BAB III
METODELOGI PENELITIAN

3.1 Subjek dan Objek Penelitian

Subjek penilitian merupakan orang, tempat, atau benda yang diamati. Subjek pada penelitian ini ialah para pengguna atau berbagai orang yang memiliki media sosial *Twitter*. Objek penelitian ialah suatu atribut yang berasal dari orang atau kegiatan yang ditetapkan oleh peneliti. Objek penelitian ini ialah klasifikasi suatu kasus yang masih terjadi di media sosial *Twitter* seperti *Cyber harassment.*

3.2 Alat dan Bahan

Alat dan Bahan yang digunakan dalam penelitian ini antara lain:

3.2.1. Perangkat keras

Perangkat keras yang dibutuhkan dan digunakan dalam pengembangan penelitian ini dengan sebuah laptop yang memiliki spesifikasi:

1. Prosesor Intel® Core™ i5-8265U
2. RAM 8 GB
3. Storage 1TB HDD
4. Display 14” Full HD
5. Graphics Nvidia GeForce MX150 2GB

3.2.2. Perangkat lunak

Perangkat lunak yang dibutuhkan dan digunakan dalam pengembangan penelitian ini, sebagai berikut:

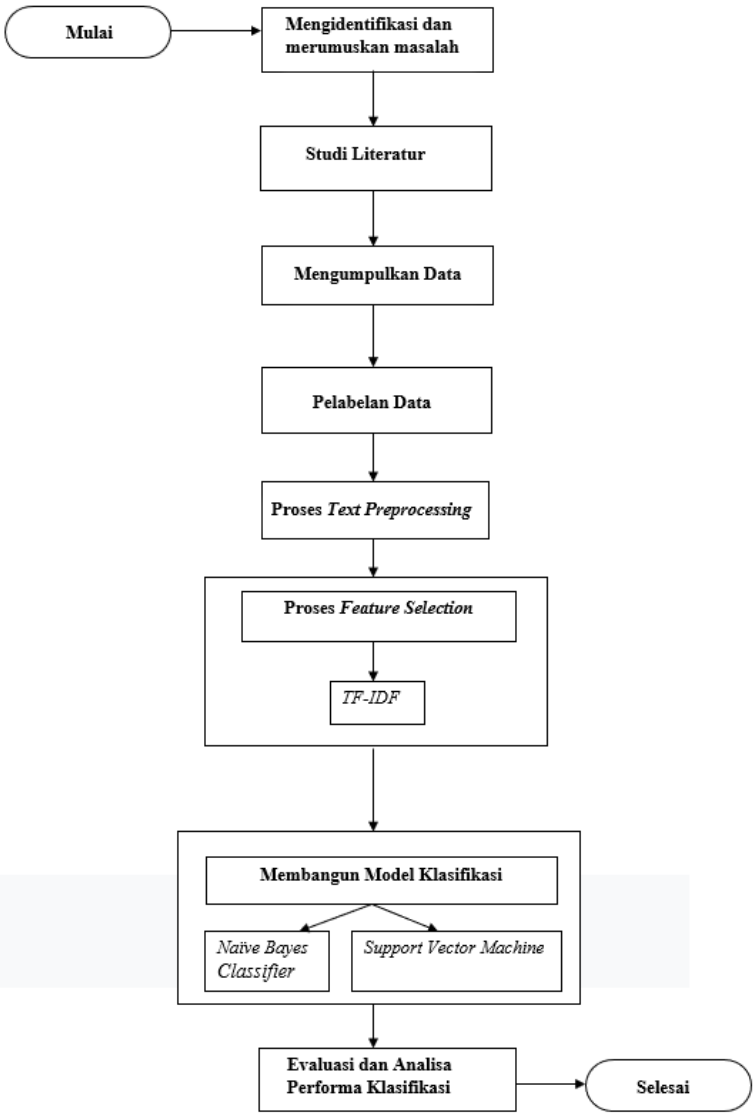
1. Sistem operasi Windows 10
2. Jupyter Notebook (Anaconda)
3. Python 3.7
4. Command Prompt
5. Browser
6. Twitterscraper
7. Rapidminer

3.2.3. Bahan

Bahan penelitian yang digunakan untuk penelitian ini ialah kumpulan data *tweet* pada media sosial *Twitter* yang dapat di kategorikan atau berlabel menjadi kasus *cyber harassment* atau tidak.

3.3 Diagram Alir Penelitian

Penelitian Perbandingan Metode *Naïve Bayes Classifier* dan *Support Vector Machine* Untuk Klasifikasi *Cyber Harassment* Pada *Twitter* dilakukan dengan beberapa tahap antara lain:



Gambar 3.1 Diagram Alir Penelitian

3.3.1. Mengidentifikasi dan merumuskan masalah

Langkah pertama yang dilakukan pertama kali yaitu mengidentifikasi dan merumuskan masalah tentang apa yang akan diteliti. Menentukan bidang, topik,

masalah penelitian serta mengusulkan metode yang akan digunakan pada penelitan. Pada tahap ini mempelajari masalah yang masih terjadi di kehidupan sehari-hari, lalu tujuan, ruang lingkup serta metodologi penelitian. Pada penelitian ini ide atau masalah sudah ada penelitian terdahulunya, namun terdapat perbedaan pada algoritma atau metode yang di usulkan.

3.3.2. Studi literatur

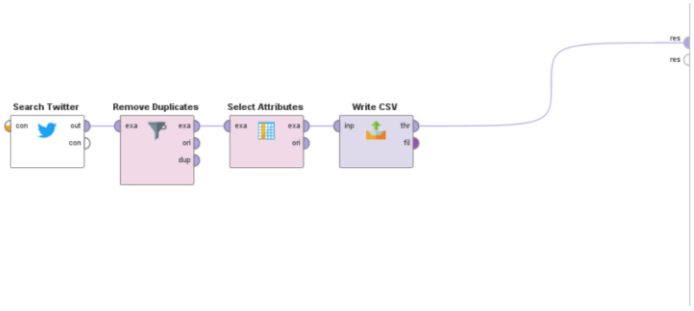
Langkah selanjutnya dengan melakukan studi literatur atau riset kepustakaan. Pada tahap ini, peneliti melakukan pengumpulan data – data yang berkaitan dengan topik permasalahan yaitu tentang klasifikasi, kasus *Cyber harassment*, tentang tweet di *Twitter* serta algoritma yang akan diusulkan yaitu *naives bayes* dan *support vector machine*. Data – data untuk penilitian ini diperoleh dari jurnal, buku elektronik, situs internet dan media elektronik. Studi literatur memiliki tujuan untuk memperkuat permasalahan yang dibahas pada penelitian ini serta menjadi dasar untuk melakukan pengembangan selanjutnya.

3.3.3. Mengumpulkan Data

Langkah berikutnya adalah mengumpulkan data, data yang dikumpulkan berupa *tweet* pada *Twitter* tentang *cyber harassment.* Mengumpulkan data ini bertujuan agar mendapatkan informasi tentang kasus *cyber harassment* ini untuk dapat di analisis dengan cara klasifikasi menggunakan metode *Naïve Bayes* dan *Support Vector Machine.* Dalam mengumpulkan data, peneliti melakukan dengan cara *crawling* menggunakan RapidMiner.

Langkah-langkah di lakukan pada proses *crawling* pada RapidMiner adalah:

1. Membuat rangkaian dengan operator *Search Twitter*, *Remove Duplicate*, *Select Atributes*, dan *Write CSV* seperti gambar 3.2.



Gambar 3.2 **Rangkaian Operator di RapidMiner**

2. Pada operator *Search Twitter* memberi parameter seperti menghubungkan API menggunakan akun *Twitter* penulis, memberi nama rangkaian, mengimput kata kunci yang dicari, jumlah data yang di targetkan, dan sebagaiannya seperti gambar 3.3.

Gambar 3.3 Format parameter *crawling*

3. Data akan tersimpan dalam bentuk *csv* atau *excel* namun penulis menggunakan operator Write CSV untuk menghasilkan data berupa *csv.*

Kriteria data yang dikumpulkan untuk penelitian ini sebagai berikut:

1. Data bersumber dari *Twitter* dalam bahasa Indonesia metode yang digunakan untuk mendapatkan data yaitu dengan menggunakan API *Twitter.*
2. Data menyesuaikan kata kunci dari berbagai jenis-jenis *cyber harassme*nt yaitu *Sexual Harassment, Racial Harassment, Appearance-related Harasmment, Intellectual Harasmment,* dan *Political Harasmment* serta memberikan pelabelan ke semua jenis yaitu pilihan ya atau tidak mengandung unsur *cyber harassme*nt.

Tabel 3.1 Sampel Tweet di Twitter

No	Tweet
1	TOFA ANJING, INILAH MANUSIA YG TERLAHIR DARI RAHIM SEORANG PELACUR. MULUTNYA KAYAK ANUS
2	Pelacur intelek.!
3	Seorang anggota DPR menjebak #pelacur untuk menaikan popularitas nya. itulah #pelacur politik yg sesungguhnya

4	Ketololan dalam bersosmed. Memperolok olok satu sama lain. Over dalam merargumen. Karena rasis itu harus mati!
5	Cemen lho Nil nil, provokator bangsa.percuma petetang petenteng pake peci hitam berjenggot,sikap kayak pelacur politik,penjilat stadium4.mikirrr.....
6	Mesum gak baik nak,

3.3.4. Pelabelan Data

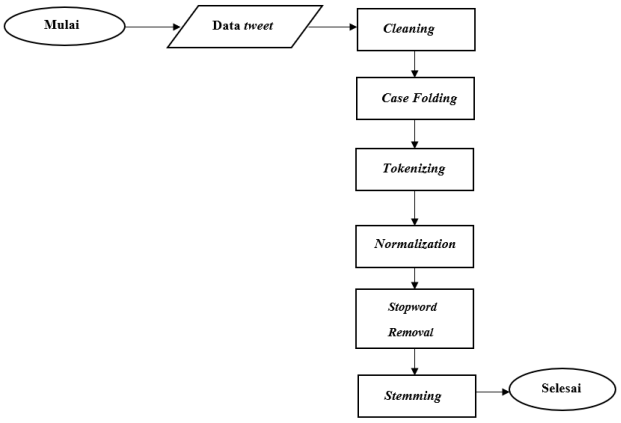
Setelah mengumpulkan beberapa data, data tersebut masih bersifat mentah maka dibutuhkan label pada data tersebut dan bertujuan agar bisa melakukan klasifikasi. Label "1" menandakan bahwa teks mengandung unsur pelecehan atau *harassment.* Sedangkan label "0" menandakan bahwa teks tidak mengandung unsur pelecehan atau *harassment.* Proses pelabelan data dilakukan secara manual yang dilakukan oleh ahli dan asisten ahli. Data *tweet* menggunakan bahasa Indonesia sehingga penelitian ini membutuhkan Kamus Besar Bahasa Indonesia (KBBI).

Tabel 3.2 Pelabelan data

No	Tweet	Label
1	TOFA ANJING, INILAH MANUSIA YG TERLAHIR DARI RAHIM SEORANG PELACUR. MULUTNYA KAYAK ANUS	1
2	Pelacur intelek.!	1
3	Seorang anggota DPR menjebak #pelacur untuk menaikan popularitas nya. itulah #pelacur politik yg sesungguhnya	1
4	Ketololan dalam bersosmed. Memperolok olok satu sama lain. Over dalam merargumen. Karena rasis itu harus mati!	1
5	Cemen lho Nil nil, provokator bangsa.percuma petetang petenteng pake peci hitam berjenggot,sikap kayak pelacur politik,penjilat stadium4.mikirrr.....	1
6	Mesum gak baik nak,	0

3.3.5. *Text Preprocessing*

Setelah melakukan pelabelan pada data maka selanjutnya langkah untuk *text preprocessing* yaitu *cleaning* dengan cara penghapusan retweet, URL, *emoticon,* dan selanjutnya *case folding, tokenizing, normalization, stopword removal* dan *stemming.*



Gambar 3.4 Proses *Text Preprocessing*

1. Cleaning

Dengan cara kata atau karakter yang akan dihilangkan seperti simbol, link URL, taggar (#), nama pengguna atau mention (@namaakun), *emoticon* dan RT (*Retweet*). Contoh proses *cleaning* dapat dilihat pada tabel berikut ini.

Tabel 3.3 *Cleaning*

No	Sebelum melakukan *Cleaning*	Setelah melakukan *Cleaning*
1	Seorang anggota DPR menjebak #pelacur untuk menaikan popularitas nya. itulah #pelacur politik yg sesungguhnya	Seorang anggota DPR menjebak pelacur untuk menaikan popularitas nya. itulah pelacur politik yg sesungguhnya

2. Case Folding

Case Folding adalah mengubah semua karakter huruf menjadi huruf kecil (lowercase). Proses *case folding* dapat dilihat pada tabel berikut ini. Proses *case folding* dapat dilihat pada tabel berikut ini.

Tabel 3.4 *Case Folding*

No	Sebelum melakukan *case folding*	Setelah melakukan *case folding*
1	Seorang anggota DPR menjebak pelacur untuk menaikan popularitas nya. itulah pelacur politik yg sesungguhnya	seorang anggota dpr menjebak pelacur untuk menaikan

		popularitas nya. itulah pelacur politik yg sesungguhnya

3. Tokenizing

Tokenizing yaitu proses penguraian deskripsi yang semula berupa kalimat-kalimat menjadi kata-kata dan menghilangkan delimiter-delimiter seperti tanda titik (.), koma (,), spasi dan karakter angka yang ada pada kata. Contoh proses *tokenizing* dapat dilihat pada tabel berikut ini.

Tabel 3.5 *Tokenizing*

No	Sebelum melakukan *Tokenizing*	Setelah melakukan *Tokenizing*
1	seorang anggota dpr menjebak pelacur untuk menaikan popularitas nya. itulah pelacur politik yg sesungguhnya	seorang
		anggota
		dpr
		menjebak
		pelacur
		untuk
		menaikan
		popularitas
		nya
		itulah
		pelacur
		politik
		yg
		sesungguhnya

4. Normalization

Normalization merupakan perbaikan dan subtitusi kata yang salah eja ataupun disingkat dengan bentuk tertentu. Subtitusi kata dilakukan untuk menghindari perhitungan dimensi kata yang melebar. Perhitungan dimensi kata akan melebar jika kata disingkat atau salah eja tidak diubah karena kata tersebut memilki kontribusi dalam mempresentasikan dokumen tetapi akan dianggap sebagai entitas yang berbeda proses penyusunan matriks.

5. *Stopword Removal*

Stopword Removal yaitu proses penghapusan kata-kata yang terdapat pada *stoplist*. *Stoplist* itu sendiri berisi kosakata-kosakata yang bukan merupakan ciri dari suatu dokumen dan bertujuan agar menghilangkan kata yang tidak berguna. Contoh *stopword removal* bahasa Indonesia adalah kata hubung "dan", "yang", dan lain-lain. Contoh proses *stopword removal* dapat dilihat pada tabel berikut ini.

No	Sebelum melakukan Stopword Removal	Setelah melakukan Stopword Removal
1	seorang anggota dpr menjebak pelacur untuk menaikkan popularitas nya itulah pelacur politik yg sesungguhnya	seorang
		anggota
		menjebak
		pelacur
		menaikkan
		popularitas
		politik
		sesungguhnya

Tabel 3.6 *Stopword Removal*

6. Stemming

Stemming ialah suatu proses pemetaan dan penguraian berbagai bentuk (variants) dari suatu kata menjadi bentuk kata dasarnya (*stem*). *Stemming* bertujuan untuk menghilangkan berbagai imbuhan seperti prefiks, sufiks, maupun konfiks yang ada pada setiap kata. Contoh proses *stemming* dapat dilihat pada tabel berikut ini.

Tabel 3.7 *Stemming*

No.	Sebelum melakukan Stemming	Setelah melakukan Stemming
1	seorang	orang
	anggota	anggota
	menjebak	jebak
	pelacur	lacur

	menaikkan	naik
	popularitas	populer
	politik	politik
	sesungguhnya	sungguh

3.3.6. Proses *Feature Selection*

Selanjutnya data akan melakukan proses *feature selection* atau seleksi fitur menggunakan pembobotan *TF-IDF*. Adanya pembobotan fitur ini sebelum klasifikasi dilakukan agar membantu meningkatkan akurasi klasifikasi. Contoh dalam proses pembobotan dari data pada tabel 3.2 di atas yang telah melalui tahap *text preprocessing*.

Tabel 3.8 *Pembobotan TF* dan DF

No.	Fitur	Dok 1	Dok 2	Dok 3	Dok 4	Dok 5	Dok 6	DF
1	anjing	1	0	0	0	0	0	1
2	ini	1	0	0	0	0	0	1
3	manusia	1	0	0	0	0	0	1
4	lahir	1	0	0	0	0	0	1
5	rahim	1	0	0	0	0	0	1
6	orang	1	0	1	0	0	0	2
7	lacur	1	1	2	0	1	0	5
8	mulut	1	0	0	0	0	0	1
9	kayak	1	0	0	0	0	0	1
10	anus	1	0	0	0	0	0	1

Tabel 3.9 *Hasil TF-IDF*

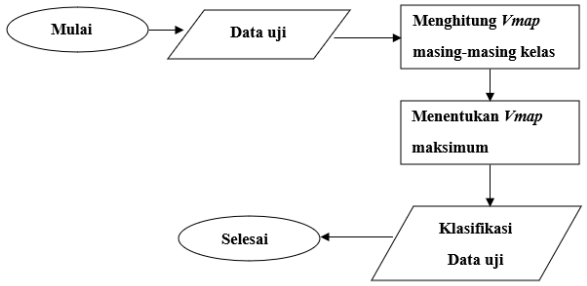
No.	Fitur	IDF	TF-IDF					
			Dok 1	Dok 2	Dok 3	Dok 4	Dok 5	Dok 6
1	anjing	Log(6/1)=0,77815	0,77815	0	0	0	0	0
2	ini	Log(6/1)=0,77815	0,77815	0	0	0	0	0
3	manusia	Log(6/1)=0,77815	0,77815	0	0	0	0	0
4	lahir	Log(6/1)=0,77815	0,77815	0	0	0	0	0
5	rahim	Log(6/1)=0,77815	0,77815	0	0	0	0	0
6	orang	Log(6/2)=0,47712	0,47712	0	0,77815	0	0	0
7	lacur	Log(6/5)=0,07918	0,07918	0,07918	0,15836	0	0,07918	0

8	mulut	Log(6/1)=0,77815	0,77815	0	0	0	0	0
9	kayak	Log(6/1)=0,77815	0,77815	0	0	0	0	0
10	anus	Log(6/1)=0,77815	0,77815	0	0	0	0	0

3.3.7. Membangun Model Klasifikasi

Selanjutnya membangun model klasifikasi menggunakan metode *Naïve Bayes Classifier* dan *Support Vector Machine* yang merupakan bagian data mining, proses ini merupakan inti dari proses ekstraksi pengetahuan dari data. Pengujian dataset pada penelitian ini dilakukan menggunakan dua algoritma klasifikasi yaitu *Naïve Bayes Classifier* dan SVM yang akan diimplementasikan menggunakan bahasa pemrograman *Python*. Pada model klasifikasi terdapat proses *training* dan *testing*, dalam analisis ini data dibagi menjadi 80% dan 20%. Proses *training* menggunakan dataset ialah proses melatih mesin tentang pengetahuan dataset yang sudah diberi label sebelumnya. Kemudian proses testing, mesin akan di uji menggunakan 20% dari total dataset untuk mengetahui tingkat akurasi dan klasifikasinya.

1. Metode *Naïve Bayes Classifier*



Gambar 3.5 Proses Klasifikasi Metode NBC

Naïve Bayes Classifier sebagai algoritma *supervised learning* seharusnya diberi pengetahuan awal sebagai acuan untuk dapat klasifikasi suatu dokumen. Ada 3 langkah dalam proses learning:

a. Membentuk fitur

Fitur dalam penilitian ini adalah kata penting yang menjadi parameter satuan data latih, yaitu dokumen yang merupakan sampel *tweet* untuk diklasifikasikan kedalam 2 kelas yaitu “1” dan “0”.

Tabel 3.10 Pembentukan fitur Data Latih

Dokumen	Fitur (kemunculan)	Kelas/Label
Dok 1	dasar(1), tolol(1)	1
Dok 2	tolol(1), rasis(1)	1
Dok 3	makan(1), besar(1)	0
Dok 4	anak(1), kecil(1), tidur(1)	0

b. Menghitung Probabilitas Kelas/Label

Setelah membentuk fitur dengan kemunculan dari data latih dan menghitung dengan rumus persamaan **3.1**.

$$P(c_i) = \frac{fd(c_i)}{|D|}$$

……**3.1**

Keterangan:

$P(c_i)$: Menentukan probabilitas c_i yang merupakan kategori kelas

$fd(c_i)$: Jumlah dokumen c_i

$|D|$: Jumlah data latih/dokumen

c. Menentukan Probabilitas setiap Fitur

Setelah mendapatkan probabilitas dari setiap kelas, selanjutnya menghitung probabilitas setiap fitur pada setiap kelas dengan persamaan **3.2**.

$$P(w_k|c_i) = \frac{f(w_{ki},c_i)+1}{P(c_i)+|W|}$$

……**3.2**

Keterangan:

$P(w_k|c_i)$: Peluang kemunculan kata-kata pada sebuah kategori/kelas, w_k adalah kata yang muncul pada sebuah kategori.

$f(w_{ki}, c_i)$: Nilai kemunculan kata w_{ki} pada kelas c_i

$f(c_i)$: Jumlah keseluruhan kemunculan kata pada kelas c_i

$|W|$: Jumlah semua kata dari semua kategori

Tahapan setelah melakukan uji data adalah sebagai berikut:

a. Menghitung Vmap

Vmap adalah perhitungan yang digunakan *naïve bayes classifier* untuk menentukan probabilitas data uji dari masing – masing kelas

berdasarkan dari proses learning. Nilai probabilitas yang terbesar akan dipilih.

$$Vmap = \underset{\{kelas\ 0, kelas\ 1\}}{argmax} \prod_{i=1}^{n} P(w_k|c_i) x P(c_i) \ldots\ldots\ldots\ldots\ldots\ldots \mathbf{3.3}$$

Keterangan:

$P(w_k|c_i)$: Peluang kemunculan kata-kata pada sebuah kategori/kelas, w_k adalah kata yang muncul pada sebuah kategori.

$P(c_i)$: Menentukan probabilitas c_i yang merupakan kategori kelas

b. Menentukan Vmap maksimum

Setelah menghitung Vmap maka membandingkan nilai Vmap kelas 1 (positif) dan Vmap kelas 0 (negatif) yang lebih besar. Sehingga dapat menyimpulkan bahwa *tweet* tersebut di klasifikasikan ke dalam kelas positif atau negatif.

2. Metode *Support Vector Machine*

Berikut merupakan langkah-langkah analisis *Support Vector Machine*:

1. Membagi keseluruhan data menjadi data training dan data testing. Data training yang akan digunakan sebagai permodelan sebanyak 80% dan data testing untuk prediksi sebanyak 20%.
2. Menentukan metode pendekatan hyperplane SVM Multikelas yaitu dengan metode SVM Multikelas satu lawan semua (SLA).
3. Menentukan fungsi kernel yang akan digunakan sebagai permodelan hyperplane SVM SLA.
4. Menentukan nilai parameter C dan nilai-nilai parameter kernel yang akan digunakan sebagai permodelan hyperplane SVM SLA.
5. Mendapatkan nilai alpha dan b.
6. Menghitung matriks kernel.
7. Melakukan prediksi klasifikasi.
8. Evaluasi performansi model klasifikasi menggunakan *confusion matrix*. Menghitung akurasi klasifikasi hasil prediksi dan memilih nilai parameter dan fungsi kernel terbaik.

Pada proses *Support Vector Machine* terdapat *hyperplane* yaitu bidang pemisah suatu kelas dengan kelas lainnya. Library yang ialah

library sklearn dengan model SVC (*Support Vector Classifacation*). Parameter yang dibutuhkan pada penelitian ini adalah parameter C *(complexity), d*(*degree*), $\gamma(gamma)$ dan $\sigma(sigma)$ secara *default*.

3.3.8. Evaluasi dan Analisa Performa Klasifikasi

Pada langkah evaluasi ini parameter yang digunakan adalah *confusion matrix.* Tujuan dari adanya evaluasi ini agar dapat melihat nilai *accurasy*, *precision , recall* dan *f1 score* dengan melihat performa model klasifikasi *Naïve Bayes Classifier* dan SVM untuk kasus *cyber harassment* pada Twitter. Setelah mendapatkan hasil dari evaluasi maka dapat di tarik kesimpulan dari penelitian ini.

3.4 Analisis Hasil

Berdasarkan kerangka pemikiran dan paradigma penelitian yang sudah dibahas pada penjelasan sebelumnya, maka peneliti merumuskan hipotesis bahwa algoritma *Naïve Bayes Classifier* dan *Support Vector Machine* memiliki persamaan dengan nilai akurasi yang tinggi namun dengan perbedaan selisih yang sedikit.