**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**PROPOSAL PROYEK AKHIR PEMROSESAN BAHASA ALAMI**



**Sentiment Analysis Hotel Based on Review with Comparing Naive Bayes and SVM**

**OLEH:**

|  |  |
| --- | --- |
| **12S17001** | **Krista Lumbantoruan** |
| **12S17042** | **Aulia S L Pakpahan** |
| **12S17045** | **Evi Fanny Dear Sidabutar** |
| **12S17056** | **Maria Anastasia Br. Simanullang** |

**11S4054 –PEMPROSESAN BAHASA ALAMI**

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**NOVEMBER 2020**

# DAFTAR ISI

[DAFTAR ISI ii](#_Toc56964587)

[DAFTAR TABEL iii](#_Toc56964588)

[DAFTAR GAMBAR iv](#_Toc56964589)

[BAB 1 PENDAHULUAN 1](#_Toc56964590)

[1.1 Latar Belakang 1](#_Toc56964591)

[1.2 Rumusan Masalah 2](#_Toc56964592)

[1.3 Tujuan 2](#_Toc56964593)

[1.4 Ruang Lingkup Proyek 2](#_Toc56964594)

[BAB 2 ISI 4](#_Toc56964595)

[2.1 Landasan Teori 4](#_Toc56964596)

[*2.1.1* *Sentiment Analysis* 4](#_Toc56964597)

[2.1.2 *Support Vector Machine (SVM)* 5](#_Toc56964598)

[*2.1.3* *Naïve Bayes* 5](#_Toc56964599)

[2.2 Tahapan Pengerjaan Proyek 8](#_Toc56964600)

[BAB 3 RENCANA KERJA 10](#_Toc56964601)

[3.1 Pembagian Tugas 10](#_Toc56964602)

[3.2 Jadwal 10](#_Toc56964603)

[DAFTAR PUSTAKA 12](#_Toc56964604)

# DAFTAR TABEL

[Tabel 1. Pembagian Tugas 10](#_heading=h.z337ya)

[Tabel 2. Jadwal 10](#_heading=h.1y810tw)

# DAFTAR GAMBAR

[Gambar 1. Metode Pengerjaan Proyek 8](#_heading=h.1ksv4uv)

# BAB 1 PENDAHULUAN

Bab ini menjelaskan mengenai latar belakang, tujuan, manfaat dan ruang lingkup dari proyek.

## **Latar Belakang**

Hotel merupakan bangunan yang memiliki kamar dalam jumlah banyak yang memang sengaja dibangun untuk dijadikan tempat istirahat atau penginapan sementara bagi para pengunjung. Hotel menyediakan jasa layanan penginapan, makanan dan minuman dengan harga yang bervariasi. Semakin berkembangnya perubahan gaya hidup, hotel saat ini tidak hanya berfungsi sebagai tempat menginap, tetapi juga dimanfaatkan orang untuk rapat, seminar, bahkan sampai pertunjukan dan pameran. Dengan begitu banyak fungsinya semakin banyak hotel yang dibangun maka semakin bingung juga untuk memilih ingin menginap atau menyewa hotel yang lebih baik dan sesuai dengan keinginan. Baik wisatawan maupun orang yang memiliki kepentingan dan mengharuskan menginap atau menyewa hotel seringkali mengalami kesulitan untuk memilih hotel. Mungkin ada beberapa aplikasi yang membantu dalam masalah ini dengan beberapa fitur andalannya seperti fitur komentar. Tetapi banyaknya komentar juga membuat calon pengunjung hotel bingung dan butuh waktu lama untuk memilih hotel mana yang sesuai dengan keinginan. Berdasarkan permasalahan tersebut, maka dibutuhkan adanya sistem analisis sentimen yang mampu mengolah komentar-komentar tersebut, sehingga data yang diperoleh dapat diubah menjadi informasi yang sangat bermanfaat. Sistem tersebut dibangun dengan tujuan membuat model *sentiment review* hotel.

Abdul Muhaimin Rahat, Abdul Kahir, Abu Kaisar Mohammad Masum [1] melakukan penelitian untuk membandingkan Algoritma Naive Bayes dan SVM pada sentimen analisis berdasarkan komentar. Pada penelitian ini mereka menggunakan dataset komentar pelanggan di twitter tentang penerbangan. Penelitian ini dilakukan untuk mendapatkan tindakan evaluasi atas komentar yang diberikan oleh para penumpang. hasil dari komentar tersebut digunakan untuk meningkatkan kualitas maskapai penerbangan kedepannya. Setiap komentar dianalisis apakah masuk ke dalam golongan positif atau negatif. Pada penelitian ini dihasilkan tingkat akurasi pada SVM sebesar 83% dan Naive Bayes 77%.

Dany Pratmanto, Husni Faqih, Warjiyono [2] melakukan penelitian mengenai *review analysis sentiment* pada hotel kelas dunia menggunakan metode *Naive Bayes Classifier* dan *Particle Swarm Optimization.* pada penelitian ini menggunakan komentar di media sosial dengan mengambil 200 data review. Pengujian model yang dilakukan dengan menggunakan *Naive bayes* menghasilkan tingkat akurasi 66,67% sementara pengujian menggabungkan *Naive Bayes* dengan *Particle Swarm Optimization* menghasilkan nilai terbaik dengan 85% tingkat akurasinya. Penelitian ini menarik kesimpulan bahwa *Naive Bayes* akan memberikan hasil terbaik jika digabungkan dengan *Particle Swarm Optimization*.

Berdasarkan penelitian terdahulu, pada proyek ini penulis akan melakukan penelitian mengenai *sentiment analysis hotel based on review* dengan membandingkan dua metode yakni algoritma *Naïve Bayes* untuk melakukan klasifikasidan *Support Vector Machine* untuk menentukan opini berdasarkan hasil analisis sentimen. Pada penelitian ini, akan dilakukan perbandingan hasil model sentiment melalui kedua metode tersebut untuk menemukan hasil atau cara yang lebih baik dalam analisis sentimen yang akan mengacu pada emosi atau perasaan. Dengan banyaknya komentar-komentar maka akan meningkatkan data tekstual yang perlu dianalisis. Analisis sentimen ini ditujukan untuk menentukan apakah polaritas korpus tekstual tersebut cenderung positif, negatif dan netral. Penelitian ini juga memberikan manfaat bagi para calon pengunjung hotel untuk dijadikan sebagai acuan untuk memilih hotel, juga dapat bermanfaat untuk manajemen hotel dalam perbaikan atau peningkatan kualitas pelayanan maupun fasilitas.

## **Rumusan Masalah**

Beberapa rumusan masalah dalam proyek ini adalah:

1. Bagaimana Naïve Bayes dan SVM dapat memberikan sentimen analisis terhadap hotel berdasarkan komentar?
2. Bagaimana perbandingan hasil metode Naïve Bayes dan SVM pada *sentiment analysis*?

## **Tujuan**

Yang menjadi tujuan dalam proyek ini adalah:

1. Mengetahui cara Naïve Bayes dan SVM dalam membangun model *sentiment analysis* berdasarkan komentar.
2. Mengetahui hasil perbandingan metode antara Naïve Bayes dan SVM dalam menemukan model *sentiment analysis*.

## **Ruang Lingkup Proyek**

Adapun yang menjadi ruang lingkup penelitian ini adalah:

1. Penelitian ini menggunakan *dataset* dari komentar wisatawan dan pengunjung terhadap sebuah hotel.
2. Algoritma yang digunakan pada proyek ini adalah *Naïve Bayes Classification* dan *Support Vector Machine.*

# BAB 2 ISI

## **Landasan Teori**

Bab ini menjelaskan tentang landasan teori yang memuat teori-teori yang digunakan oleh penulis dalam proyek Pemrosesan Bahasa Alami yaitu teori mengenai Sentiment Analysis, Support Vector Machine dan Naïve Bayes.

### ***Sentiment Analysis***

Menurut E. Haddi [3], *Sentiment Analysis* merupakan proses investigasi pada komentar-komentar pada sebuah produk, tempat, penerbangan dan yang lainnya di internet untuk mengetahui pendapat atau perasaan mereka terhadap sesuatu. *Sentiment Analysis* merupakan sebuah tugas klasifikasi untuk mengklasifikasikan orientasi teks menjadi positif atau negatif. *Sentiment Analysis* merupakan proses menganalisa pendapat seseorang terhadap sesuatu yang dikenakan, dikunjungi, dibeli atau semacamnya berdasarkan komentar yang diberikan. *Sentiment Analysis* ini mengacu pada penggunaan pemrosesan bahasa alami yang menerapkan *text analytic*s untuk menemukan berbagai sumber data dari beberapa *platform* sosial media dan internet dengan tujuan mendapatkan pendapat pengguna terhadap *platform* tersebut.

*Sentiment Analysis* memiliki beberapa tipe diantaranya, *Fine-Grained Sentiment Analysis* merupakan analisis untuk mengelompokkan opini atau pendapat ke dalam beberapa kategori yaitu positif, negatif, netral. Tipe kedua adalah *Intent Sentiment Analysis* yang bertujuan mengidentifikasi motivasi di balik komentar pengguna apakah komentar tersebut merupakan, saran, pendapat, pertanyaan, keluhan atau bahkan penghargaan. Dan tipe yang terakhir adalah *Aspect-Based Sentiment Analysis* fokus pada elemen yang spesifik dari produk dan layanan. Analisis tipe ini memungkinan untuk menghubungkan sentimen spesifik dari aspek produk atau layanan [4]

Di dalam *sentiment analysis*, teks ulasan *(review)* sangat penting dalam pengambilan keputusan berdasarkan analisis yang dilakukan. Oleh karena itu ketika mengumpulkan ulasan *(review)* dan mengolahnya menjadi sebuah masukan dalam sistem *sentiment analysis*, maka harus terlebih dahulu dipahami terminology apa saja yang berhubungan dengan *sentiment analysis*. Dan kemudian informasi tersebut akan diklasifikasi menjadi dua bagian utama yaitu positif dan negatif. Ada berbagai jenis metode yang dapat digunakan dalam *sentiment analysis*, salah satu diantaranya adalah klasifikasi dengan menggunakan *machine learning.* *Machine Learning* adalah sebuah teknik pendekatan yang digunakan di dalam *artificial intelligence* untuk menggantikan atau menirukan tingkah laku manusia dalam menyelesaikan masalah atau untuk melakukan otomasi. Metode yang paling banyak digunakan di dalam *machine learning* adalah algoritma *Naïve Bayes, Support Vector Machine, Logistic Regression, dan Neural Network.* Dalam proyek ini, algoritma yang akan digunakan adalah *Support Vector Machine* dan *Naive Bayes*.

### ***Support Vector Machine (SVM)***

Support Vector Machine (SVM) adalah sebuah algoritma yang merupakan bagian dari *machine learning* yang dapat digunakan untuk klasifikasi atau masalah regresi. Algoritma ini menggunakan teknik yang disebut trik kernel untuk mengubah data kemudian hasil transformasi menemukan batas optimal antara kemungkinan keluaran. Secara sederhana, ia melakukan beberapa transformasi data yang sangat kompleks, lalu mencari tahu cara memisahkan data berdasarkan label atau keluaran yang telah ditentukan. SVM pertama kali diusulkan oleh Cortes dan Vapnik pada tahun 1995 yang menunjukkan banyaknya keuntungan unik dalam memecahkan masalah pengenalan pola dimensi kecil, nonlinier dan tinggi [5]. *Support Vector Machine* atau SVM merupakan *universal learner* dimana SVM ini menentukan format untuk *input* dan *output.* Keluaran dari algoritma ini bisa menyatakan positif atau negatif. Cara kerjanya, dokumen teks yang tidak cocok akan di*transformasikan* menjadi format yang terstruktur kemudian dimasukkan ke *machine learning.* Skor teks tersebut kemudian akan dihitung dan dijadikan sebagai inputan ke SVM. SVM ini sudah sangat sering digunakan untuk melakukan kategorisasi pada teks [1].

### ***Naïve Bayes***

Naïve Bayes merupakan sebuah algoritma yang menggunakan metode supervised learning serta metode statistik untuk klasifikasi. Klasifikasi Bayes mengasumsikan model probabilistik dasar dan memungkinkan kita untuk menangkap ketidakpastian tentang model dengan cara dasar dengan cara menentukan probabilitas hasil. Klasifikasi Bayes dapat memecahkan masalah diagnostik dan prediksi. Metode Naïve Bayes merupakan sebuah metode klasifikasi yang menggunakan perhitungan probabilitas dengan cara membandingkan nilai probabilitas suatu sampel berada di suatu kelas dengan nilai probabilitas di kelas lainnya dengan menggunakan konsep dasar teorema Bayes. Metode ini dapat digunakan untuk text classification, spam filtering, dan hybrid recommender system [6].

Multinomial Naïve Bayes merupakan versi Naïve Bayes yang dirancang untuk dokumen teks. Naïve Bayes sederhana memodelkan dokumen sebagai ada dan tidaknya kata tertentu, sementara Multinomial Naïve Bayes secara eksplisit memodelkan jumlah kata dan menyesuaikan perhitungan dasar untuk disepakati.

Untuk menggunakan Naïve Bayes classifier dalam teks, perlu diketahui posisi kata, secara sederhana dengan memberikan index pada setiap posisi kata dalam dokumen seperti pada Persamaan (2.11).

Posisi ← semua posisi kata dalam dokumen test

C\_NB=argmax┬cϵC⁡〖P(c)〗 ∏\_(i∈position)▒〖P(w\_i |c)〗

Keterangan:

C\_NB : Probabilitas kelas C

P(c) : probabilitas kelas c

P(w\_i |c) : Probabilitas frekuensi kata i diberikan kelas c (2.11)

Pada Persamaan (2.12), perhitungan Naïve Bayes, seperti perhitungan untuk pemodelan bahasa dikerjakan dalam logspace, untuk menghindari underflow dan menambah kecepatan.

C\_NB=argmax┬cϵC⁡log⁡〖P(c)〗 +∑\_(i∈position)▒logP(w\_i |c)

Keterangan:

C\_NB : Probabilitas kelas C

P(c) : probabilitas kelas c

P(w\_i |c) : Probabilitas frekuensi kata i diberikan kelas c (2.12)

Dengan menentukan fitur dalam log space, persamaan tersebut menghitung kelas yang diprediksi sebagai fungsi linear dari fitur-fitur input. Classifier yang menggunakan kombinasi linear dari input untuk membuat klasifikasi disebut linear classifier.

Dalam Naïve Bayes terdapat istilah prior probabilities yang dapat diperoleh dengan menggunakan Persamaan (2.13), yaitu menghitung jumlah dokumen dalam data train dengan kelas c dan total jumlah dokumen.

P(c)=N\_c/N\_doc

Keterangan:

P(c) : probabilitas kelas c

N\_c : jumlah dokumen dalam data train yang merupakan kelas c

N\_doc : total jumlah dokumen (2.13)

Untuk mempelajari probabilitas P(w\_i |c) yang dihitung sebagai bagian dari berapa kali kata wi muncul pada semua kata dalam semua dokumen dengan kelas c, semua dokumen dengan kategori c digabungkan dalam satu teks ‘kategori c’ besar. Kemudian pada Persamaan (2.14), frekuensi digunakan dalam dokumen yang digabungkan ini untuk memberikan peluang yang memiliki maximum likelihood estimate.

P(w\_i |c)=(〖count (w〗\_(i ),c))/(∑\_(w∈V)▒〖〖count (w〗\_ ,c)〗)

Keterangan:

P(w\_i |c) : Probabilitas frekuensi kata i diberikan kelas c

〖count (w〗\_(i ),c) : jumlah kata muncul pada kelas c

∑\_(w∈V)▒〖〖count (w〗\_i,c)〗 : jumlah semua kata dalam semua dokumen dengan kelas c (2.14)

Dimana vocabularyV merupakan gabungan dari semua jenis kata dalam semua kelas, bukan hanya kata dalam kelas c.

Pada Persamaan (2.15) dapat diketahui bahwa jika terdapat kasus dimana probabilitas fitur adalah 0, maka Naïve Bayes akan secara ‘naif’ mengalikan semua kemungkinan fitur bersamaan, probabilitas nol dalam likelihood term untuk setiap kelas akan menyebabkan probabilitas kelas menjadi nol.

P(w\_i│c)=(〖count (w〗\_(i ),c))/(∑\_(w∈V)▒〖〖count (w〗\_i,c)〗)=0

Keterangan:

P(w\_i |c) : Probabilitas frekuensi kata i diberikan kelas c

〖count (w〗\_(i ),c) : jumlah kata wi muncul pada kelas c

∑\_(w∈V)▒〖〖count (w〗\_i,c)〗 : jumlah semua kata dalam semua dokumen dengan kelas c (2.15)

Laplace smoothing (add-one) merupakan sebuah solusi sederhana untuk menangani kasus tersebut. Laplace smoothing dilakukan dengan cara menambahkan angka 1 seperti pada Persamaan (2.16):

P(w\_i│c)=(〖count (w〗\_(i ),c)+1)/(∑\_(w∈V)▒〖〖count (w〗\_i,c)+1〗)

Keterangan:

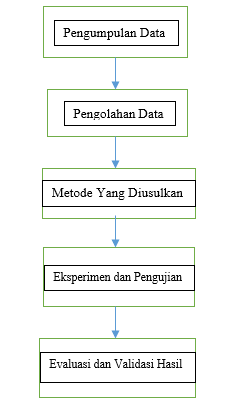
P(w\_i |c) : Probabilitas frekuensi kata i dari kelas c

〖count (w〗\_(i ),c) : jumlah kata muncul pada kelas c

∑\_(w∈V)▒〖〖count (w〗\_i,c)〗 : jumlah semua kata dalam semua dokumen dengan kelas c (2.16)

## **Tahapan Pengerjaan Proyek**

Menjelaskan tentang metodologi pengerjaan proyek yang dilakukan. Metodologi ini merupakan langkah-langkah yang akan dilakukan, berikut penjelasannya.



Gambar 1. Metode Pengerjaan Proyek

Berikut ini keterangan pada Gambar 1.

1. Pengumpulan Data

Pertama sekali akan dilakukan pengumpulan data-data yang berhubungan dengan masalah proyek. Dimana akan kucari kumpulan dataset dari berbagai sumber penelitian orang lain dengan tujuan berbeda tetapi dapat dimanfaatkan. Dimana data yang diperoleh merupakan data *review customer* yang telah datang untuk menginap.

2. Pengolahan Data

Pada tahapan ini, ketika data sudah dikumpulkan maka data tersebut akan diolah dengan cara *Data Preprocessing*. *Data Preprocessing* dilakukan untuk pembersihan dan persiapan data dengan tujuan menghilangkan konsistensi data, data yang tidak lengkap, dan redundant data yang terdapat pada dataset yang ada.

3. Metode yang Diusulkan

Metode yang diusulkan untuk pengerjaan proyek *sentiment analysis* ini ialah dengan penerapan metode *Naïve Bayes Classification* dan *Support Vector Machine* (SVM).

4. Eksperimen dan Pengujian

Tahapan untuk melakukan penerapan metode *Naïve Bayes Classification* dan *Support Vector Machine* (SVM). Kemudian dilakukan pengujian kedua metode terhadap dataset untuk mengetahui perbedaan komentar positif dan negatif pada *review customer*.

5. Evaluasi dan Validasi Hasil

Setelah melakukan tahapan sebelumnya maka akan dilakukan evaluasi dan validasi dengan mengukur akurasi dari hasil kedua metode yang diterapkan. Kemudian akan disimpulkan hasil perbandingan kedua metode pada model *sentiment analysis.*

BAB 3 RENCANA KERJA

# 3.1 Pembagian Tugas

Berikut adalah tabel pembagian tugas setiap anggota kelompok.

Tabel 1. Pembagian Tugas

|  |  |  |
| --- | --- | --- |
| Nama | Jabatan | Tugas |
| Krista Lumbantoruan | *Team Leader* | Beperan untuk pembagian tugas dan memantau perkembangan proyek. Bertanggung jawab untuk perencanaan dan pengkoordinasian pengerjaan proyek yang akan dikerjakan. |
| Aulia Pakpahan | *Programmer* | Berperan dalam mengimplementasikan code untuk membangun sistem dan melakukan pengujian terhadap sistem yang telah dibangun. |
| Evi Fanny Dear Sidabutar | *Programmer* | Berperan dalam mengimplementasikan code untuk membangun sistem dan melakukan pengujian terhadap sistem yang telah dibangun. |
| Maria Anastasia Br. Simanullang | *Data Analyst* | Beperan dalam mengidentifikasi, menafsirkan serta menganalisis data dan strategi yang efisisen untuk digunakan dalam pengerjaan proyek. |

# 3.2 Jadwal

Jadwal pengerjaan proyek yang dilakukan dapat dilihat pada tabel berikut.

Tabel 2. Jadwal

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No | Kegiatan | Week | | | | | |
| 11 | 12 | 13 | 14 | 15 | 16 |
| 1 | Menentukan kelompok proyek |  |  |  |  |  |  |
| 2 | Menentukan topik proyek akhir |  |  |  |  |  |  |
| 3 | Pengerjaan dan pengumpulan Proposal |  |  |  |  |  |  |
| 4 | Pengerjaan Proyek diawali dengan pengumpulan data |  |  |  |  |  |  |
| 5 | Menganalisis Data untuk pengolahan Data |  |  |  |  |  |  |
| 6 | Pengerjaan dan pengujian terhadap penerapan metode menggunakan SVM dan *Naïve Bayes*. |  |  |  |  |  |  |
| 7 | Evaluasi dan Validasi Hasil |  |  |  |  |  |  |
| 8 | Pengerjaan laporan akhir proyek |  |  |  |  |  |  |
| 9 | Persiapan pengumpulan final laporan akhir dan video presentasi |  |  |  |  |  |  |
| 10 | Pengumpulan final laporan akhir dan video presentasi |  |  |  |  |  |  |

# DAFTAR PUSTAKA

[1] A. M. Rahat, A. Kahir, and A. K. M. Masum, “Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset,” *Proc. 2019 8th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2019*, pp. 266–270, 2019, doi: 10.1109/SMART46866.2019.9117512.

[2] S. Aji *et al.*, “Review Sentiment Analysis of World Class Hotel Using Naive Bayes Classifier And Particle Swarm Optimization Method,” no. January, 2019, doi: 10.4108/eai.24-10-2018.2280546.

[3] E. Haddi, X. Liu, and Y. Shi, “SENTIMENT ANALYSIS OF HOTEL REVIEW USING NAÏVE BAYES ALGORITHM AND INTEGRATION OF INFORMATION GAIN AND GENETIC ALGORITHM AS FEATURE SELECTION METHODS,” *Int. Semin. Sci. Issues Trends Bekasi*, 2014.

[4] T. Annisa, “Mengenal peran sentiment analysis beserta cara kerjanya,” 2020.

[5] H. Yu, J. Vaidya, and X. Jiang, “Privacy-preserving SVM classification on vertically partitioned data,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3918 LNAI, pp. 647–656, 2006, doi: 10.1007/11731139\_74.

[6] C. Matthews, “An introduction to natural language processing through prolog,” *An Introd. to Nat. Lang. Process. Through Prolog*, no. May, pp. 1–306, 2016, doi: 10.4324/9781315845593.