**PROPOSAL PROYEK DATA MINING**

***Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine***



**Disusun oleh:**

1. 12S17028 – Andri Reimondo Tamba
2. 12S17029 – Silvany Angelia Lumban Gaol
3. 12S17041 – Dewi Purnama Napitupulu

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS TEKNIK INFORMATIKA DAN ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**NOVEMBER 2020**

# DAFTAR ISI

[DAFTAR ISI ii](#_Toc56688164)

[1 Business Understanding 1](#_Toc56688165)

[**1.1 Determine Bussiness Objective** 1](#_Toc56688166)

[**1.2 Situation Assesment** 1](#_Toc56688167)

[**1.3 Determine Data Mining Goal** 1](#_Toc56688168)

[**1.4 Produce Project Plan** 3](#_Toc56688169)

[2 Data Understanding 4](#_Toc56688170)

[**2.1 Collect Initial Data** 4](#_Toc56688171)

[**2.2 Describe Data** 5](#_Toc56688172)

[**2.3 Explore Data** 5](#_Toc56688173)

[**2.4 Verify Data Quality** 6](#_Toc56688174)

[Referensi 8](#_Toc56688175)

# 1 Business Understanding

## **Determine Bussiness Objective**

## Proyek kali ini akan dilakukan dalam lingkup penelitian. Tujuan dari proyek ini adalah menolong para pembaca berita untuk lebih mudah memilih berita yang akan di bacanya. Hal tersebut dapat direalisasikan dengan mengklasifikasikan berita kedalam beberap kategori. Penelitian ini menggunkan beberapa metode yaitu Support Vactor Machine (SVM) yang bertujuan adalah untuk membangun OSH (Optimal Separating Hyperplane), yang membuat fungsi pemisahan optimum yang dapat digunakan untuk klasifikasi. Penelitian ini juga menggunakan metode Naïve Bayes Classifier yang menyediakan data training. Tujuan adanya data training adalah untuk menghasilkan model dari NBC untuk mengetahui ketepatan dari klasifikasi selain itu pada data training juga memperhatikan waktu yang diperlukan pada pembentukan model. Jadi penelitian ini bermaksud untuk melakukan perbandingan antara kedua metode tersebut diatas untuk mengetahui tingkat akurasi dari masing-masing metode. Dengan demikian dapat disimpulkan metodde yang paling baik dalam melakukan klasifikasi.

## 

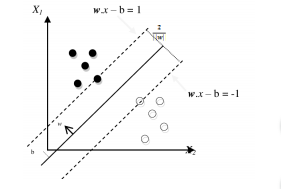
## **1.2 Situation Assesment**

## Pada tahun 2006 pertumbuhan dan pertukaran informasi sudah mencapai lebih dari 550 triliun dokumen dan 7,3 juta Internet page baru tiap harinya. [1] Dengan demikian berdampak pada artikel berita yang diunggah di internet sangat banyak dalam rentang waktu yang cepat. Selama ini pengkategorian berita masih menggunakan tenaga manusia dan secara manual.Kategori yang banyak beserta waktu yang cepat akan menyulitkan editor untuk mengkategorikan, terutama artikel yang tidak terlalu berbeda secara jelas. Beberapa kategori yang penggunaan bahasanya tidak berbeda terlalu jauh seperti nasional, internasional, sains, ekonomi, tekno, *health*, dan properti mengharuskan seorang editor mengetahui isi artikel yang akan diunggah secara keseluruhan untuk selanjutnya dimasukkan ke dalam kategori yang tepat. Akan lebih efisien apabila kategori berita dimasukkan secara otomatis dengan computer menggunakan metode tertentu.

## **1.3 Determine Data Mining Goal**

Text mining adalah cara agar teks dapat diolah dengan menggunakan komputer untuk menghasilkan analisis yang bermanfaat. Praproses dalam text mining diantaranya adalah tokenizing, case folding, stopwords, dan stemming. Diantara keempat langkah tersebut yang paling penting adalah proses stemmingyang merupakan proses menghilangkan imbuhan pada suatu kata untuk mendapatkan kata dasar dari kata tersebut. Penelitian kali ini akan menggunakan Text Mining untuk pengklasifikasian Betrita Indonesia. Berikut Startegi utama yang digunakan :

1. SVM adalah metode *supervised learning* yang digunakan untuk klasifikasi serta regresi. Klasifikasi dan regresi keduanya adalah sub kategori dari *supervised learning.* Klasifikasi adalah sesuatu yang dapat didefinisikan sebagai memprediksi label sedangkan regresi adalah tentang memprediksi kuantitas. Jadi tugas utama *classifier Support Vector Machine* adalah melakukan klasifikasi, yaitu mengklasifikasikan data dalam kelas yang berbeda dengan menggambar hyperplane yang membedakan antara kelas berbeda yang kita plot dalam ruang n-dimensi [2]



Gambar 1 *Hyperplane* pada SVM

Sumber : *ICTACT Journal On Soft Computing:Detection On Twitter Data Using Support Vector Machine(*Ashima Garg and Neelam Duhan, 2020)

*Hyperplane* yang digambar oleh SVM digambar dengan bantuan fungsi matematika yang disebut kernel. Titik data yang paling dekat dengan *hyperplane* disebut vektor dukungan dan metode ini disebut *Support Vector Machine.* Kernel yang akan digunakan dengan penelitian yang ditunjukkan di persamaan di bawah ini:



1. Naïve Bayes merupakan algoritma supervised learning yang menggunakan teorema Bayes untuk masalah klasifikasi. Metode ini sebagian besar digunakan dalam klasifikasi teks yang mencakup kumpulan data untuk pelatihan dimensi tinggi. Ini adalah salah satu algoritma klasifikasi paling sederhana dan paling kuat yang membantu membuat model pembelajaran mesin cepat yang dapat membuat prediksi dengan cepat. Ini banyak digunakan untuk analisis teks, analisis sentimen dan artikel klasifikasi. Terdapat beberapa teorema Bayes, sebagai berikut : [3]

Prob (A | B) dikenal sebagai (probabilitas posterior) dari hipotesis A yang terjadi jika suatu kondisi telah disediakan.

Prob (B | A) itu juga disebut probabilitas kemungkinan itu adalah probabilitas bukti E ketika kita mengasumsikan bahwa hipotesis yang diberikan benar.

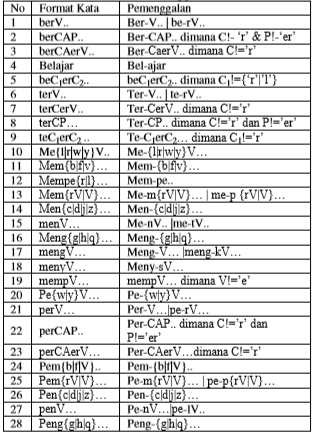
Prob (A) itu adalah probabilitas yang diketahui sebelumnya dari A dan tidak termasuk kondisi apapun.

Prob (B) itu adalah probabilitas yang diketahui sebelumnya dari A dan tidak termasuk kondisi apapun.

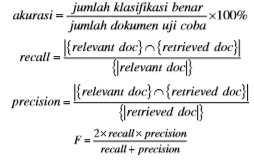
## **1.4 Produce Project Plan**

Project plan merupakan komponen rencana proyek yang dilakukan selama pengerjaan proyek. Pada proyek berikut, yang menjadi tahapan perencanaan dalam pengerjaan proyek:

* Pengumpulan dataset, berupa artikel bahasa Indonesia dan melakukan analisis terhadap data.
* Melakukan tahapan pra pemrosesan data untuk mencegah adanya redundansi pada data.
* Mengukur kketepatan dan waktu klasifikasi SVM terhadap data training.
* Melakukan pemenggalan kata, seperti pada gambar berikut:



* Melakukan pengukuran performa, berupa: akurasi, recall, serta precision.



* Melakukan tahapan analisis yang meliputi klasifikasi pada teks menggunakan SVM.

# 2 Data Understanding

## **2.1 Collect Initial Data**

Tahapan Collect Initial Data merupakan tahapan dimana analis melakukan proses identifikasi terhadap data dan mampu menjabarkan secara jelas dan singkat mengenai data yang diperoleh dalam tahapan pengerjaan data mining selanjutnya. Dalam mengklasifikasi artikel berita untuk menemukan algoritma mana yang terbaik dalam klasifikasi artikel berita, dataset yang digunakan merupakan kumpulan artikel berita yang dapat dikumpulkan dari berbagai sumber, terutama pada internet. Penulis telah mencari dan menetapkan data berupa artikel berita berformat .csv yang telah dikumpulkan dari kaggle: <https://www.kaggle.com/aashari/indonesian-news-articles-published-at-2017> sebagai dataset yang digunakan dalam pengerjaan proyek.

Dataset artikel berita merupakan data sequence yang bersifat dependency terhadap antar kata pada suatu teks. Namun, dikarenakan data berbentuk teks memiliki bentuk yang lebih terstruktur. Astikel berita tersebut kemudian akan melakukan pemenggalan kata berdasarkan beberapa kategori, seperti: *Inflection Suffixes*, *Derivation Suffixes*, dan *Derivation Prefixes.* Adapun setiap kata yang dipenggal akan diklasifikasi termasuk kelas kata yang mana. Dari sini dapat kita lihat bahwa atribut data pada setiap kata merupakan data nominal, dimana setiap kelas kaya tidak dapat diurutkan namun memiliki jumlah tetentu pada setiap artikelnya

## **2.2 Describe Data**

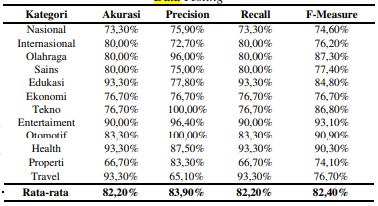
Pada tahapan ini, akan dijabarkan sedikit mengenai data yang digunaan, serta akan ditinjau permasalahan yang ditemukan dalam data, yang kemudian akan diidentifikasi sebelum melakukan tahapan deployment pada proyek yang akan dikerjakan. Pada topik ini, dataset yang digunakan merupakan artikel berita berformat .csv yang dapat diperoleh dari internet. Data yang digunakan berjumlah 500 data. kesulitan yang dihadapi dalam menggunakan data ialah:

1. data yang diperoleh masih noisy,
2. masih terdapat beberapa values yang kosong
3. masih menggunakan tanda baca, dan lainnya.

Maka dari itu, dibutuhkan tahapan pre-pemrosesan pada data untuk menghasilkan data yang bersih, dan agar proses data mining pada data dapat dilakukan dengan lebih cepat. Adapun kesulitan lain yang ditemui pada data yang digunakan adalah dataset yang digunakan merupakan artikel berita. Dimana data berupa teks merupakan data berbentuk sequence yang memiliki bentuk yang lebih terstruktur dari data lainnya.

## **2.3 Explore Data**

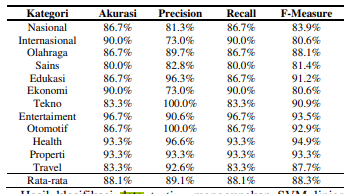
Pada proses ini dilakukan dengan menggunakan fungsi statistik, matematik, dan divisualisasikan dalam bentuk grafik. Hal ini dilakukan untuk mempermudah dalam pemahaman dan pola dasar data. Selanjut nya dilakukan explore data untuk menemukan data yang bermanfaat untuk data penelitian, dan dapat digunakan untuk proses penambangan data. Pada tahap ini juga dilakukan pra-proses data misalnya pembersihan data outlier dan penanganan data yang kosong, sehingga siap untuk diolah dan dianalisa.



Pada gambar diatas diamati bahwa dataset memiliki 5 kolom dan 12 baris. Dari gambar dapat dilihat niali dari Akurasi, Precission, Recall dan F-Measure terhadap baris masing masing data. Berdasarkan rata-rata *akurasi, recall, precision*, dan *F-Measure* pada Tabel memperlihatkan hasil yang cukup baik. Masing-masing nilainya adalah 82,2%, 83,9%, 82,2%, dan 82,4%. Untuk tingkat akurasi paling tinggi dihasilkan oleh kategori berita edukasi, health, dan travel dengan nilai akurasi 93,3%. Berbeda dengan tiga kategori tersebut kategori berita properti menghasilkan akurasi yang paling rendah yaitu 66,7%. Untuk ukuran *precision* kategori tekno dan otomotif bernilai 100% sebaliknya travel menjadi yang paling rendah yaitu 65,1%. *Recall* paling tinggi terdapat pada kategori edukasi, health, dan travel sedangkan untuk paling rendah adalah properti. Untuk ukuran gabungan dari *precision* dan *recall* yaitu *F-Measure* memperlihatkan bahwa kategori entertaiment adalah yang paling tinggi sedangkan yang paling rendah adalah properti. EDA juga berfungsi untuk mengoptimalkan pengetahuan mengenai data. Salah satu metode tradisional dalam EDA adalah visualisasi dalam bentuk grafik.

## **2.4 Verify Data Quality**

Pada proses ini dilakukan untuk menemukan kualitas dari dataset yang akan digunakan apakah dataset tersebut terdapat *outlier* dan *missing value* atau tidak. Setelah mengetahui bahwa dataset yang dimiliki mengandung *outlier* ataupun *missing value*, atau dengan mengisi baris data yang mengandung *outlier* tersebut dengan suatu nilai yang konstan. Skenario yang diambil dalam penelitian ini adalah menghapus kolom yang mengandung *missing value* dan juga *outlier* dengan mempergunakan fungsi dropna() dan drop().



Pada gambar diatas didapati setelah melakukan *testing* data hasil klasifikasi data testing menggunakan SVM linier pada Tabel ini menunjukkan performa yang cukup baik dengan masing-masing nilai dari akurasi, *precision, recall*, dan *F-Measure* adalah 88,1%, 89,1%, 88,1%, dan 88,3%. Kategori berita entertaiment menjadi kategori dengan tingkat akurasi yang paling tinggi yaitu 96,7%, sebaliknya sains menjadi kategori dengan tingkat akurasi yang paling rendah yaitu 80,0%. Untuk precision dengan nilai paling baik adalah kategori tekno dan otomotif sebesar 100%, sedangkan kategori internasional dan ekonomi adalah kategori dengan precision terendah sebesar 73,0%. Hasil recall tertinggi adalah kategori entertaiment dengan nilai sebesar 96,7% dan terendah adalah kategori sains dengan nilai sebesar 80,0%. Nilai F-Measure menunjukkan performa yang paling baik adalah kategori health 94,9%, sedangkan yang paling rendah adalah kategori internasional dan ekonomi 80,6%. Kemudian data tersebut diverify setelah melewati *testing data*.

# Referensi

|  |  |
| --- | --- |
| [1] | D. Ariadi and K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Confix Stripping Stemmer," *JURNAL SAINS DAN SENI ITS,* pp. 2337-3520, 2015. |
| [2] | R. Suman and J. Slngh, "Sentimen Analysis of Tweets Using Support Vector Machine," *International Jurnal of Computer Science and Mobile Apllications,* vol. 5, no. 10, pp. 83-91, 2017. |
| [3] | M. L and P. R, "Twitter Sentimen Analysis Using Machine Learning Methods," *International Conference for Emerging Technology (INCET),* pp. 1-5, 2020. |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |