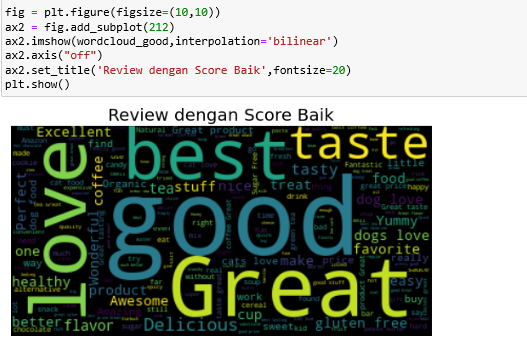


Gambar 9. Kode program wordcloud

Pada pengimplementasian review dengan score buruk akan digunakan *background* berwarna putih dan pengimplementasian review dengan score baik akan digunakan *background* berwarna hitam.



Gambar 10. Wordcloud review buruk

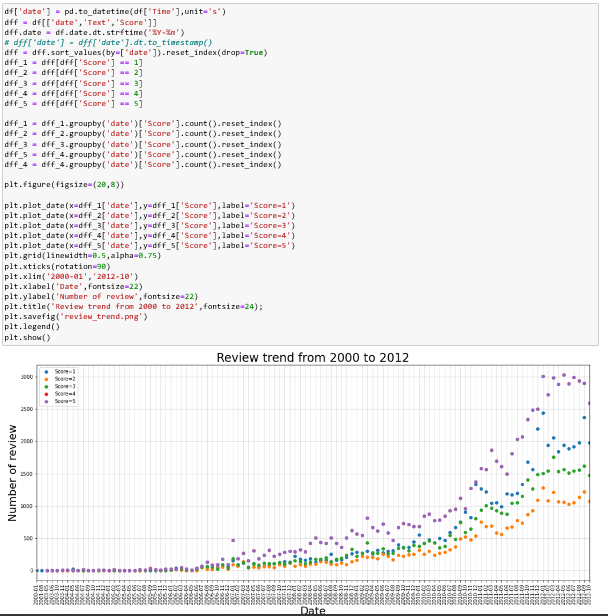


Gambar 11. Wordcloud review baik

Analisis produk dengan pengguna juga dilakukan pada implementasi ini dimana akan ditampilan id dari pengguna dan barang yang sudah dibeli.



Gambar 12. Kode program analisis produk dengan pengguna

Tim proyek melakukan analisi pada *review* pengguna dari tahun ketahun dimana rentang waktu nya adalah dari tahun 2000 sampai 2012 pada grafik dapat dilihat bahwa review pengguna dengan score 5 semkain meningkat setiap tahunnya.

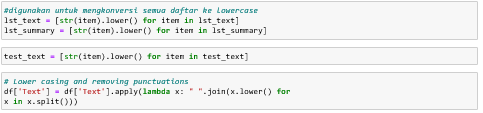
Gambar 13. Visualisasi review pengguna 2000-2012

## **3.4 Implementasi *Text Pre-Processing***

Pada bab ini akan dijabarkan mengenai tahapan pemrosesan bahasa alami yang akan diterapkan seperti *convert into lower cases, remove punctuations, remove stop words, remove HTML tags, remove special characters, lemmatization,* dan *stemming.*

### **3.4.1 Implementasi *Convert Into Lower Cases***

Proses *case folding* bertujuan untuk mengubah seluruh huruf kapital menjadi huruf kecil. Terdapat beberapa cara yang dapat dilakuka pada tahap *case folding,* seperti mengubah teks menjadi *lowercase,* menghapus angka, menghapus karakter kosong.



Gambar 14. Kode program lower case

### **3.4.2 Implementasi Remove Stopwords**

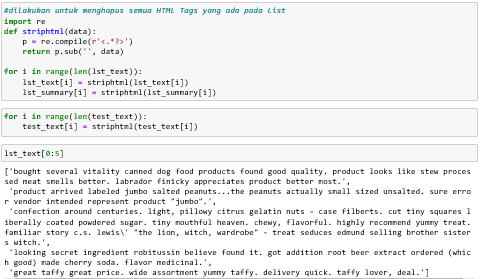
Kinerja klasifikasi akan menjadi lebih optimal dengan menghapus kata-kata yang jarang muncul dan akan mengurangi data sparsial dan menyusutkan ruang fitur secara substansial.



Gambar 15. Kode program remove stopwords

### **3.4.3 Implementasi Remove HTML Tags**

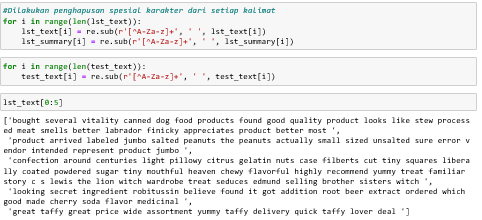
Pada implementasi juga dilakukan penghapusan pada tag html yang terdapat pada setiap daftar kata yang ada.



Gambar 16. Kode program Remove HTML Tags

### **3.4.4 Implementasi *Remove Special Characters***

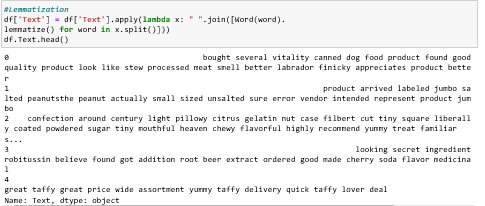
***Remove Special Characters*** bertujuan untuk menghilangkan karakter seperti angka, tanda baca dan hal lain yang buka merupakan alfabet a-z karena tanda baca yang terdapat didalam suatu kata nantinya akan sulit untuk diklasifikasikan karena tidak sesuai dengan kata yang terdapat pada kumpulan teks.



Gambar 17. Kode program remove special characters

### **3.4.5 Implementasi *Lemmatization***

Pada implementasi dlakukan lematisasi yang digunakan untuk mengubah kata yang memiliki imbuhan menjadi kata dasarnya.



Gambar 18. Kode program lematisasi

### **3.4.6 Implementasi *Stemming***

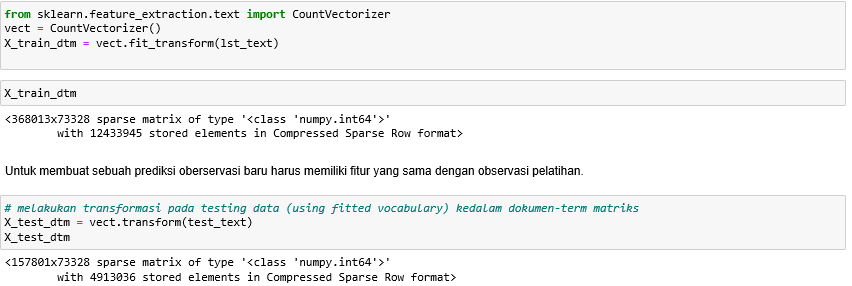
Proses *stemming* dilakukan untuk merubah kata-kata dalam dokumen menjadi kata dasarnya. Pada teks yang menggunakan bahasa inggris proses yang dilakukan hanya menghilangkan sufiks. Implemnetasi stemming pada proyek ini menggunakan stemming snowball.



Gambar 19. Kode program stemming

## **3.5 Implementasi Feature Extraction**

*Feature extraction* merupakan sebuah proses mengekstrak informasi-informasi yang paling relevan untuk dikategorikan dengan meminimalkan variasi antar pola dalam suatu kategori dan sekaligus memaksimalkan variasi antar kategori yang berbeda dari data mentah yang tersedia. *Feature extraction* yang digunakan pada penelitian ini yaitu menggunakan metode *Bag of words.*



Gambar 20. Kode program feautre extraction

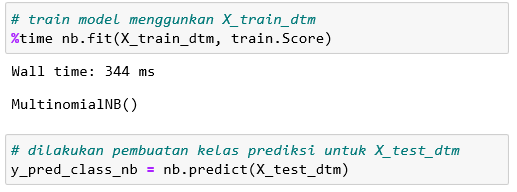
Keterangan :

1. vect.fit(lst\_text) belajar dengan kosakata dari training data
2. vect.transform(lst\_text) belajar dengan kosakata yang pas untuk membangun document-term matrix dari training data
3. vect.transform(test\_text) belajar dengan kosakata yang pas to build a document-term matrix dari testing data (dan mengacuhkan tokens yang tidak pernah dilihat sebelumnya.

## **Implementasi *Modelling***

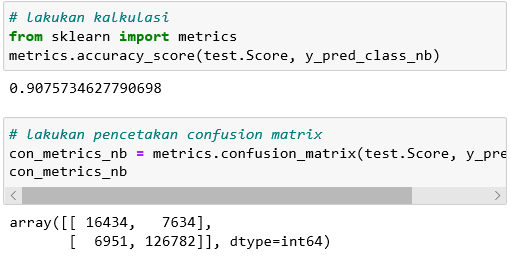
Pembangunan sentimen analisis untuk *food review amazon* menggunakan *Naïve Bayes.* Implementasi ini menggunakan *naïve bayes* dimana akan dilakukan import library MultinomialNB untuk mengimport library Multinomial Naïve Bayes.

Pada implementasi akan dilakukan pembuatan kelas prediksi untuk X\_test\_dtm



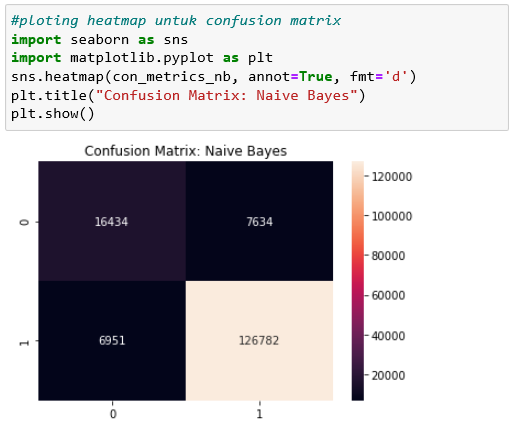
Gambar 21. Kode program implementasi klasifikasi

Pada implementasi akan dilakukan kalkulasi untuk melihat nilai akurasi dan akan dilakukan pencetakan *confusion matrix.*



Gambar 22. Kode program kalkulasi akurasi

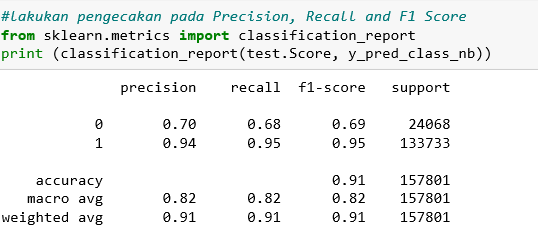
Iplementasi ini menggunakan heatmap untuk *confusion matrix* dimana pembauatannya dapat dilihat pada implementasi dibawah ini.



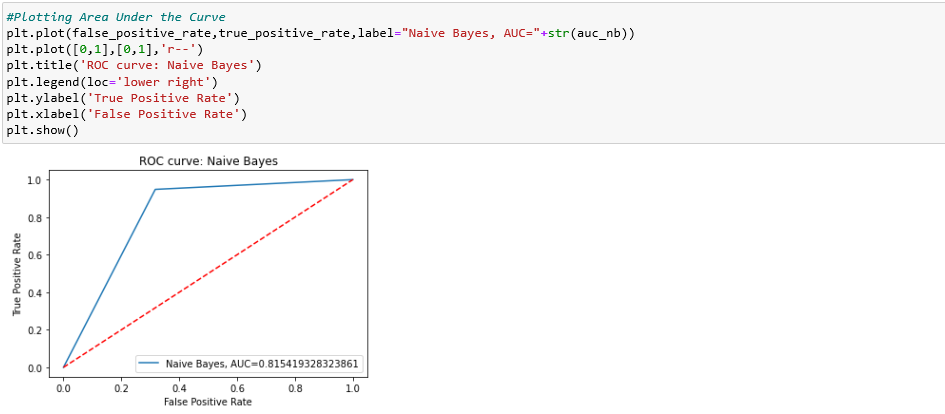
Gambar 23. Implementasi heatmap

## **Implementasi Evaluation**

*Evaluation* atau pengujian dilakukan dengan membagi data ke dalam data *training* dan data *testing*. Metode pengujian merupakan penilaian seberapa baik proses identifikasi pada sistem dengan menggunakan rumus umum perhitungan *accuracy*, *precision*, *recall*, dan *F-score.*



Gambar 24. Kode program evaluasi



* *Accuracy* adalah tingkat kedekatan antara nilai prediksi dengan nilai aktual.

Rumus yang digunakan untuk perhitungan *accuracy.*

* *Precision* adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem
* *Recall* adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.

Rumus yang digunakan untuk perhitungan *recall*

* *F1-Score* adalah *harmonic mean* dari *precision* dan *recall*.

# **BAB IV. PENUTUP**

Pada bab ini dijelaskan mengenai pembagian tugas dan tanggung jawab, kesimpulan, dan saran pada kelanjutan pengembangan proyek yang dilakukan.

## **4.1 Kesimpulan**

Dalam Analisis sentimen review restoran Fish Streat terdapat 2 proses, yaitu proses training dan proses testing. Proses training bertujuan untuk membangun model analisis sentimen dari data latih dengan mengitung probabilitas. Proses testing bertujuan untuk menentukan klasifikasi sentimen terhadap data yang akan di uji cobakan. Pengukuran akurasi dari sistem yang dibuat menggunakan confusion matrix. Berdasarkan hasil implementasi dan pengujian yang telah dilakukan pada analisis sentiment didapatkan tingkat akurasi sebesar 0,91%. Berdasarkan penelitian yang telah dilakukan, dalam analisis sentimen penggunaan metode Naïve Bayes Classifier akan menimbulkan tingkat akurasi yang tinggi. Semakin banyak data yang diujikan maka tingkat akurasi dari hasil klasifikasi akan semakin tinggi. Hal ini juga dipengaruhi oleh jumlah data latih yang digunakan pada metode Naïve Bayes Classifier. Kedepannya diharapkan dalam penelitian ini dapat menggunakan metode lain sebagai pembanding dari metode Naïve Bayes Classifier.

## **4.2 Saran**

Berdasarkan pengerjaan proyek yang telah dilakukan, masih terdapat beberapa kekurangan yang perlu diperbaiki. Saran yang diberikan adalah dengan menyediakan antarmuka pengguna agar dapat digunakan oleh perusahaan untuk memudahkan memahami informasi yang diberikan oleh perusahaan atau organisasi.

## **4.3 Pembagian Tugas**

Pada subab ini akan dijelaskan mengenai pembagian tugas tiap anggotan kelompok.

Tabel 2. Pembagian Tugas

|  |  |  |
| --- | --- | --- |
| Team Members | Roles | Task |
| Reza Oktovian Siregar | Data Engineer | Bertanggung jawab untuk membersihkan, memproses data yang sudah dikumpulkan. |
| Angelia Regina Ginting | Data Engineer | Bertanggung jawab untuk membersihkan, memproses data yang sudah dikumpulkan. |
| Christina Clara | Team Leader | Bertanggung jawab untuk membersihkan, memproses data yang sudah dikumpulkan. |

# **Referensi**

|  |  |
| --- | --- |
| [1] | S. K. Hendry Ardian, "Analisis Sentimen Pada Review Produk Kosmetik Bahasa Indonesia Dengan Metode Naive Bayes," Jurnal ENTER, Pontianak. |
| [2] | K. A. Mohamad Syahrul Mubarok, "Aspect-based sentiment analysis to review products using Naïve Bayes, Muhammad Dwi Aldhi," International Conference on Mathematics: Pure, Applied and Computation, 2017. |
| [3] | R. G. M. D. M. H. Lei Zhang, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis," Hewlett-Packard Development Company, Chicago, 2011. |
| [4] | L. K. O. S. Ni Putu Sri Merta Suryani, Penggunaan Metode Naïve Bayes Classifier pada Analisis Sentimen Facebook Berbahasa Indonesia, Majalah Ilmiah Teknologi Elektro, 2019. |
| [5] | E. Junianto and R. Rachman, "Penerapan Metode Naive Bayes Classifier Untuk Mendeteksi Emosi pada Komentar Media Sosial," *Jurnal Responsif,* vol. 2, no. 1, pp. 1-8, 2020. |
| [6] | A. P. Wijaya and H. A. Santoso, "Naive Bayes Classification pada Klasifikasi Dokumen Untuk Identifikasi Konten E-Government," *Journal of Applied Intelligent System,* vol. 1, no. 1, pp. 48-55, 2016. |
| [7] | S. M. Derwin Suhartono, "Natural Language Processing," Binus University, [Online]. Available: https://socs.binus.ac.id/2013/06/22/natural-language-processing/#:~:text=Ingason%20dkk.%20(2008)%20mengemukakan,bentuk%20dasar%20dari%20sebuah%20kata.&text=Lemma%20adalah%20bentuk%20dasar%20dari,arti%20tertentu%20berdasar%20pada%20kamus.. [Accessed 08 Januari 2020]. |
| [8] | I. M. Yulietha, S. A. Faraby and Adiwijaya, "Klasifikasi Sentimen Review Film Menggunakan Algoritma Support Vector Machine," in *e-Proceeding of Engineering*, Bandung, 2017. |
| [9] | E. P. P, D. Puspitaningrum and A. Mirfen, "Identifikasi Tanda Tangan dengan Pendekatan Support Vector Machine," *Jurnal Sains, Teknologi dan Industri,* vol. 12, no. 2, pp. 225-231, 2015. |
| [10] | F. Z. Alif, *Ekstraksi Fitur untuk Pemilihan Topik Spesifik Review Film dalam Menghasilkan Aspect-Based Sentiment Analysis,* Medan, 2020. |
| [11] | R. Putra, "PENERAPAN ASSOCIATION RULE DENGAN ALGORITMA A-PRIORI UNTUK SISTEM REKOMENDASI PADA PERUSAHAAN E-COMMERCE X," UNIVERSITAS KATOLIK PARAHYANGAN, Bandung, 2018. |
| [12] | W. A. Triyanto, "ASSOCIATION RULE MINING UNTUK PENENTUAN REKOMENDASI PROMOSI PRODUK," Universitas Muria Kudus, 2014. |
| [13] | R. A. H. A. M. Nugroho Wand, "Pengembangan Sistem Rekomendasi Penelusuran Buku Dengan Penggalian Association Rule Menggunakan Algoritma Apriori (Studi Kasus Badan Perpustakaan Dan Kearsipan Provinsi Jawa Timur)," Institut Teknologi Sepuluh Nopember (ITS), Surabaya, 2012. |
| [14] | S. Mujilahwati, "PRE-PROCESSING TEXT MINING PADA DATA TWITTER," *SENTIKA 2016,* pp. 18-19, 2016. |
| [15] | A. . M. Pravina, . I. Cholissodin and . P. . P. Adikara, "Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. III, no. 3, p. 4, 2019. |
| [16] | R. M. Reese, in *Natural Language Processing with Java*, 2015. |
| [17] | W. . A. Luqyana, . I. Cholissodin and . R. . S. Perdana, "Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. II, no. 11, p. 4, 2018. |
| [18] | P. Bajaj, "Creating linear kernel SVM in Python," GeeksforGeeks, 20 June 2018. [Online]. Available: https://www.geeksforgeeks.org/creating-linear-kernel-svm-in-python/. [Accessed 22 November 2020]. |
| [19] | F. Rahutomo, D. S. E. Ikawati and O. A. Rohman, "Evaluasi Fitur Word2Vec pada Sistem Ujian Esai Online," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika),* vol. 04, no. 01, pp. 36-45, 2019. |
| [20] | A. F. Niasita, P. P. Adikara and S. Adinugroho, "Analisis Sentimen Pembangunan Infrastruktur di Indonesia dengan Automated Lexicon Word2Vec dan Naive-Bayes," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 3, no. 3, pp. 2673-2679, 2019. |